AUTHORSHIP PROFILING IN A FORENSIC CONTEXT

ANDREA NINI

Doctor of Philosophy

ASTON UNIVERSITY

March 2014

SUMMARY OF THESIS

Authorship profiling in a forensic context

Andrea Nini

Doctor of Philosophy

March 2014

There are several unresolved problems in forensic authorship profiling, including a lack of research focusing on the types of texts that are typically analysed in forensic linguistics (e.g. threatening letters, ransom demands) and a general disregard for the effect of register variation when testing linguistic variables for use in profiling. The aim of this dissertation is therefore to make a first step towards filling these gaps by testing whether established patterns of sociolinguistic variation appear in malicious forensic texts that are controlled for register. This dissertation begins with a literature review that highlights a series of correlations between language use and various social factors, including gender, age, level of education and social class. This dissertation then presents the primary data set used in this study, which consists of a corpus of 287 fabricated malicious texts from 3 different registers produced by 96 authors stratified across the 4 social factors listed above. Since this data set is fabricated, its validity was also tested through a comparison with another corpus consisting of 104 naturally occurring malicious texts, which showed that no important differences exist between the language of the fabricated malicious texts and the authentic malicious texts. The dissertation then reports the findings of the analysis of the corpus of fabricated malicious texts, which shows that the major patterns of sociolinguistic variation identified in previous research are valid for forensic malicious texts and that controlling register variation greatly improves the performance of profiling. In addition, it is shown that through regression analysis it is possible to use these patterns of linguistic variation to profile the demographic background of authors across the four social factors with an average accuracy of 70%. Overall, the present study therefore makes a first step towards developing a principled model of forensic authorship profiling.

Keywords: forensic linguistics, authorship profiling, authorship analysis, threatening texts, stylometry, register variation, stylistics

Acknowledgments

I would like to thank everybody who helped me through these past four years of my PhD. Thanks to my parents who supported my journey. Thanks to my girlfriend who was there for me when I needed it. Thanks to all of my friends who made these years more special than they might have otherwise been. Above all, however, thanks to my grandfather, who taught me patience.

"Beyond the obvious facts that you are a bachelor, a solicitor, a freemason, and an asthmatic, I know nothing whatever about you."

— Sherlock Holmes

(Conan Doyle, A., 1905, "The Adventure of the Norwood Builder", in *The Return of Sherlock Holmes*)

Table of Contents

List of Figures

List of Tables

1 Introduction

Authorship profiling is defined in this dissertation as *the task of determining information about the background of the author of an anonymous text based on the language of the text.* Even though some research has been carried out on authorship profiling, there is currently a demand for more research on *forensic* authorship profiling, which is the application of authorship profiling in the forensic context. The present dissertation aims at meeting this demand through an experimental analysis of texts that are similar to the type of texts that forensic linguists usually examine. The goal of this Chapter is to introduce the present work by contextualising it within the past research. A summary of previous research in authorship profiling and its gaps are presented and then the steps adopted by the present study to address these gaps are outlined.

1.1 The state of the art of authorship profiling in a forensic context

In a typical forensic authorship profiling case, the forensic linguist is asked to identify linguistic markers that can reveal any information to law enforcement about the identity of the author of an anonymous text. These types of cases have been previously defined by Grant (2008: 222) as *single text problems*. Grant (2008) suggests that the only method currently used for authorship profiling is an analysis that is grounded in the linguist's expertise to find the sociolinguistic clues that help establish a profile. One of the most quoted examples of this type of work is Roger Shuy's analysis of a ransom note that successfully pointed to the author's background being an educated male from Akron, Ohio on the basis of a dialect item and a pattern of misspellings (Leonard, 2005). Grant (2008) states that for the present state of the art, single text problems can only be approached using *ad hoc* methods and cites a case in which he was involved that was similarly successful in profiling a man of Jamaican origin based on a dialectal item. This kind of method was initially explained in the non-academic publication of Foster (2001), and Grant (2008: 227) states that this type of method is currently 'the only language based approach which might be applied to single text problems'. Foster's method consists in selecting unusual patterns of words and/or phrases from the anonymous text and then searching for those items in databases, reference corpora or on the internet to identify what other texts the author has been influenced by.

Although this method can be unquestionably useful, as at least two successful cases have demonstrated, Foster's approach to authorship analysis in general has been strongly criticised for lack of objectivity and for its foundations in literary criticism rather than in linguistic science (Chaski, 2001). Foster's method clearly lacks structure and experimental confirmations, as well as being heavily case-specific and primarily based on the linguist's intuition. For the discipline to move forward it is necessary to develop a scientifically grounded methodology that can allow for the drawing of profiles of

anonymous authors using objective tested techniques based on linguistic science. Unfortunately, the research on a systematic methodology that can be used to determine clues as to an anonymous author's general demographics is almost non-existent within forensic linguistics as well as traditional branches of linguistics. Although sociolinguistics has done extensive research on the relationship between linguistic variables and social variables, this type of research has never been applied to authorship profiling. Most of the research available for authorship profiling has been carried out mainly within computer science and psychology.

In computer science, the most significant research has been restricted to studies that involved character and word level features to classify texts by demographics such as gender or age (e.g. Argamon *et al.*, 2009). Even though these methods are promising, computational studies face the problem of lacking theoretical foundations for their findings. These studies generally present a pragmatic solution to a classification problem rather than an understanding of the processes that lead to certain social groups adopting specific linguistic patterns. The problem with sacrificing theory is that it is not possible to know why certain markers are successful whereas others are not and why, therefore, the same markers would work in a new case that is not included in the tested data set. Although slightly more theoretically grounded work has been done when computer scientists have worked together with psychologists (Pennebaker and Stone, 2003; Argamon *et al.*, 2005; Newman *et al.*, 2008), the problem with these studies is typically that they employ a naïve conception of the linguistic variables analysed. Indeed, the most significant problem of all the studies is their failure to account for register variation, even though this type of variation has been proved by many linguistic studies to be the most significant type of linguistic variation. Most of the times, this lack of awareness leads to fallacies in the experiment design that in turn lead to erroneous conclusions. For example, it is common to find studies on authorship profiling in which a number of texts that form the training data set are combined together without controlling for their communicative situations and, consequently, register variation (e.g. Newman *et al*., 2008; Argamon *et al*., 2003). Failure to control for register variation can result in incorrect generalisations being made, because a social difference for a certain linguistic variable in, for example, academic prose, might not hold true in threatening letters. Biber (2012) cites a number of examples of past studies in which measurements on general non-balanced corpora of English produced findings that were initially believed to apply to the English language as a whole but then found to be heavily register-dependent on a second analysis.

The importance of accounting for register variation is due to the fact that there seems to be no doubt that for most of the linguistic variables examined in previous profiling studies there is much greater difference between, for example, a conversation and an academic text than between an academic text written by a male and an academic text written by a female. This point is illustrated by Biber and Conrad (2009), who show how the frequency of contractions varies far more according to register than to social class. Even though not many studies have dealt with the question of how register variation is compared to other types of variation, the validity of this finding is reflected by the fact that generally

the effect sizes found by register variation studies are far larger than the effect sizes found by corpus studies of sociolects. Certain authors have even proposed the idea that social variation for frequency variables is dependent on register variation, since it is likely that social groups have unequal access to registers in society. This idea was elaborated independently but in a similar fashion by both Finegan and Biber (2001) and Hasan (2009) under the names of, respectively, *Register Axiom* and *Semantic Variation* (or *Codal Variation*). In simple terms, both theories propose that social groups produce different frequencies for linguistic forms because they bring to the text their experience with language, which in turn is shaped by their different access to registers. For example, when confronted with writing an academic essay, a person who does not have experience with academic writing but only with conversation is far more likely to bring elements of conversations to it than another person who has had experience of academic writing for many years. In turn, this difference in experiences is reflected in different frequencies of linguistic items that correspond to the registers that these two hypothetical individuals have experienced. Because social groups have different access to registers, the variation that is measured and that is attributed to sociolects is indeed register variation that is skewed by social group. Accounting for register variation when analysing texts is therefore extremely important, even though this is neglected in many past studies.

Another substantial gap in the present research on forensic authorship profiling is the lack of research on those texts that are typically studied by forensic linguists in cases of profiling, such as threatening letters, abusive letters, ransom demands, extortion letters, and similar texts that constitute criminal offences. The present work defines these kinds of text as **malicious texts** and uses the following working definition:

> A malicious text is a text that is a piece of evidence in a forensic case that involves threat, abuse, spread of malicious information or a combination of the above.

These texts are typically analysed by forensic linguists in real cases of extortion, blackmail, ransom, threat, abuse, stalking and so forth. However, for these texts, neither profile analyses nor register variation analyses has ever been applied before. In general, virtually no study has analysed the linguistic and/or extra-linguistic characteristics shared by these texts. Of the several sub-types of malicious texts, only threatening texts have received some attention in the literature, especially from a pragmatic perspective (Fraser, 1998; Solan and Tiersma, 2005; Shuy, 1996). This lack of profiling research of malicious texts combined with the general disregard of register variation constitutes a significant problem, since it is not possible to assume that a particular linguistic variable that is a good discriminator of a social variable in, for example, blogs is also a good discriminator of that same social variable for a threatening letter. The step from blogs or other genres to malicious texts can be made only thanks to both a valid linguistic theory that accounts for register variation and an empirical analysis of malicious texts.

Finally, another missing element in the research in authorship profiling is the lack of a summary of previous research on the most important links between language variation and social variables. A clear picture of the main patterns of linguistic variation for general demographics such as gender, age, or social class would in fact help both the applied research and the theoretical research on profiling. On one hand, applied researchers would benefit from such a review by having a battery of markers that can be run on new data sets. On the other hand, the theoretical researchers would benefit from such a review by using this list of findings to arrive at general theories that explain why these linguistic patterns can reveal the demographics of the author. Without any doubt, better practice in forensic authorship profiling can be developed when both of these dimensions are combined.

1.2 Aims of the present study

The brief review of the situation and state of the art of authorship profiling has identified a number of issues in the current state of the field:

1. The lack of a systematic summary of the relationship between linguistic variation and a range of social variables, including gender, age, level of education and social class;
2. The lack of integration of linguistic theory into current research on authorship profiling and, consequently, a general disregard for the importance of register variation;
3. The lack of research in authorship profiling based directly on malicious texts, such as threatening letter, ransom demands, etc.;
4. The lack of an objective methodology or protocol for authorship profiling in the forensic context.

The present work aims at making substantial steps towards filling in these four gaps. The first step towards this goal is to understand how much is already known about language variation and social variables. The present project therefore starts in Chapter 2 with a survey of the most significant studies from as many disciplines as possible that have found a link between language use and social structure, thereby addressing the first gap outlined above.

The next step of the project is to collect a valid data set that can be used to test whether these linguistic patterns found in the literature review can be used to profile malicious texts. In order to carry out this part of the project, a set of malicious texts of known authorship in compatible registers should be gathered. However, this step is problematic because for many real malicious forensic texts the authorship is unknown. Furthermore, even if this corpus were available, it should be rather large in order to have many texts of compatible registers so that register variation can be accounted for. Additionally, gathering authentic malicious texts is rather difficult because many texts are confidential. To avoid these problems, a data set of fabricated malicious texts produced by a stratified sample of subjects in controlled experimental conditions was created. By controlling the texts that the individuals

produced it is possible to reliably know the details of the authors of the texts as well as to perfectly control the register of the text. This data set is described in Chapter 3.

The advantages of this data set are that the sample is controlled both for the communicative situations in which the participants are writing and for the social characteristics of the participants. Because, however, this data set is fabricated, a drawback of the present study is that these texts are different from real malicious texts and therefore the findings obtained might not be valid for real malicious texts. As a way to compensate for this drawback, another data set was compiled consisting of a corpus of authentic malicious texts that appeared in real forensic cases. This corpus was then compared to the corpus of fabricated malicious texts in order to verify whether the fabricated texts are similar to the authentic texts. The description of the corpus of authentic malicious texts is in Chapter 3 while the comparison between the two data sets is described in Chapter 4.

After the data sets are described and the fabricated texts are validated against real data, Chapter 5 addresses the main research question of the present work, that is, to what extent the relationships found in previous studies between certain linguistic patterns and some general demographics of the author are valid for malicious texts. Since the fabricated corpus is controlled for register, Chapter 5 also addresses the question of determining to what extent the relationship between linguistic variation and social variation is affected by register variation. The scope of the present project is on the four most general and common social variables that can be easily studied on every individual: gender, age, social class and level of education. The study of other important social variables such as geographical origin or ethnicity has been abandoned since it would have required more resources than what is available for the project. The four social variables listed above are from now on referred to as **social factors**.

The last step of this dissertation is to transform the findings of Chapter 5 into a model that can be used by forensic linguists and by future researchers who want to expand upon the present work. To this end, in Chapter 6 the patterns of variation found in Chapter 5 are inserted into predictive regression models that show to what extent and with what reliability these patterns of linguistic variation can be used to profile the social background of the authors of the fabricated malicious texts. This type of analysis not only provides information on the validity of previous studies for malicious texts but also constitutes a first step towards the development of a systematic method for profiling malicious texts.

Finally, Chapter 7 summarises the findings of this study with the goal of providing new hypotheses and directions for future research.

2 Literature review

This Chapter consists in a literature survey for each of the social factors considered for empirical exploration in the present work: gender (Section 2.1), age (Section 2.2), level of education (Section 2.3) and social class (Section 2.4). Each literature survey covers a large number of significant studies that provide evidence for a relationship between one or more linguistic variables and one of the social factors. For each social factor a separate survey of several key studies is presented. The goal of each of the surveys is to find out which links between linguistic patterns and the social factors are established in linguistics as well as in other fields where these links have been studied empirically. A study was reported in this review only when it contributed either theoretically or empirically to the understanding of the relationship between a linguistic pattern and a social factor. Among the studies that were reported, a sub-set was selected for replication in the empirical part of the present work. The studies that were considered for replication are the studies that were conducted using linguistic variables that were calculated in a method explicit enough to be replicated. At the conclusions Sections of each of the social factor surveys, a summary of the main linguistic patterns found is given together with a list of the linguistic variables that constitute them and that will be tested in the empirical part of the present work.

2.1 Literature survey on Gender

Gender is perhaps the most studied social factor in many disciplines that examine the link between language use and the social world. From a very general point of view, at least two kinds of 'gender' can be distinguished: biological gender, which can also call be defined as 'sex', and sociological gender. However, reflections on this differentiation are seldom reported in the literature. Indeed, even this superficial difference is subject to disagreement. For example, in their essay, Bing & Bergvall (1998) point out that not just gender but also sex could be considered as a socially derived categorisation. They quote medical references that underline the fact that sex is a biological continuum that develops thanks to many factors and that the categorisation of sex in a binary way is a cultural phenomenon rather than a natural one. The very fact that people look for how men and women speak differently, or perform differently at maths and so on just reinforces the gender polarisation of these two categories and it fails to acknowledge the reality that sex is a continuum. This opinion is also shared by some branches of modern social psychology. A study by Carothers & Reis (2013) confirms the widely-accepted opinion among behavioural scientists and psychologists that the differences between males and females are indeed of the continuum type rather than of the taxonomy type. Considering *sex* and *gender* as continuums rather than taxonomies implies that an individual of male (female) sex/gender categories does not always consistently present all the behavioural categories of male (female) sex/gender. As Carothers & Reis (2013: 17) suggest:

> 'there are average sex differences for each "symptom" of gender, but they are not consistent or big enough to accurately diagnose group membership […] there are not two distinct genders, but instead there are linear gradations of variables associated with sex, such as masculinity or intimacy, all of which are continuous'. (Carothers & Reis, 2013: 17)

Another challenge to the conception of simple binary classification of gender has been presented by Bamman *et al.* (2012). In this paper, the authors analysed the most frequent words in a corpus of short messages taken from more than 14,000 users of Twitter for gender patterns. The results they obtained confirmed many established findings that are also reviewed below. Bamman *et al.* (2012) also found a certain number of outliers, that is, for example, males using female language. However, instead of treating these outliers as statistical aberrations, the authors decided to explore them more closely. They therefore carried out a second study that consisted in the application of a statistical clustering method to the data with the aim of discerning patterns without the imposition of the gender categories. The result of this second study suggested that the outliers were typically people that within the corpus had more messages sent to and received from people of a gender different from their own. In other words, the individuals that presented a more gendered style were the ones that were more likely to have in their social networks more people of their own gender and that therefore interacted more

often with people of the same gender. Although the researchers clarified that the individuals that more often interacted with people from the opposite gender presented the opposite gender's style, the reasons as to why this relationship was found can be only hypothesised. These 'outliers' can present the above pattern for at least two reasons: (1) they may employ the opposite gender style simply because they interact with the opposite gender more often in that particular genre and therefore learn the language of that genre in the other's gender way; (2) or they employ the opposite gender style because their personal gender orientation, at least for that kind of social interaction, lies in the border and they therefore prefer to interact with the opposite gender more often. Although these possibilities were not explored, the authors could nonetheless conclude that the most useful theory that should be considered when studying language and gender is one that conceives gender as not a dichotomy of male/female but as the product of interaction. Gender, in their opinion, should be considered as an action or construction of the persona that co-varies with other social variables of that same persona. The authors therefore called for more research that considers gender and other social variables as combined rather than in isolation from each other.

All the issues presented above should be accounted in any study on language and gender. For the present literature review, however, the problematic definition of 'gender' is resolved by treating gender as being either the biological male/female distinction or the sociological cluster of behaviours. In fact, the aim of this survey is to assess the state of the art regarding the degree of success in understanding the link between language and gender, independently from which type of gender is investigated. That being the aim, starting from the social sciences and the main theories of how gender should be modelled, this review is concerned with any study in variationist sociolinguistics, corpus linguistics, computational linguistics and computer science that involves an empirical experiment or that propose a theory of the link between language and any type of gender, biological or sociological.

2.1.1 Social sciences and sociolinguistics

One of the ground-breaking and most important pieces of work on gender and language within the social sciences is Lakoff (1973). Although mainly theoretical and based on introspective and anecdotal evidence, this work managed to consolidate a research paradigm that conceives the two genders as two different cultures produced by the process of socialisation. This socialisation argument in turn proposes that society, in the form of parents or peers, exerts pressure on the individual to develop a certain cluster of behaviours and language styles according to the individual's characteristics and it does so mainly by using language itself. Lakoff (1973) proposes that gender is a category which society takes great care to distinguish socially and therefore linguistically as it is based on evident biological factors.

In her discussion, Lakoff (1973) lists a number of linguistic features that she presupposes are characteristic of one or the other gender. In her opinion, female gender lexicon contains a more thorough

taxonomy of adjectives and a 'weak' set of expletive particles. At the syntactic level, Lakoff (1973) mentions a higher frequency of tag questions and polite forms of request for females, such as declaratives with rising intonation. In general, Lakoff (1973) concludes that all these features are not just characteristic of female language but that they signal powerlessness and lack of commitment. It is because females are generally identified with these two qualities in Western culture that they then find themselves adopting the above-mentioned styles of communication.

After the publication, Lakoff's (1973) claims inevitably called for more empirical investigation. One of the first studies carried out within this framework was Crosby & Nyquist's (1977). These authors designed three experiments aimed at testing Lakoff's (1973) proposals, which involved the analysis of speech samples produced by: (1) dyads in conversation; (2) recorded inquires at an information booth; (3) conversations between clients and police personnel. In two of the experiments, Lakoff's (1973) 'female register' features were indeed found to be more common in females than males. However, since the use of these features was also particularly influenced by social role, the researchers concluded that 'female register' is more likely to be the result of social roles rather than gender alone. Indeed, they expand on Lakoff's (1973) hypothesis by proposing that females employ the 'female register' more often because they are often associated with those powerless social roles that often employ it.

A very similar conclusion was reached by a study conducted by O'Barr & Atkins (1980). The researchers started their investigation from the observation that manuals for lawyers had special sections on how to treat female witnesses during a court trial. The entries in these manuals state that sometimes females can be treated differently from males and that certain facets of their behaviours can be used to affect jurors. Incidentally, these manuals' descriptions roughly coincide with Lakoff's (1973) 'female register'. In an empirical exploration, O'Barr & Atkins (1980) analysed the speech of male and female witnesses in court searching for Lakoff's (1973) features. The result of their analysis pointed out that the females as well as the males regarded as belonging to a low social status were equally using these features. The authors therefore hypothesised that the real effect proposed by Lakoff (1973) is actually a correlation between 'female features' and powerlessness in western culture. As conjectured by Lakoff (1973), females were on average scoring higher than males on this 'powerless style' and this is explained by the researchers by the fact that women find themselves more often than men in powerless social situations. According to this study, the phenomenon that is measured by linguistic analysis is not the degree of 'femaleness' but the degree of 'powerlfulness', which is, incidentally, likely to be correlated with gender. O'Barr & Atkins (1980) thus support Lakoff's (1973) hypothesis and Crosby & Nyquist's (1977) experiment. However, their analysis was not replicated in the present study as its methodology was rather subjective and difficult to replicate.

Coming from an analogous perspective, Poole (1979) conducted a series of structured interviews to find out whether gender can be regarded as similar to social class in determining the *code* of an individual. This notion of *code* used by Poole refers to Bernstein's notion of *elaborated/restricted*

code (Bernstein, 1962), which is further explored in the literature review section on social class. Poole's (1979) argument is that if social class can be considered as having a powerful force in socialising an individual in respect to language, then gender socialisation is also likely to have an influence. After collecting elicited spoken data, Poole (1979) ran a discriminant function analysis on a large set of variables previously used to measure codes. She found that one discriminant function distinguished groups by class and that another function separated genders. Poole interpreted these results as giving further credit to Bernstein's hypothesis, as well as providing evidence that the same mechanisms of socialisation that apply to social class might be also valid for gender, thus providing more evidence to support Lakoff's (1973) initial ideas. These findings support another study carried out by Poole (1976) few years before in which the researcher found similar effects in a study of a different data set consisting of 80 first year university students.

Research conducted in Australia by systemic functional linguists, such as Cloran (1989), seems to support the findings above. Her work can be contextualised within the general paradigm of *codal variation* (cf. Section 1.1). This paradigm is based on one hand on the sociology of Bernstein and his notion of code and on the other hand on Halliday's systemic functional linguistics. As also proposed by Lakoff (1973), Cloran (1989) posits that gender can be thought of as one of the possible social groupings that are being installed in the mind of the child during the process of socialisation. Gender is therefore a culturally dependent social construct based on an initial biological distinction of the sexes. As systemic functional linguistics suggests, social constructs like gender or social class are mainly constructed and passed to the new generations through language and they contribute to the personality of the new social individual. However, it is not possible to find a single linguistic component that performs the job of passing social constructs, since the task is spread more generically across language as a whole. More specifically, it is the patterning of certain meanings often produced in a certain context that shapes the minds and it is therefore in the grammar, 'the powerhouse of meaning', that these different clusters of being, behaving and saying are found (Halliday and Matthiessen, 2004: 21).

After reviewing the literature and establishing that male and female children are generally treated differently by their parents, Cloran (1989) set up a study to find out to what extent this is the case. The empirical experiment that she reports is an analysis of speech produced by 24 dyads of mothers-children divided into socioeconomic groups which was recorded during daily activities across nearly a month. The language was coded using a SFL system of semantic options devised by Hasan. The variables were then examined in a principal component analysis followed up by an ANOVA. The results showed that certain linguistic behaviours differentiated mothers of boys from mothers of girls. Specifically, the mothers of boys exhibited a more controlling behaviour and encoded points of view less often. In general, Cloran's (1989) findings were in line with previous research and confirmed what the authors expected to find. Cloran's (1989) study could not be included in the present work as no thorough explanation of the coding system is given by the authors.

Following a similar paradigm, Mulac & Lundell (1994) conducted a study to find out whether culturally stereotyped gender styles appear even in writing and whether there is a conscious perception of these styles in lay people's perception. They were concerned with writing as they assumed that written language is less susceptible to large stylistic variation because of the standardisation effects that individuals generally receive from formal education. In this study, the researchers recruited 148 students from a university in California and asked them to write a description for each of two landscape pictures that were projected in a class. They then randomly selected the essays written by 20 males and 20 females within a17-25 age span.

The first part of the study consisted in asking a set of judges to assign a gender to each of these anonymised essays. The result of this study seemed to show that lay people could not guess the gender of the writer better than chance. However, when these same judges were asked to rate the personality of the writer, they gave scores on dimensions such as *socio-intellectual status* or *aesthetic quality* or *dynamism* that were gendered and that mirrored what the authors expected based on cultural stereotypes regarding gender.

In the second part of the study, Mulac & Lundell (1994) studied the writings of these subjects in terms of style. Nine coders were trained to analyse the texts for a set of variables that were judged by the authors as being good discriminators of gender according to previous studies. The variables were analysed using a discriminant function analysis, which successfully classified the texts with a reclassification accuracy of 75%. The features that were found to distinguish males and females were consistent with previous research. Males showed more usage of terms that refer to quantity or locations whereas females used more uncertainty verbs (*it seems to be…*) and references to emotions. Although the authors stated clearly that these variables separate groups with a significant overlap, they pointed out that a difference does exist. Furthermore, the authors found a significant correlation between these two clusters of gendered features and the personality dimensions that were introduced above. In other words, for example, texts that present more 'female' features are more likely to receive a high score in personality dimensions typically associated with the female gender, such as *socio-intellectual status*. This is interpreted by the authors as a suggestion that the two linguistic styles are primary correlated with different personalities and behaviour patterns and only secondarily correlated with gender. This study could not be considered for replication for the present study as not enough information was given about the variables that the authors used.

The theoretical explanation for the differences found was later on re-addressed in Mulac *et al.* (2001). The authors claimed that the studies so far have shown that genders are sub-cultures and that gender is a 'social system that reinforces behavioural expectations for group members' (Mulac *et al.*, 2001: 122). Since the difference is in the behaviour, the differences between the two groups do not originate from different repertoires of linguistic structures but from the different ways of employing these resources in the same context. In the author's words:

> 'Boys and girls (as well as men and women) may share a common vocabulary but use that vocabulary in dissimilar ways. For example, both men and women may know a wide variety of terms for referring to emotional states, but women may be more likely to produce these terms in interpersonal communication' (Mulac *et al.*, 2001: 122).

This theoretical position coincides with Mulac's & Lundell's (1994) conclusions on the relationship between personality, behaviour patterns and gender.

Most of the social scientists whose studies were reported above worked with theories that explain the language-gender connection in a similar fashion. All influenced by Lakoff's (1973) essay, the empirical explorations carried out by these researchers show that at least some evidence in support of Lakoff's (1973) hypothesis exists. However, an important criticism that should be pointed out consists in the fact that the theory of socialisation seems to take into account only the cognitive-behavioural aspect of the construction of personality. There is enough evidence to suggest that dimensions of personality are not just the result of social learning but also a function of genes (Gleitman *et al*., 2004). In other words, the aspects of behaviour that distinguish females or males can be due to social learning and socialisation but also to inner traits of personality influenced by genetic factors. In this regard, Cloran (1989) specifies that the inner temperament of the child interacts with the social background and is managed by the parents according to social rules. A more accurate theory of socialisation should therefore presuppose that an initial temperament is altered according to which label society has prepared for the individual during the socialisation process.

2.1.1.1 Labovian sociolinguistics

In quantitative Labovian sociolinguistics, the understanding of the connection between variation of linguistic variables and gender has been one of the main concerns. Mainly focusing on phonetic variation, sociolinguists have extensively studied the differences between males and females in several societies and cultures. A comprehensive and theoretical review of the research was produced by Chambers (1992). In this paper, Chambers tries to find a consistent and comprehensive theory that would explain why in almost every sociolinguistic study females tend to present fewer non-standard features and stigmatised forms. In a similar fashion to the introduction to Section 2.1, Chambers (1992) starts his argument by pointing out that *gender* and *sex* should be distinguished, with *gender* referring to the social construct, as in the hypotheses put forward by Lakoff (1973), and *sex* defining the human biological parameter. Even though those two dimensions often coincide, the author points out that it is very important to understand the real nature of the independent variable. Chambers' (1992) idea is that there are two sources of variation: *gender-based variation* and *sex-based variation*. The former kind is dependent on the access that one of the genders has to social roles, social mobility and social networks in a particular culture. The gender that has more access will develop a well-rounded sociolinguistic competence and a more comprehensive repertoire of variants. This will allow them to benefit from the

advantage of adapting the speech to the situation. In our society, because of the socialisation that they receive, it is females who develop the highest levels of competence. On the other hand, in other societies, where the gender roles are reversed, it can be exactly the opposite.

However, Chambers (1992) also acknowledges that in modern western societies this gender-based variation hypothesis fails to predict the behaviour of middle-class individuals. For this social group it is still possible to find gender differences even though the social mobility is rather similar. Furthermore, Chambers (1992) adds that another point that should be covered by a valid sociolinguistic theory is the vast body of psychological and neuroscientific research on the differences between males and females. According to this research, the lack of lateralisation of the faculty of language in females gives them a clear advantage in verbal skills. To reconcile these positions, Chambers (1992) proposes a second hypothesis, the *sex-based variation*, which predicts that women present a higher level of sociolinguistic competence given by their tendency to be superior to men in terms of verbal skills. Chambers (1992) is however careful in pointing out that the magnitude of these differences is small and that therefore the individual differences often overcome the group differences.

Together with Chambers' (1992), a similar comprehensive review of the sociolinguistic research is Wodak & Benke (1998). These authors conclude that most of the studies seem to show that women use more standard forms than men for phonetic variables and that this finding applies across languages and cultures. However, as opposed to Chambers (1992), the authors are not too optimistic on the status of the theoretical explanations of this finding. Even though the cultural explanation appears to be more sensible than the biological explanation, they insist that it is still a naïve way of explaining a complex phenomenon like language behaviour. Wodak & Benke (1998) call for an explanation that can take into account the social context and group ideologies.

This position is also maintained by Eckert & McConnell-Ginet (1992). These authors raise the point that much of the research in social science and sociolinguistics starts from the underlying ideology that sex or gender are variables that are added up to the picture. Quite often sociolinguists classified individuals as being, for example, middle class plus male plus forty years old. The authors' position is that it is the *community of practice* rather than the addition of these classifiers that influences the linguistic forms that are acquired and therefore used in a particular context. The differences between the genders that are typically found are more likely to be a result of the fact that these two groups engage in different activities and are exposed to different varieties, for example, 'men are more likely than women to be members of football teams, armies, and boards of directors […] women are more likely to be members of secretarial pools, aerobics classes, and consciousness raising groups' (Eckert & McConnell-Ginet, 1992: 472). This argument shows how simplistic a pure variationist research can be at times and raises the point that gender is construed in the interaction. It is implied therefore that real linguistic differences can be only found in the complexity of these interactions. An objection to this argument is, however, that it is entirely possible that the direction of the correlation is inverted. Since men and women present on average different orientations, preferences and perhaps language, then they

are more likely to enjoy different communities of practice. The solution to this problem is to be found in controlled experimental empirical research and in replications on many data sets.

It is important to note that the research so far in variationist sociolinguistics focused almost entirely on a phonological, phonetic or syntactic variable for which it was possible to establish different ways of saying the same thing in accordance with the Labovian sociolinguistic framework. Little work has been done on relative frequencies of linguistic features. These forms have not been studied traditionally and other disciplines like computational linguistics and corpus linguistics are only now exploring their correlations with social dimensions. However, as it is shown in the sections below, traditional sociolinguistics' position regarding the avoidance of frequency variables can to some extent be challenged.

2.1.2 Corpus linguistics

As opposed to the general sociolinguistic approach, corpus linguistics studies reviewed in this section focused more extensively in the analysis of frequency of linguistic features in large corpora of naturally occurring texts.

A small but significant first contribution to the study of gender within corpus linguistics was Biber *et al.* (1998). In a section of this book, the authors studied a corpus of letters written by males and females across time to verify empirically whether certain claims relating to gender styles are accurate. Their corpus consisted of 276 personal letters distributed across centuries and grouped according to the recipient (F to M, F to F, M to F and M to M). The first claim that the authors tested was the alleged high frequency of emphatics in female writings. Although there was no mention of statistical testing, the authors concluded that women did use more emphatic forms than men, in particular in the 20th century. It is worth noting, however, that even though the total corpus consisted of 276 letters, for the 20th century group only 14 letters were written by females.

The second test that they carried out consisted in verifying whether the variable Dimension 1 successfully separated the two gender groups. Dimension 1 is a linguistic variable found by Biber (1988) as a result of a factor analysis adopted to study the variation in speaking and writing in the English language. This variable has two poles: (1) a high score corresponds to the Involved pole, which includes linguistic features such as mental verbs, pronouns, demonstratives and sentence relatives; (2) a low score corresponds to the Informational pole, which includes linguistic features such as nouns, high average word length, adjectives and prepositions. When analysing the corpus of letters using this variable, Biber *et al.* (1998) found that women were more Involved than men and, consequently, men were more Informational than women. The researchers also found a particular accommodation effect, since males tended to be more Involved when writing to females than when writing to other males.

That Dimension 1 seems to be a good discriminator of gender is also supported by Heylighen & Dewaele (1999). In an attempt to reconcile several findings of sociolinguistics, these researchers

suggested a new way of measuring a concept that they define as *deep formality* that indeed overlaps with Biber's (1988) Dimension 1. The authors define deep formality as '[the] attention to the form for the sake of unequivocal understanding of the precise meaning of the expression' (Heylighen & Dewaele, 1999: 3). They argue that deep formality is similar to the concept of 'formal' in mathematics, that is, 'context-independent' and 'non-fuzzy'. Firstly, the authors review a number of studies showing that deep formality is pervasive and possibly the main source of linguistic variation in a large number of languages, if not even universally. Secondly, the authors move to an assessment of how deep formality correlates with situational and personality dimensions. In their empirical studies, the authors found that there is a difference in deep formality between males and females, with females being less formal than males, and found this difference to apply even across languages. However, the researchers found that this difference disappeared in formal writings, such as essays, as the situational aspect of these genres was a better predictor of deep formality than the gender of the authors. For an explanation of this phenomenon, the authors cite psychological and sociolinguistic evidence of women being generally more Involved in conversations as opposed to men being more Informational. The authors also propose cognitive explanations by citing research that supports the hypothesis that women and men tend to present different cognitive orientations on average. Although still a speculation for the authors, a summary of the justifications for such differences given by them is the following:

> 'women would be more sensitive to the immediate social and physical context, whereas men would tend to consider problems in a more detached way […] this would explain women's involvement in the social context of a conversation, and the concurrent reduction of deep formality in their speech' (Heylighen & Dewaele, 1999: 30).

The findings that Heylighen & Dewaele (1999) present in their report are compatible with the rest of the literature review and they indeed offer a comprehensive reassessment of the picture. The value of their report is to provide a theory that explains a significant finding of modern corpus linguistics: the seemingly universal opposition between Informational against Involved/Interactional features represented by Biber's (1988) Dimension 1. However, although the adoption of the concept of deep formality to explain Dimension 1 would be tempting, there is not enough empirical research to accept many of the claims put forward by Heylighen & Dewaele (1999). In this study, therefore, Dimension 1 and deep formality, as well as all the other similar variables, are treated separately.

Similar results related to Dimension 1 and/or deep formality are found again by Schmid (2003). The researcher carried out an analysis of the spoken texts in the British National Corpus (BNC), aiming at verifying whether the two genders do live in different cultures, as per Lakoff's (1973) hypothesis. The methodology consisted in obtaining the relative frequencies of a number of words that the author or previous literature considered as being 'gendered'. As anticipated above, the results are consistent with the findings obtained by other authors, with females using more hedges, colour words and temporal

adverbs and few abstract nouns if compared with males. Furthermore, in a final remark, Schmid (2003: 218) concludes that

> 'much more than men, women seem to be engaged […] in what is usually regarded as prototypical spontaneous speech […] [that is,] high involvement in the interaction and little spatial, temporal and emotional distance between the speech participants'. (Schmid, 2003: 218)

This, the author continues, is found also in the type and frequency of parts of speech that are used by one or the other gender, with males using a more nominal style than females.

This general Involved-Informational dichotomy and its correlation with genders already pointed out by other researchers was found again in a recent study by Saily *et al.* (2011). The authors analysed a part-of-speech tagged version of the Corpus of Early English Correspondence consisting of more than two million words of personal letters produced by about 660 writers of both genders. Although designed for historical linguists, the corpus was nonetheless suitable to a sociolinguistic investigation on genre and gender variation regarding the frequency of nouns and pronouns.

Saily *et al.* (2011) confirmed in their study that males used more nouns and fewer pronouns than females and that this finding was statistically significant for every span of time between 1415 and 1681 except for one year. Consistently with Biber *et al.* (1998), independently of the gender of the writer, the personal letters addressed to males were found to use a higher frequency of nouns than the personal letters addressed to females, although this finding could not be tested for statistical significance.

The research presented by Saily *et al.* (2011) confirms once again that males tend to use a more Informational or nominal style, whereas females tend to use a more Involved, pronominally rich style. However, as the researchers admit, the analysis was not controlled for the writers' social class and level of education, their relationship with the sender and topic of the letter and this striking difference found could therefore be at least influenced by these and other social parameters. The fact that letters received by males consistently showed a more nominal style could confirm the fact that the accommodation between interactants in language interaction is rather significant even for writing and that therefore it is important to take into account the recipient of the communication.

Within corpus linguistics, a slightly different area of research which was included in this review as regarded to be significant regarding threatening, abusive or malicious texts is the study of swear words. The seminal work in this area is McEnery (2006), whose comprehensive research not only included a history of swearing in British English but also a theory-supported investigation on how swear words are used in contemporary society by different social groups. For his analysis, McEnery (2006) used a sub-corpus of the BNC comprising ten million words of transcribed speech equally distributed across age, sex and social class of the speakers. When analysing the corpus for correlations between swear words and sex, the author found no significant difference between males and females in the

general frequency of swearing. However, when examining which types of swear words the two categories used, McEnery (2006) did find a significant difference in terms of the intensity or type of swear words used. Based on the observations of the corpus and on a scale of offence elaborated by the British Board of Film Classification, it seemed that males tend to use stronger swear words more often and females tend to use weaker swear words more often. Furthermore, the investigation of the grammatical categories in which these swear words appear showed that males and females significantly vary in the way they employ their swear words, with males using more adverbial boosters ('fucking awesome') and emphatic adverbs/adjectives ('he fucking did it') and females using more general expletives ('oh, fuck!'), premodifying intensifying negative adjectives ('the fucking idiot') and idiomatic set phrases ('give a fuck') (McEnery 2006: 27). However, as McEnery (2006) notes, this difference can be due to the fact that the strongest swear words show a tendency to appear in the adverbial boosters and emphatic adverb/adjective categories independently of sex. That being so, although it can be concluded that a significant difference between the sexes is found in terms of the intensity of swear words and grammatical category, it is not possible to conclude whether the real correlation is between sex and intensity or sex and grammatical category.

An interesting exploration that McEnery (2006) conducts is on the interaction between the sex of the speaker, the swear words used and the sex of the hearer. When considering same-sex dyads of speaker and hearer, swearing happens more frequently than in cases in which the dyad consists of two individuals of different sexes. A more thorough analysis revealed that indeed the different combinations of dyads (F-F, M-F, F-M, M-M) show preferences for the use of certain swear words and avoidance of other swear words. For example, words such as *bloody*, *bitch* or *cow* are more likely to be directed at females by females, whereas words such as *fucking*, *gay* or *arsehole* are more likely to be directed at males by males (McEnery, 2006: 33). In general, it appears that females are less likely to hear swear words in language spoken to them and, quite significantly, more less likely to hear strong swear words. Since males/females produce stronger/weaker swear words but also hear and are targeted by stronger/weaker swear words, the conclusion that McEnery (2006) draws is that these two phenomena are likely to be linked. McEnery (2006) also studied the interaction of sex, age and social class and the use of swear words. However, this discussion will be the object of the following sections on the respective social variables.

McEnery (2006) not only examined how swearing is stratified across social parameters but also tried to explain the variation observed in terms of social theory. The author describes the gender differences found as an effect of Bourdieu's 'theory of distinction'. As the author suggests, Bourdieu's theory could be summarised in one claim: 'features of culture are used to discriminate between groups in society, establishing a social hierarchy based on a series of social shibboleths' (McEnery, 2006: 9). Applied to the language of swearing, this model supports the hypothesis that swear words are a symbol that conveys social information. Since the seventeenth century, the middle class started to distinguish themselves from the other social classes through the avoidance of swear words, which in turn slowly

became associated with the lower classes (McEnery, 2006: 11). As McEnery (2006) suggests, after years of hierarchical associations with certain *overt* and *covert* prestige, the lack or presence of swearing is used as a symbol to project an identity in society. The connection between variation in the use of swear words and social categories such as gender according to this theory is therefore one of social distinction. Males and females vary in the kinds of swear words used because they choose to use or avoid to use certain linguistic symbols of social hierarchy that are connected to the prestige with which they are willing to identify themselves.

McEnery's (2006) findings are however in contrast with the findings of Rayson *et al.* (1997), who showed that swear words were more common in male speech in their 10 million word sub-corpus of spoken language of the BNC. Incidentally, Rayson *et al.* (1997) also confirmed previous findings related to the more Informational character of male speech by noticing that male speakers were more likely to use numbers, determiners and the preposition *of* as opposed to the more Involved character of female speech, characterised by more pronouns. Rayson *et al.* (1997: 6) concluded that their data 'bear out the hypothesis that male speech is more factual and concerned with reporting information, whereas female speech is more interactive and concerned with establishing and maintaining relationships'. A tendency was noted by Rayson *et al.* (1997) for women to use more proper nouns in general and specifically to refer to people. Males were found instead to use proper nouns to refer to places.

The conclusions generated by the studies reviewed in this section were confirmed in larger data sets by computer scientists and computational linguists, as outlined in the Section below.

2.1.3 Computer science and computational linguistics

The computational field has recently produced many studies that try to categorise texts according to the gender of the author. However, these studies generally do not propose explanations for the differences and are therefore just limited to finding correlations. Although their lack of explanations results in a lack of the necessary validity that would allow a generalisation to other cases, the computational studies are useful as they have as their own advantage the large size of their data sets and the sophistication of techniques.

For example, Koppel *et al.* (2002) analysed some texts from the BNC using machine learning algorithms using as variables 405 common function words, all the parts of speech and the most common part-of-speech two-grams and three-grams. In their argumentation on the selection of the variables, the authors claim that these features are independent of content and therefore they are more likely to capture the style of the authors. On a corpus constituted by 566 fiction and non-fiction documents from the BNC equally distributed in gender groups, the machine learning algorithm performed much better in classifying the texts by genre rather than by gender. This was a predictable result, given that it is widely established that the features that the authors have selected significantly vary with genre (Biber, 1988). When controlling for genre, the classification algorithm for gender achieved about 80% accuracy. The

features that the algorithm isolated as being more distinctive of one gender or the other are generally consistent with the Dimension 1 polarisation described above.

In a similar study, Argamon *et al*. (2003) analysed an analogous sample taken from the BNC consisting of fiction and non-fiction documents. Their aim was to verify whether gender differences could be found even in formal written documents. Their corpus included 604 texts equally divided by genre and controlled for authorial origin for a total size of 25 million words. Their analysis consisted in a frequency count of basic and most frequent function words, part-of-speech tags and part-of-speech two-grams and three-grams. The counts were processed by a machine learning algorithm aimed at classifying the texts by author gender. The results were compatible with Koppel's *et al*. (2002) findings, as their algorithm obtained an accuracy of 80%. In terms of features, the results were consistent with Biber's *et al*. (1998) proposal: males were more Informational, thus using more determiners and prepositions, whereas females were more Involved, thus using more pronouns. Moreover, the authors also measured other variables used in Biber's (1988) Dimension 1 obtaining a perfect compatibility with the hypothesis put forward by Biber *et al*. (1998). Another result of the study that is compatible with Koppel's *et al.* (2002) work is that these variables do not just distinguish males and females but also fiction and non-fiction texts. This finding can be explained by the fact that these variables load significantly on Dimension 1 and that this factor has been shown to account for a significant amount of variance in English texts as well as in other languages (Biber, 1995).

Other studies support the hypothesis that these results hold even for informal genres. Schler *et al.* (2006) obtained results that are compatible with Koppel *et al.* (2002) and Argamon *et al.* (2003) in a more vast corpus of 300 million words of blogs stratified for age and gender of authors. As shown previously, even for blogs on average females presented a higher frequency of pronouns and negations, whereas males presented a higher frequency of determiners and prepositions.

Newman *et al.* (2008) found similar results after examining another large sample of texts. The authors collected data sets from previous studies and then analysed them with their own computer program, the Linguistic Inquiry and Word Count (LIWC). This program classifies the words of a text for their grammatical and psychological meaning by looking at its own internal dictionary. Although rather linguistically naïve, the program has been shown to produce consistent results if the data set is sufficiently large to correct the inevitable mistakes that the program produces when tagging a text. In Newman's *et al.* (2008) study, LIWC was used to analyse a collection of texts that included: written texts composed as part of psychology experiments coming from universities in the US, New Zealand and England (in the form of: stream of consciousness, diaries, short essays, free responses to questions or description of images); full texts of fictional novels; essays written for university evaluation; and transcribed free conversations from research interviews. In total, the data set amounted to almost 46 million words produced by 11,609 participants. Using a MANOVA the authors concluded that some of their LIWC variables were statistically significant and showed a gender effect. However, the effect size for many of the significant variables was considerably low. Interestingly, the highest effect sizes were

found for the features that have been extensively noticed to vary between males and females, that is: frequency of pronouns, frequency of prepositions, long words and frequency of articles.

When controlling for context of production, the authors found that in conversation the effect sizes were higher than in other contexts. Consistently with the findings of other studies, the authors concluded that the reason for this difference can be explained by the fact that the less formal the context the greater the opportunity for the person to express themselves and to select the topics and style they prefer. Overall, the authors conclude, the study confirms many if not most of the studies conducted recently.

Although the findings in this section all point to the same direction, a strong criticism of certain assumptions of all these studies was suggested by Herring & Paolillo (2006). As most of the computational studies are almost entirely linguistic theory free, they inevitably fail to account for some of the basic understanding of how language works, such as the fact that genre differences significantly influence language variation. Based on previous empirical and theoretical literature on gender and language, Herring & Paolillo (2006) developed a study aimed at verifying two hypotheses: (1) whether males and females write blogs posts differently; and (2) whether, in general, authors of the blog sub-genre 'diary', whose purpose is to report and comment on their life events, write differently than authors of the blog sub-genre 'filter', whose purpose is to report and comment on events external to their life. More specifically, the authors tried to replicate the findings presented in Koppel *et al.* (2002) and similar studies in which males are found to present more Informational features and females more Involved features. Using the sets of words provided by Koppel *et al.* (2002), the authors examined 127 blog posts equally divided by the two genders and the two sub-genres 'diary' and 'filter' and concluded that the main effect found was related to genre, rather than gender.

The relative frequencies of the words found in Koppel's *et al.* (2002) study to distinguish between males and females were actually distinguishing between 'diary' blogs and 'filter' blogs in Herring & Paolillo's (2006) study. Since, however, these two sub-genres of blogs were skewed regarding the gender that preferred them, the conclusion to be drawn is that females use more 'female words' because they are actually writing more within their preferred blog sub-genre of 'diary' and that, therefore, males use more 'male words' because they are actually writing more within their preferred blog sub-genre of 'filter'. This explanation is compatible with the present knowledge available on linguistic variation and the pervasiveness of genre and also in line with theoretical research in sociology and sociolinguistics, for which a division between females and orientation towards relationships and people on one side and males and orientation towards objects and information on the other is postulated (Herring & Paolillo, 2006: 3). It is also the authors' opinion that if Koppel *et al.* (2002) were to employ a more thorough sub-genre division of their data set they would eventually find that their gender effect would be indeed a genre effect. This study has the merit of suggesting a linguistically motivated understanding of the reason as to why differences between males and females are found.

2.1.4 Some considerations from neuroscience and psychology

In this Section, the focus of the survey moves towards the main findings obtained by neuroscience and psychology. The reason for this change of field of study lies in the attempt to create a picture of the situation that is as wide and comprehensive as possible. This Section thus reports some of the key review papers that outline the most significant findings on the relationship between gender and the mind.

Within neuroscience, males-females differences in the brain have been extensively studied for years. In a review of the research, Springer & Deutsch (1997) summarised the findings by stating that females are on average better at performing verbal tasks, whereas males are better at visual-spatial tasks. These conclusions were drawn from a large collection of experiments that involved males and females across several years. The tasks that involve language proficiency (e.g. verbal fluency, speed of articulation, use of grammar) see females scoring higher than men. On the other hand, men achieve better results in tasks that require visual-spatial abilities (e.g. maze performance, picture assembly, mental rotation). According to the review, this has to do with the fact that the hemispheres of the brain are on average organised differently in males and females.

However, in the quest of understanding the nature of these differences, since the analysis of brain activity, composition and chemistry are inconclusive, some researchers have hypothesised that the differences are indeed given by socialisation effects (Kaiser *et al.,* 2009). This explanation can also explain why there are inconsistencies in the findings, as the independent variable might not be biological sex but sociological gender. Whatever the explanation of the patterns is, however, researchers agree on the fact that these differences are around one fourth of a standard deviation in magnitude and that therefore there is a very significant overlap between the two genders (Springer & Deutsch, 1997; Cosgrove *et al.,* 2007). The differences can be found only when large samples are compared, as individual variation is usually considerably influential.

A further point raised by research in neuroscience is that these differences might actually be connected with hormone levels. Male and female hormones' role in affecting the male or female patterns of behaviour in mammals has already been thoroughly established. However, research hints towards the possibility that the different hormones also correlate with different cognitive orientations. For example, some studies found that female individuals that for some reason received higher levels of testosterone in early years tend to produce higher scores in visual-spatial abilities tasks. Other experiments show that females perform better at verbal ability tasks in the phases of their age where there is a higher level of oestrogens and progesterone. The correlation therefore seems to be between levels of testosterone and visual-spatial skills on one side and levels of progesterone or oestrogens and verbal skills on the other.

However, this argument appears to be rather controversial. Some studies that showed these correlations failed to be replicated. In a more recent review of the research on the influence of hormones

on cognitive tasks, Halari *et al.* (2005) concluded that, although differences in the averages in the performances in cognitive tasks between males and females are evident, these seem not to be dependent on hormone levels. The conclusion on this topic is that the effects of hormones on cognition are unknown, with certain studies finding effects and other studies failing to replicate these effects. There are many possible explanations for these contradictory outcomes and, even though the general agreement is that there is some influence, at the present stage the extent of this influence is still unknown.

For the purposes of the present study, it is interesting to note that there is some weak indication that testosterone can influence Dimension 1 features. Using LIWC, Pennebaker *et al.* (2004) studied how injections of testosterone influenced the writing style of two individuals. The researchers found that the injections of testosterone were significantly correlated with a decrease in the rate of usage of pronouns and other socially relevant categories of words, such as feeling and communication verbs. These features, as indicated by the review of the literature above, are part of the Dimension 1 Involved pole and have been generally found to be indicative of female gender in many experiments. Even though these findings have not been replicated yet and are hard to generalise due to the small sample used, the results do point to the predicted direction.

Within psychology, researchers have investigated the relationship between the use of first person pronouns and depression. This link is given by the fact that depressed people tend to act a strategy of rumination about themselves more often than non-depressed people and tend therefore to use more first person pronouns (Rude *et al.*, 2004; Pennebaker, *et al.*, 2003). This strategy of rumination was found to be more often employed by females than males in a study by Fast & Funder (2010). In this study the researchers linked the use of first person pronouns to tendency to depression for females and tendency to narcissism for males. They also noted that in general it was females who used significantly more first person pronouns.

This body of research adds another perspective to the wider picture. If the use of personal pronouns is linked to rumination, depression and/or narcissism or even testosterone levels, then it is possible that the measurement of frequency of pronouns is only secondarily and incidentally linked to gender. If this link is found to be valid, it could provide an explanation as to why the differences in other studies explored in this review found limited effect sizes in the linguistic variables between males and females. A hypothesis that could be formulated is that the real differences in the linguistic patterns adopted by people depend on their personality and/or hormone levels and that genders are different to the extent that on average different genders are prone to different personality orientations and/or hormone levels.

2.1.5 Discussion

This literature survey has pointed out several perspectives. In general, however, all the studies considered for empirical exploration for the present work can be summarised under three main linguistic patterns of variation. This classification is presented in Table 2-1 below.

Table 2-1 – Summary of the studies reviewed for gender. Variables in bold represent variables that increase if the gender is female whereas underlined variables are variables that increase if the gender is male.

Study	Genre of data	N of participants	Average or min-max text length	Year of data	Summary of linguistic variables	Country
Pattern 1: Rapport/report orientations On average, males prefer a nominal *report* discourse orientation whereas females prefer a clausal/deictical *rapport* discourse orientation. These two orientations could be due to socialisation effects or average biological differences in brain organisation or a combination of both effects.						
Poole (1979)	Structured interviews	96	N/A	1979	**uncommon adjectives/ adjectives; personal pronouns/ total words; I/total words; total adverbs; I/total personal pronouns; *I think*/total words; automatisms**; <u>total adjectives/total words; unusual adverbs/adverbs; total prepositions/total words; proportion of *of*/*in* and *into*; language mazes; *ah*-disturbances/verbal tics</u>	Australia
Rayson *et al.* (1997)	Casual conversation	N/A	N/A	1990	Frequency of **pronouns; proper nouns;** <u>numbers; determiners; the preposition *of*</u>	UK
Biber *et al*. (1998)	Personal letters	80	412	1900-1990	Frequency of **emphatics**; Dimension 1 score (low for males, high for females)	UK/America

Heylighen & Dewaele (1999)	Speech	80	N/A	1979	(Deep) Formality (high for males, low for females)	Netherlands, UK
Koppel *et al*. (2002)	Formal non-fiction writings	N/A	34,320	1960-1974	Frequency of: **pronouns; *and*; *for* and *with*** determiners; prepositions	UK
Argamon *et al.* (2003)	Formal written documents	N/A	42,000	1990	Frequencies of: **personal pronouns, negative particles; contractions; present tense verbs**; *its*; determiners; Determiners/Nouns; Attributive Adjectives; Noun-*of*; prepositions; long words	UK
Newman *et al*. (2008)	Conversations and general written documents including: stream of consciousness essays, essays about emotions, published books	11,609	300 - 8000	1980-2002 (books from the 1800)	Frequency of: **social words, pronouns, third person pronouns**; words longer than six letters; articles	England, New Zealand, USA
Saily *et al.* (2011)	Personal letters	660	200 - 2000	1415-1681	Nouns:**pronouns** ratio	England
Pattern 2: Distribution of expletives Males and females on average tend to produce language in different ways in order to show their affiliation and/or their detachment with certain values or social groups. For gender, a point of distinction lies in the use of swear words.						
Rayson *et al.* (1997)	Casual conversation	N/A	N/A	1990	Frequency of swear words	UK

McEnery (2006)	Conversation	N/A	N/A	1990	Frequency of: Strong SWs; Very Strong SWs; **Mild SWs; Very Mild SWs**	UK
Newman *et al*. (2008)	Conversations and general written documents including: stream of consciousness essays, essays about emotions, published books	11,609	300 - 8000	1980-2002 (books from the 1800)	Frequency of swear words	England, New Zealand, USA
Pattern 3: Powerless register Males and females on average present a different distribution of power in society, with males being in powerful positions more often than females. Therefore, since individuals have different familiarities with the powerless registers depending on their sex, the language that an individual produces will manifest powerless elements depending on their sex.						
Crosby & Nyquist (1977)	Recorded conversation in laboratory; Recorded spoken requests for information	122	N/A	1977	Frequency of: **empty adjectives; tag questions; hedges; politeness forms**	USA

All the studies gathered under the first pattern, the rapport/report orientations pattern, agree in finding the same pattern across many data sets, registers and time spans: on average, females tend to use more pronouns, verbs and deictic features whereas males tend to use more nouns, adjectives and other nominal and informational features. This opposition could be summarised by Biber's (1988) Dimension 1 Involved/Informational poles, with females on average producing more Involved discourse and males on average producing more Informational discourse. Even though not all the studies mention an explanation for this pattern, several propose that socialisation patterns in modern western society tend to separate individuals according to their sex to different roles which then in turn shape their language. The rapport/report roles are therefore responsible for the rapport/report discourse orientations that the two genders on average manifest. Few reviewed studies in psychology and neuroscience have also found that characteristics such as tendency to depression or hormone levels influence to some extent cognitive processes, behaviour, and potentially language use towards effects that are similar to the ones found for gender. If this effect is indeed larger than the gender effect, then it might be the case that the rapport/report distinction is primarily correlated to personality and only secondarily to gender. Until further experiments are carried out, however, it is not possible to establish which independent variable is the primary one.

The second pattern concerns a difference in expletives use that has been observed in three studies. McEnery (2006) proposes that this difference could be linked to *Bourdieu's theory of distinction*, according to which social groups tend to develop linguistic ways to differentiate themselves from other social groups or social values that they do not want to be associated with.

Finally, the third pattern is represented by only one study, Crosby and Nyquist (1977), who propose that average female language production is characterised by politeness features and that these differences in degree of politeness are dependent on the unequal distribution of power in society.

In conclusion, the review of the research on language use and gender generated three patterns of linguistic variation correlated with gender. In Chapter 5, the presence or absence of these patterns is tested in the data set collected for the present study to understand to what extent these linguistic patterns are found in texts resembling a forensic scenario. The present Chapter moves to the review of the research of language variation with age.

2.2 Literature survey on Age

This survey covers the most significant findings related to the link between language variation and age. Similarly to gender, the concept of 'age' has not been problematized by many studies. At least two dimensions of 'age' could be said to exist: a biological/chronological age and a social age. Eckert (1998), for example, criticises researchers that often report only the chronological age of the participants of their studies and thus consider this aspect of age as the most significant. Eckert (1998) claims that 'social age', for which chronological age is just an approximation, is likely to be more significant in determining linguistic variation than the mere chronological dimension. Socially significant events in an individual's life such as certain birthdays (sixteen or eighteen), religious status changes (bar and bat mitzvah), changes in institutional/family/legal statuses (marriage, retirement, naturalisation) tend to affect the socialisation patterns and therefore the acquisition and production of linguistic variables (Eckert, 1998: 156). Furthermore, Eckert (1998: 157) suggests the possibility that social age is a vehicle through which other social variables, such as gender or social class, are conveyed. For example, important social landmarks in society are different according to which gender class the individual belongs to. Similarly, although individuals starts their job only when they reach a certain age, this age threshold seems to be lower for the working class people, who therefore have an earlier access to the linguistic influence of their co-workers' registers and styles.

In a similar fashion to gender, this literature review is aimed at being comprehensive and will therefore include any study that looked at the relationship between language and age, independently on how age was calculated. As opposed to gender, however, age as a dependent variable has not been thoroughly investigated in linguistics research, with some exceptions in sociolinguistics where it was mostly alternation variables that were investigated. Most of the research related to age in other branches of linguistics focuses on language acquisition at early stages of development. Only few studies are dedicated to the examination of how language varies throughout all the phases of life. Interesting findings in this area were produced by researchers in psychology and psycholinguistics. In particular, psycholinguistics has been rather prolific in expanding their already consistent body of research on language development and language impairment to the decline of language faculties in later adulthood. Traditional variationist sociolinguistics offers only a limited theoretical contribution to this work, as most of the research carried out within this field is related to phonetic variables and/or alternation variables. In contrast with the findings reported for gender, the theoretical discussions within the social sciences regarding the differences in language production according to age groups are limited.

2.2.1 Psycholinguistics

The first and one of the most productive research fields to be reviewed in this survey is psycholinguistics. Although in psycholinguistics many studies analysed the development of language

abilities in the early years, until quite recently not many studies looked at the effects that the ageing process can have on the individual, especially during late adulthood. The breakthrough research in this area was led by Kemper and colleagues, whose most significant papers are discussed in this section.

Among the first studies related to language variation and ageing, Kemper (1987) located her research within previous findings that indicate that individuals older than 70 tend to show a decrease in their ability of processing left-branching sentences, that is, sentences that contain embedded or subordinated clauses before their Predicate. The sample of her study was twofold: a longitudinal sample of eight diaries written by eight subjects and a cross-sectional sample of ten diaries written by ten different subjects of different ages. The sample for each individual consisted in the longest diary entry for each half-decade. In terms of size, the samples ranged between 150 and 1300 words. Kemper's analysis of sentence types and complexity revealed that for both the longitudinal sample and the cross-sectional sample older age corresponded to a decrease of frequency of relative clauses, that-clauses, wh-clauses, infinitives, and double and triple embeddings. Moreover, a reduction in the number of clauses per sentence was noted. With certain tentativeness linked to the small sample size, the author concluded that the difference observed can be explained as the result of the reduction of working memory capacity that generally affects the ageing brain.

Kemper *et al.* (1989) continued this research in how syntactic complexity decreases with ageing by carrying out an experiment aimed at replicating the previous findings even across genres. Similarly to her other study, the researchers' aim consisted in verifying that elderly adults' working memory capacity decrease is likely to influence the degree of complexity that they present in language. This time the experiment involved a larger sample of 30 young adults (from 18 to 28) and 78 elderly adults (from 60 to 92). Both groups were almost equally distributed in terms of gender. Most of the characteristics that could influence sentence complexity, such as general level of health, vision and hearing, level of education, employment and personal interests were controlled. The participants produced three samples from three different genres: a structured interview, an oral open question and a short written essay. The analysis resulted in interesting findings for both level of education and age. As far as age is concerned, consistently with the researchers' predictions, the authors found that among the variables analysed the mean number of clauses per utterance and the frequency of left-branching clauses significantly decreased with age. The researchers also found that this decrease was equally spread across the life span, rather than affecting the participants only from a certain age.

The authors proposed two explanations for the differences found, expressing their view that the most likely explanation of the effect can be indeed a combination of the two explanations. The first explanation would theorise that, with age increasing, working memory decreases and therefore the production of complex sentences is to a certain degree impaired. This effect can be reasonably assumed since the measurements of sentence complexity were negatively correlated with the scores that the participants obtained in a working memory capacity test. The second explanation, however, proposes

that this difference in language is not the result of impairment but the effect of a linguistic style. Elderly subjects avoided complex sentences because their linguistic experience suggested to them that these are the kinds of sentences that are more difficult to understand. To confirm this hypothesis, the researchers presented the essays written by the participants to a series of judges. The essays written by elderly adults were indeed considered the clearest and most interesting essays overall. Furthermore, a correlation test showed that the judges' positive opinions were significantly negatively correlated with the level of complexity of the sentences of the essays that were judged clear and interesting.

Kemper contextualised these and other findings within a wider theory in Cheung & Kemper (1992). The authors examined a data set consisting of oral narratives produced by elderly adults aged 60-90 years. The researchers then tested a series of hypotheses using a structural equation model in order to verify which theoretical model can successfully predict the interrelations between age, working memory, verbal abilities and linguistic complexity. The most fitting of the proposed models was the one that explained Linguistic Complexity as a measurement that is comprised of three factors: Length, Amount of Embedding and Type of Embedding. The model revealed that Linguistic Complexity was positively correlated with measures of working memory capacity, thus confirming that more working memory needs to be available when producing more complex sentences. Since elderly adults tend to suffer from working memory decrease, they consequently tend to produce less complex structures. The authors also found that measures of vocabulary were positively correlated only with level of education and not with age, thus showing that vocabulary is not affected by the ageing process. Kemper's *et al.* (1989) finding regarding the reduction of left-branching clauses in elderly adults' language is therefore explained by this theory as being the effect of limited availability of working memory. Since left-branching clauses are more difficult to manipulate, they require a type of effort that is easily avoided when the working memory is even slightly impaired.

Kemper and her colleagues continued the series of studies on language production and ageing on several different samples and found results compatible with previous research. In Kemper & Sumner (2001) the authors examined speech samples gathered from 100 young adults (18-28) and a sample of 100 old adults (63-88). The participants were subjected to a battery of verbal fluency and working memory tests. The researchers analysed the language samples by looking at two measures of linguistic ability that were found to vary with age and cognitive abilities in previous studies. Both of these measures related to some extent to grammatical complexity.

The first measure was Development Level, or D-Level. This variable is a measure of clause complexity based on which type of clause is produced. Clause types that are known to be developed later by children score higher than simpler clauses. The second measure was Propositional Density, or P-Density. P-Density is 'a measure of the extent to which the speaker is making assertions (or asking questions) rather than just referring to entities' (Brown *et al.,* 2008: 3). It is approximated to how much information is packed in a sentence, relative to the number of words. Since propositions roughly equal

the number of verbs, adjectives, adverbs, prepositional phrases and conjunctions, usually P-Density can be calculated computationally by counting these parts of speech.

The analysis pointed to older adults presenting a richer vocabulary and higher type-token ratio and to young adults presenting a higher working memory capacity and using more complex sentences, measured with a complexity index that included D-Level and P-Density.

Two more thorough and wide-ranging studies on the relationship between ageing and these two linguistic variables were Kemper *et al.* (2001) and Kemper *et al.* (2001). These studies were based on a large scale longitudinal experiment often referred to as the Nun Study. This study involved the participation of two convents of nuns in a series of experiments on ageing, dementia and Alzheimer's disease and on how these factors influence cognitive abilities. The Nun Study offered a unique perspective as it allowed the researchers to examine how cognitive faculties gradually decline with age. It also allowed the researchers to test hypotheses regarding the influences of dementia on language production. Among the tests that have been carried out, Kemper and her colleagues analysed the language that the nuns produced in autobiographies during their life. A total of about 150 nuns participated in the study from age 17-32 to age 78-90. Every year the participants were assessed by a battery of tests designed to study the effects of Alzheimer's disease on cognitive function such as short-term memory and visuospatial ability. Furthermore, the participants provided the researchers with autobiographies of their lives at the time they joined the convent as well as two, three or four other times at different stages of their lives until the end of the study. Participants that at any stage were found to present signs of dementia were analysed in a different group from the ones that did not show any evidence of dementia. The language samples were then analysed for D-Level and P-Density. Both these measurements were scored manually for the last ten sentences of each sample and presented good inter-coder reliability scores.

The results were consistent with previous research in showing that for both the group that showed signs of dementia and the group that did not there was a significant decrease of D-Level and P-Density with age. For the group that presented signs of dementia, the scores for these two measurements were on average lower and the rate of decrease with age was steeper. Based on previous literature that pointed to similar conclusions, the authors suggested that low scores on D-Level and P-Density in younger age can to some extent predict development of dementia in older age. Although level of education was shown not to influence significantly these two linguistic scores, significant differences for the scores were found between the two convents. The authors argue that this might point to the fact the different intellectual lifestyle between the two convents can be of some significance.

The Nun Study provided a large data set that was often re-examined by other researchers or within other research paradigms. Kemper herself looked at the data set from a linguistic-oriented point of view in Mitzner & Kemper (2003). The aim of this paper was to understand how ageing and dementia effects on language production differ according to the medium of production. A sub-sample of the Nun

Study was therefore selected for which both spoken and written texts were available. This sample consisted of 118 nuns aged between 78 and 91. This set of subjects was also assessed with a series of tests aimed at measuring cognitive abilities and the ability to perform daily activities. The researchers applied the same methodology employed in Kemper *et al.* (2001) and examined the same linguistic markers.

Not surprisingly, the authors found that written language and spoken language differed significantly for almost all the markers tested. They found that written texts presented more left-branching clauses, higher type-token ratio, longer utterances, more clauses per utterances and fewer sentence fragments. The P-Density of the written samples was higher than the one for spoken samples but the D-Level for written and spoken samples was not significantly different. In terms of the effects of age and dementia on linguistic production, the researchers found that written language was more likely to manifest the linguistic impairment than spoken language. However, this impairment was not correlated with age, as found by Kemper *et al.* (2001), but only with working memory capacity. Independently of age, the subjects that manifested a lower score on the cognitive ability tests also produced more main clauses, fewer right-branching clauses, higher type-token ratio, shorter utterances both in terms of number of words and number of clauses, lower P-Density and lower D-Level scores. Except for the finding on type-token ratio, these results largely confirm the findings of Kemper and other researchers in different studies. The lack of a significant age effect can be explained by the fact that the sample they used consisted of old adults that had already abundantly passed the stage in which impairment usually manifests. It is hypothesised by the researchers that 'the age-related decline may slow down or asymptote in late life' (Mitzner & Kemper, 2003: 471). This study thus confirms that working memory decrease has a more significant and visible effect on language production of written texts. Furthermore, this study strengthens the position that the decrease noted in previous studies regarding language complexity with age is indeed likely to be related to working memory capacity.

Among the studies conducted by Kemper on language and ageing, another finding concerned the processing of noun phrases. Kemper *et al.* (2011) carried out a laboratory experiment with 40 young adults (mean age: 21) and 40 old adults (mean age: 76) on the production and planning of complex sentences. The sample of participants was assessed with a battery of tests aimed at measuring working memory capacity, processing speed and vocabulary level. The participants were shown fragments of sentences on a computer screen and asked to complete the sentences while also tracking a rotor. The subjects were shown a simple noun phrase subject of a simple verb phrase (e.g. 'John transferred') at the centre of the screen and they were asked to complete this sentence with another noun phrase functioning as a direct object (e.g. 'some plans') plus a prepositional phrase functioning as indirect object (e.g. 'to Lee'). The second noun phrase was randomly chosen to be complex or simple. Complex noun phrases included embedded sentences, modifiers or embedded prepositional phrases, whereas simple noun phrases consisted of simply a noun. Although all the subjects were more likely to shift the

complex noun phrases at the end of the sentence, older adults were significantly more likely to do so. Older adults who also scored lower in the tests for working memory capacity and vocabulary were even more likely to shift the complex noun phrases at the end of the sentence or fail to complete the sentence altogether.

The explanation for this phenomenon lies in the fact that working memory decrease impinges on the ability of keeping in mind and tracking the participants of a clause. In an experimental setting such as the one of Kemper *et al.* (2011) in which the participant is asked to dedicate part of their working memory capacity in tracking an object, this effect was even more evident. Older people or individuals with lower working memory or vocabulary levels are more likely to find this task difficult.

Although all of Kemper's experiments seem to suggest the same findings, it is likely that not all the possible explanations for the observed decrease in grammatical complexity and P-Density were considered. Kemper and her colleagues tended to hypothesise that the decrease in production of grammatically complex sentences is related to working memory capacity because a number of studies showed that older people that experience decrease in working memory find grammatically complex sentences harder to process. This explanation is reinforced by the fact that in the Nun Study the authors found that P-Density and D-Level decreased more steeply in cases of individuals with symptoms of Alzheimer's disease. Furthermore, all these findings on the decrease of language abilities are consistent with neurological evidence of age-related deterioration of the fronto-striate loop in the frontal lobes of the brain (Harley *et al.*, 2011: 138). The review of the literature produced by Harley *et al.* supports Kemper's predictions that Alzheimer's disease speeds up this process of cognitive impairment related to the deterioration of the frontal lobe. Since the frontal lobe is related to language production and, more specifically, in the production of grammar (Pennebaker, 2011: 29), Kemper's assumptions seem to hold.

However, in Kemper *et al.* (1989), the researchers mention that the differences in sentence complexity found could be also due to different 'styles' adopted by older adults. In later works the authors seem not to expand on these observations when conducting further empirical work. The two short text samples that Kemper *et al.* (2001) present to the reader show that the differences between the same text written at age 19 and at age 80 could be regarded as stylistic differences. As Kemper notes, it seems likely that the text written by the participant when she was 19 is more emotionally concerned than the one written by the same participant when she was 80. The fact that two styles were chosen could have therefore resulted in different scores for D-Level and P-Density. Statistically speaking, a correlation between D-Level and tests of working memory capacity does not mean that low D-Levels scores are caused by low working memory capacity. The relation of causation between the two is imposed by the theory the researchers are using. At least another explanation of the data could be that ageing sees a working memory decrease and at the same time an increased likelihood of adoption of a more Informational style. In a smaller scale replication of Kemper's studies, Labov & Auger (1993) found no effect of age in the complexity of syntax in a longitudinal sample of 12 adults' speech. Aware

of the differences in methodologies and contexts, the authors' tentative but sensible conclusion is that much depends on topic and text-type. In conclusion, although the 'style' option is less theoretically supported than the 'working memory' explanation option, it seems that a more linguistically sound analysis of language and context of production is necessary to understand the causes of the observed decline of grammatical complexity with age.

A final note on Kemper's study regards P-Density. This measurement has been widely successful in many studies in being able to distinguish individuals affected by Alzheimer's disease by healthy individuals. The potential of this measurement is important and some researchers have tried to find base-rate knowledge of the distribution of this measurement for clinical purposes. In a similar attempt, a problem with P-Density was found by Spencer's *et al.* (2012) study. The researchers analysed 635 texts produced by 127 women longitudinally from 50 to 60 years old. These women answered a questionnaire including an open response five times within this time period. In this sample, the authors found that for P-Density the within-individual difference in scores was significantly higher than the between-participants scores. Although they did not find any age effect, this was predicted by the literature, as P-Density was noticed to decrease largely in people older than 60. The researchers concluded that at the present moment P-Density cannot be used for clinical purposes as it is unknown how much within-individual variation should be expected. The authors also raised a concern regarding text length. Many studies that took into account P-Density used texts of any length and sometimes analysed only a short fragment. As Spencer *et al.* (2012) showed, however, the last batch of responses which was characterised by a lower than average text length presented a variance in P-Density that was 68% higher than the other variances. As expected, therefore, limited text size resulted in inaccurate estimation of the variable. At the moment it is unknown how long a text should be for a reliable estimation of P-Density. However, since the calculation of P-Density involves the counting of some basic parts of speech, perhaps future studies similar to Biber's (1993) work on the reliability of measurements of parts of speech in corpus linguistics will help to improve the accuracy of P-Density measures.

Although the present survey largely focused on Kemper and her colleagues' findings, a number of other studies carried out by several researchers independently replicated many of the effects reported above. For example, after analysing a larger sample consisting of short descriptive essays produced by 240 participants (age range: 80-86), Bromley (1991) agreed with Kemper's conclusions that older people produce less sentences presenting complex syntax as well as less subordinating conjunctions.

More recently, Engelman *et al.* (2010) replicated the Nun Study measurements of P-Density on a new study named Precursors Study, which involved medical students of an American university. Engelman *et al.* (2010) successfully confirmed that low P-Density in young adults' writing was correlated to the development of Alzheimer's disease in later life. The researchers also propose the possibility that P-Density can be considered as a proxy to *cognitive reserve*. This concept was elaborated

in gerontology to account for socioeconomic differences in the development of Alzheimer's disease. As the authors point out, 'those with greater cognitive reserve will maintain better function at similar levels of brain disease due to an increased capacity to compensate for damage. Educational or occupational attainment, literacy, and IQ scores are commonly used proxies for cognitive reserve' (Engelman *et al*., 2010: 706).

Another more recent replication of Kemper's study in a different data set found results that are compatible with previous findings. Rabaglia & Salthouse's (2011) collected language samples consisting in elicited short texts from about 900 adults from 18 to 90 years old. The participants were subjected to a battery of tests aimed at measuring cognitive abilities such as processing speed and working memory capacity. After running an exploratory factor analysis on several lexical and grammatical variables, the authors found that two factors emerged that represented respectively lexical complexity and grammatical complexity. They also found that these two factors were correlated with age in the predicted direction, that is, lexical complexity increasing with age and grammatical complexity decreasing with age. In particular, the grammatical complexity result was replicated even in a sentence completion experimental task conducted on a sub-set of the participants.

In an attempt to find the cause of these changes, the authors statistically controlled the two factors for age and examined the correlations of the factors with the results on the cognitive abilities tests. They found that lexical complexity was significantly correlated with the vocabulary test. Since the vocabulary test is a proxy to crystallised world knowledge, the authors proposed that as people age their knowledge of the world increases and, therefore, they are able to express more concepts with refined and more specialised vocabulary. As for the grammatical complexity findings, however, the authors concluded that there is no definitive explanation and that there are at least two possibilities: either the decrease in grammatical complexity is correlated to some other cognitive ability that was not measured in this study, including, perhaps, a language-specific working memory capacity; or the decrease is caused by a social-behavioural factor. This latter term was used quite vaguely by the authors and included at least two phenomena: the decrease likelihood that elder people are exposed to more complex syntax; and the fact that for older people there is a greater gap with the school years. As it will be shown below, this explanation for the decrease of grammatical complexity fits well with other findings in social psychology and computational linguistics. Among the social-behavioural factors, another explanation that could be hypothesised is the same as the one that applies to the increase in vocabulary size, that is, the increase in world knowledge. As Kemper *et al.* (1989) found in her study, older people tended to produced essays that were clearer and better to read. Perhaps the loss of complicated syntax could be due to the understanding of the older and more experience individuals that complex syntax is not easy to understand for the reader/hearer. Furthermore, it is likely that if more words for complicated concepts are available, then less complex syntax is needed to express the same concepts. This inverse correlation between lexical complexity and grammatical complexity was

proposed by Halliday as the opposition between Lexical Density and Grammatical Intricacy (Halliday, 2004). Quite puzzlingly, Rabaglia & Salthouse's (2011) did not verify whether P-Density decreased with age. Although this study is rather thorough in the theory and in the explanations, the lack of details on the methodology of the analysis does not allow an accurate replication in the present study.

Finally, not all the studies in psycholinguistics on language variation and ageing focused on sentence complexity. Byrd (1993) tested the hypotheses that elderly adults' general decrease in higher cognitive abilities can affect efficient writing of prose texts. The author analysed short essays of about 300 words written by 200 subjects, equally divided in young adults (mean age: 20) and old adults (mean age: 68). The essays represented three kinds of prose ordered in increasing difficulty of composition: narrative description of a place, a comparison essay, and an argumentative essay. The subjects were also tested for their cognitive abilities through a battery of vocabulary and working memory tests. For all the variables analysed, the researcher found that genre effect was stronger than age. However, when inserting as predictors the scores on the cognitive ability tests in a multiple regression, the analysis showed that the low levels of degree of cohesion in the text are predicted by low working memory capacity and old age. Contrary to other studies, type-token ratio was not found to vary with age but only with level of education. The author's explanation for this finding was based on a three-fold model of composition writing. In a model of writing that corresponds to the three stages of planning, translating into words, and reviewing, the planning process reflected in the measures of cohesion seems the stage most affected by decrease of working memory and therefore by age. Unfortunately, Byrd (1993) did not test Kemper's measures of sentence complexity in his own data set and did not mention how his study fits within the other body of research on the effect of decrease of working memory on language production.

2.2.2 Social Psychology

As opposed to the psycholinguistics researchers' concern of examining the link between cognitive abilities and language, research on social psychology focuses on the studies on personality and social behaviour. The only study reported in this field is Pennebaker & Stone (2003).

The researchers collected a large sample of individuals consisting of more than 3000 subjects from eight to 85 years old. This data set was constituted of data sets coming from previous studies carried out across English speaking countries. The data sets consisted of: 1239 essays on emotional disclosure; 877 essays on superficial topic, such as plans for the day; 326 interviews; and 809 writing tasks dealing with an emotional event or experience. On average, each individual contributed with 1151 words. This data set was analysed with Pennebaker's LIWC programme. All the scores for each linguistic category were statistically controlled for sex and genre.

The finding of the study only partially confirmed the initial hypotheses. As predicted by previous studies in social psychology, older age was associated with more positive emotion words and less negative emotion words and also with fewer words related to time. However, contrary to expectations, concern on the past was not found and, indeed, significantly more present and future tenses were found as opposed to past tense verbs. Again contrary to expectations, a measurement of verbal ability such as the frequency of words longer than six letters showed a significant increase with age. Moreover, first person pronouns were noticed to drop with age. It is important to note that although many of these results were highly significant, almost all of them presented a small magnitude.

The explanations given by the authors regarding these findings are related to social psychology theories. As people get older, they tend to use less negative emotion words and more positive emotion words and this provides further evidence to the widely accepted finding in psychology that levels of neuroticism and depression tend to decrease with age. The fact that the use of first person pronouns was found to decrease is consistent with this theory. The correlation between the use of first person pronouns and neuroticism and depression was already noted in the survey of gender. The finding related to words longer than six letters was surprising and even more so as it was the finding that reported the largest magnitude. However, since the use of long words is correlated with level of vocabulary knowledge (Pustet, 2004), the fact that the frequency of long words increases with age is not surprising, as vocabulary knowledge was also shown in other study to increase with age. Furthermore, as the authors admit, the sample that was used for the oldest age group was likely to consist of better educated and healthier people than a true random sample of the population for that age group and this might have skewed the results. Finally, the most surprising finding was related to the higher use of future tenses and present tenses as opposed to the hypothesised past tenses. However, no psychological explanation was proposed by the authors to account for this finding. Overall, Pennebaker & Stone's (2003) study has the advantage of bringing another point of view on the study of language and ageing.

Although the problem with the magnitude of the results and the rather naive linguistic analysis that they employed do not make a strong argument for their conclusions, Pennebaker & Stone’s findings seem to be indeed consistent with the literature on personality and ageing. In a large-scale study on more than one million and two hundred subjects, Soto *et al.* (2011) measured correlations between scores on the self-assessed Big Five dimensions of personality and age with the aim of testing the hypotheses and preliminary findings already identified by previous literature on a larger data set. Soto's *et al.* (2011) study found that the personality dimension of Neuroticism significantly decreased with age and that on average the scores on Neuroticism were higher for women, as already noted in the section on gender. Furthermore, they noted that the trait called Openness to Experience had a positive correlation with age, although smaller in magnitude than the one for Neuroticism and less widely replicated by other studies than the finding on Neuroticism.

These findings are relevant to the present work since they seem to tie up several aspects together. As reported in the section of gender, there is a link between high rate of first personal pronouns and high scores on the Neuroticism dimension. The fact that the ageing process was reported by this study and other studies as a factor that tends to reduce the predisposition to Neuroticism in individuals can be an explanation of the fact that Pennebaker & Stone (2003) found low rates of first person pronouns and negative emotion words in their data set. Indeed, in Soto's *et al.* (2011) study, the finding on the decrease of Neuroticism was among the ones that presented the highest magnitude.

On the other hand, the researchers found that the trait Openness to Experience increased with age. This finding can provide an explanation to the apparent contradictory result obtained by Pennebaker & Stone (2003) that older people were more likely to use words longer than six letters. Openness to Experience is the trait of personality that corresponds to intellectual activity and success. The fact that this trait increases with age can explain why older people reduce the number of first person pronouns as well as increasing the number of longer words. Furthermore, the positive correlation between Openness to Experience and deep formality (and, therefore, Biber's (1988) Dimension 1) has already been found by Heylighen & Dewaele (1999) in a series of studies. This finding creates another link between personality, language and age. This connection is related to the correlation between age and Dimension 1 found by computer scientists which is reported below.

2.2.3 Computational linguistics

Among the social variables examined in computational linguistics studies, age is rather common, although not as common as gender. As seen for gender, however, although the computational studies have the advantage of working with large data sets and sophisticated software, these studies often lack theoretical explanations. The only study reported in this survey is the study of Schler *et al.* (2006), already reported in the survey on gender. This study was conducted on a vast data set comprising about 1,500,000 blog posts authored by about 71,500 authors. Since these blog posts were harvested automatically, the authors were automatically classified according to the gender and age they reported in their blog. The researchers examined the corpus for both content and style features and aimed at verifying whether male and female as well as different age groups wrote blog posts in different ways and/or about different topics.

Based on their observations of gender, the authors found that ageing corresponds in a similar fashion to an increase in Informativeness or, in other words, to a shift towards the 'male style'. Using a machine learning algorithm, the researchers could correctly classify the age of an unknown blog post with an accuracy of about 70%.

Although this section presents only one study, it is important to note that this finding is significant in linking other variables presented in other sections. Indeed, Biber's (1988) Dimension 1

Informativeness pole is another formulation of high Lexical Density, which was shown to be in a negative relationship with Grammatical Intricacy, in turn a proxy to many of the grammatical complexity variables considered by Kemper and her colleagues. Furthermore, Dimension 1 is also a proxy to lack of pronouns, less adverbs and intensifiers. The significance of this study lies therefore in the fact that many of the results obtained by other studies were replicated in a completely different and vast data set in the guise of the opposition between Informational and Involved discourse within Dimension 1.

2.2.4 Corpus Linguistics

Within corpus linguistics, the most important contribution to the present work in relation to the study of language and age is the work of McEnery (2006) on swear words already introduced in the section on gender. Starting from previous literature's conclusions that swear words are markers of identity for teenagers, McEnery (2006) investigated the relationship between use of swear words and age in his own corpus of swearing derived from the spoken section of the BNC. He found that indeed, for both males and females, the frequency of swear words decreased with age. More specifically, the frequency of swear words reached a peak at 25 and from there it slowly decreased.

Taking into account the strength of the swear word, the same pattern held. Younger people tended to use stronger swear words than older people. Since the grammatical category and the strength of the swear words are related, this finding can also be extended to grammatical category. In general, for grammatical category, the main finding was that 60+ individuals were more likely to use the Idiom type, one of the weakest categories, and avoid the Personal type. The category Personal, which is the category that typically contains the strongest swear words, in McEnery's data set decreased with age until disappearing after 60 years old. As McEnery notes, however, it is clearly not possible to establish whether the pattern found is a pattern related to ageing or a change in general culture trends. Also, as specified for gender, it is important to understand that this age effect is interrelated with gender and social class effects.

A relevant study on the border between sociolinguistics and corpus linguistics is Barbieri (2008). The author's intention was to explore the patterning of lexicogrammatical items at different age stages, thus aiming at filling the gap of traditional sociolinguistics, which has often ignored this dimension of variation. Barbieri examined a sample of the American Conversation corpus consisting of 400,000 words of casual conversation divided in two sub-corpora according to age, the 15-25 (younger speakers) and the 35-60 (older speakers) age groups. The method that the researcher used consisted in a bottom-up approach that did not impose any theoretical model. Barbieri applied a key word analysis in order to find the words that occurred significantly more often in the sub-corpus of younger speakers

as opposed to the sub-corpus of older speakers. The results she found are generally consistent with McEnery's (2006) findings.

In terms of swear words, as already pointed out, the younger speaker group used swear words more often. These words and other slang words such as *cool*, *dude*, *man*, *bucks*, *booze* were at the top of the key word list for the younger speakers group. As found within social psychology by Pennebaker, in Barbieri's corpus the younger speakers used first person singular pronouns more often than older speakers. Older speakers, on the other hand, used third person singular and plural pronouns as well as first person plural pronouns more often. Modal verbs were found to be more frequent in the older speakers's sub-corpus, except for the modals *can*, *shall* and *need to*. The fact that older speakers were found to use the modal *will* more often than the younger speakers could be a further piece of evidence for Pennebaker & Stone's (2003) argument that older speakers tend to shift their focus to the future tense.

The most interesting and intricate findings concerned the use of adjectives and adverbs. In the younger speaker's sub-corpus more evaluative adjectives were found, such as *crazy*, *awesome*, *shitty*, *hot*. This group was also more likely to use intensifiers, that is, 'adverbs that boost the meaning of other clausal elements' (Barbieri, 2008: 71). The kinds of intensifiers used were also distinctive to a certain extent. For example, younger speakers were more likely to use *really* than *very* to intensify. In relation again to adverbs, younger speakers seemed also to favour the more common, informal and innovative epistemic stance adverbs, such as *kind of*, *sort of*, *probably*, *actually*, as opposed to the more traditional *certainly*, *usually*, *typically*, which were favoured by older speakers. Although this study was conducted on a corpus of American English, Barbieri cites the work of Rayson *et al.* (1997) on the spoken sub-corpus of the BNC that found compatible results. For the sake of the present work, the list of words found by Barbieri was expanded using the list provided by Rayson *et al.* of British equivalent adjectives and adverbs (e.g. *bloody*, *massive*, *brilliant*).

The explanation that the author proposed for these findings considers the fact that all these features are to some extent related to the notion of stance. Swear words, slang, intensifiers, stance adverbials, modals and evaluative adjectives are all elements that interactants use to express their stance with other interactants in communication. It seems therefore that 'through the active use of a much wider variety of stance-linked devices, American youth are able to convey a more overt, explicit, and salient stance than adults do' (Barbieri, 2008: 78). Furthermore, it seems likely that younger speakers tend to use the elements that are linguistically more innovative, as opposed to the more traditionalist markers preferred by older speakers. This can be seen, for example, in the kind of epistemic stance adverbs used by the two groups. The same phenomenon can also explain why older speakers use more modal verbs, which represents the traditional way of marking stance. Finally, it is also likely that adults and older adults seem to be less prone to display feelings and attitudes and, when doing so, they tend to use less vague and innovative means.

2.2.5 Discussion

Although the quantity of the reviewed research on age is not as wide as for the one for gender, some interesting findings were reported. The various studies reviewed above could be grouped in four general patterns of language variation. The reviewed studies organised by pattern are summarised in Table 2-2 below.

Table 2-2 - Summary of the studies reviewed for age. Variables in bold represent variables that increase with age whereas underlined variables are variables that decrease with age.

Study	Genre of data	N of participants	Average or min-max text length	Year of data	Summary of linguistic variables	Country
Pattern 1: Syntactic complexity Syntactic complexity at the clausal level decreases with age and this often corresponds to an increase in lexical complexity. This effect could be due to either a decrease in working memory capacity with older age or to a shift in the way information is packaged due to increased experience with language.						
Kemper *et al.* (1989)	Oral interviews, oral open questions and written essays	108	N/A	1989	<u>average number of clauses per utterance; frequency of left-branching clauses</u>	USA
Byrd (1993)	Narrative descriptions, comparison essays and argumentation essays	200	300	1993	<u>number of cohesive ties per sentences; number of cohesive ties per sentence weighted by intervening sentences</u>	New Zealand
Kemper *et al.* (2001a)	Speech samples	200	N/A	2001	<u>average sentence length; D-Level score; P-Density;</u> **type-token ratio**	USA
Kemper *et al.* (2001b)	Written autobiographies	150	N/A	From 1930 to 1996	<u>D-Level score; P-Density</u>	USA
Pennebaker & Stone (2003)	Emotional disclosure essays and interviews	3000	1151	2000	Relative frequency of: **words longer than six letters**	USA, New Zealand, England
Rabaglia & Salthouse (2011)	Descriptive essays	900	N/A	2011	<u>number of embedded clauses; number of subordinate clauses; number of left-branching clauses;</u> **words longer than 5 letters; ratio syllables/words; logarithm of word frequency for content words; type-token ratio**	USA

Pattern 2: Dimension 1 Older age is correlated to a higher use of Informational features and with less frequent Involved features (using Biber's (1988) terminology).						
Pennebaker & Stone (2003)	Emotional disclosure essays and interviews	3000	1151	2000	Relative frequency of: first person pronouns	USA, New Zealand, England
Schler *et al.* (2006)	Blog posts	37,478	N/A	2004	Relative frequency of: personal pronouns; negations; **determiners; prepositions**	N/A
Pattern 3: Realisation of stance Generally speaking, younger people tend to use stronger linguistic stance than older people. This pattern could be either due to language change or to change in personality, life-goals and attitude with ageing (Bourdieu's theory of distinction).						
McEnery (2006)	Conversation	N/A	N/A	1990	Frequency of: swear words	UK
Barbieri (2008)	Conversation	139	N/A	1990	Relative frequencies of: swear words; slang; evaluative adjectives; innovative stance adverbs; **modal verbs; traditional stance adverbs**	USA
Pattern 4: World-view change As people get older their view change towards more positive feelings and towards looking to the future. This effect could be due to an average decrease in neuroticism and depression that naturally happens with age.						
Pennebaker & Stone (2003)	Emotional disclosure essays and interviews	3000	1151	2000	Relative frequency of: negative emotion words; past tenses; words related to time; **positive emotion words; future tenses**	USA, New Zealand, England

The most replicated pattern is the one firstly examined by psycholinguistic research – the decrease of syntactic complexity with age. Even though four of the studies were conducted by the same group of researchers, the review pointed out that the same findings were reproduced by other researchers using different data sets. These findings generally indicate that ageing is characterised by a decrease in the complexity of sentences produced. The most widely accepted explanation for this finding is that there is a correlation between sentence complexity and working memory. The correspondence between ageing and sentence complexity is therefore a second-order correlation given by the fact that working memory capacity decreases with age. This explanation has been challenged with another linguistically-based explanation. Indeed, the decrease in syntax complexity has often been found to correlate to an increase in lexical complexity, which could indicate that it is a style shift rather than a memory deficiency that underlines the linguistic pattern. Since older age corresponds to more experience of language use, it is possible that experienced users of language are more capable of packaging information in smaller and more easily parsable chunks.

This explanation would also fit the second pattern listed in Table 2-2, the increase of nominal complexity. The finding that older people have higher Informational scores on Dimension 1 could also suggest that the implementation of a more syntactically simplified but lexically rich and nominal style by older and more experienced adults is responsible for the apparent decrease in complex sentence production.

A third pattern that can be noted regards the management of stance. In corpus linguistics, two studies were reported in which it was found that the frequency of swear words is affected by age. More generally, Barbieri (2008) concludes that the younger group she studied was more likely to produce stronger and more innovative elements of linguistic stance.

The final pattern included in Table 2-2 is connected to the previous pattern. In social psychology it was found that there is change in linguistic attitudes across life-stages that is likely to reflect a change in a more positive perspective towards life. An established finding in psychology is that older adults are on average less prone to depression and neuroticism. This relationship was found in language in the form of fewer negative emotion words and more positive emotion words.

In conclusion, the review of the research on language use and age can be summarised in four general patterns of language variation, of which one is strongly supported by many studies that have been carried out in the past. In Chapter 5, the presence of these patterns is tested in fabricated malicious communication to understand to what extent these linguistic patterns are found in texts resembling typical malicious texts. The present Chapter moves to the review of the research of language variation and level of education.

2.3 Literature survey on Level of Education

As opposed to the research reviewed for gender and age, the directly relevant research on level of education is rather limited. The most significant findings related to level of education originate from research in language development and teaching of English as first language. However, the majority of the work in these areas was not concerned with finding out how certain markers of linguistic development are maintained or changed throughout life and/or with further education. Although a leap needs to be taken to expand these markers to authorship profiling, the present work takes the stance that the markers that distinguish communicatively proficient school children from less communicatively proficient school children are also likely to distinguish communicatively proficient adults from less communicatively proficient adults. The extent to which this is the case is an empirical question that is answered after the analysis of the data of the present work in Chapter 5.

Other markers in this survey are taken from some of the studies that were considered in the section on age where the authors reported the level of education of the individual as a control variable. Some of these reported correlations regarding linguistic variables and level of education can be regarded as hypotheses of how level of education affects language use.

2.3.1 Teaching of English

The studies here reported are taken from the study of the teaching of English as a first language. These studies typically examine how the English language is taught from primary school to age 18 and measure the proficiency of students in communicating using English.

The first study to be examined in this section is Loban's (1967) large scale report of a longitudinal study of American school children. Although Loban’s findings are not recent, this study has the merit of being one of the first works assessing the problem of formalising the acquisition of linguistic competence for school children learning English as a first language. Loban’s (1967) study was also one of the first studies to show findings indicating differences in achievements between different socioeconomic classes. Loban’s study was conducted over 13 years ago and it involved 211 school children that were examined for the whole time they spent in formal education from kindergarten to age 18. The students were assessed each year with a range of tests, including oral interviews, tests of listening ability, written language tests, and IQ tests. To study the linguistic differences and their stages of development, the students were also grouped according to the marks that they received from the teachers.

One of the most important finding in this study was that the average length of communication units increased almost linearly with years of education. A communication unit was defined as an independent clause plus its dependent clauses, although various adjustments were made by the researchers to accommodate spoken language. When grouping students according to their proficiencies

at school, Loban found that the average length of communication units for a student was an excellent approximation to the student's evaluation of language skill by their teacher. Furthermore, although this measure increased with years of schools, students with lower marks were always producing scores that were lower than the ones produced by students with higher marks.

Loban investigated this finding more thoroughly by checking for differences in terms of the number of clauses and dependent clauses per communication units. He found that, although there was no significant difference in terms of which kinds of dependent clauses were used by the high proficiency students and low proficiency students, there was a highly significant difference regarding how many dependent clauses were used per communication unit, with an advantage for the high proficiency students. Loban also found that the use of subordinating connectives (e.g. *however*, *because*, *although*, *therefore*) were learned more quickly, used more frequently and used more accurately by the high proficiency group as well as by the higher socioeconomic groups. Finally, assessment of non-standard English features by ethnicity and social class showed that high proficiency students were more likely to avoid non-standard features of English.

Similar findings relative to syntactic complexity were found in a study by Hunt (1971). Two-hundred and fifty school children were selected across different grades and asked to rewrite the same passage constituted of simple sentences in a better way in their own words. The same task was submitted to 25 skilled adult writers and 25 firemen that had been away from formal education for 10 years. Hunt found that the number of subordinated or embedded clauses used to package the fragmented information given in the experimental passage increased with years of schooling. More linguistically mature individuals used less and less often immature devices of combining information such as coordinated clauses and instead adopted subordinated clauses or even more advanced compression strategies, such as noun phrase packaging. Hunt proposed that a good approximation measure for establishing linguistic maturity is the average length of the terminable units (t-units). This linguistic construction is defined as Loban's 'communication unit', that is, an independent clause plus its subordinated or embedded clauses. In general, Hunt found that clause length and t-unit length increase with linguistic maturity and that these are positively correlated with IQ.

It is again Hunt (1983) that expands on t-unit research by exploring how t-unit variables are connected to each other. As the author reports, sentence length has been often regarded as a reliable measure to assess language development in school settings. However, since then, sentence length has been found to be too dependent on an author's style rather than competence. Indeed, skilled writers can manipulate punctuation to create shorter sentences and slow the pace of reading to obtain certain rhetorical effects. As seen in the two studies above, t-unit length as opposed to sentence length is a good approximation to communicative competence. However, the combination of these two elements in a t-units per sentence ratio allows the measurement of another component of literacy development, that is, the ability to punctuate. Being able to punctuate following the Standard English pattern is equal to

maintaining a ratio sentence to t-units of roughly 1, or, in other words, to maintain a good pace between clauses and breaks. This is a skill that is acquired only late in formal education and mastered completely only by skilled adult writers (Hunt, 1983: 102).

Hunt also notes that in mature elaborate written texts t-units do not just contain the main clause but also a number of subordinate clauses. Consistent with this, Hunt found evidence in the literature that the number of clauses per t-unit grows with the school age of individuals and reaches its peak in skilled adults. Although it does seem the case that t-unit length increases because more subordinate clauses and embedded clauses are produced, Hunt suggests that t-unit length also increases because clause length increases with education. More mature writers not only adopt more subordinate or embedded clauses, but also write clauses that are on average longer. Research reported by Hunt (1983) supports the hypothesis that the contribution to long t-units for skilled adults is not simply given by the number of subordinate clauses but also by the higher clause length. The interaction between all these measures sheds light on how linguistic maturity is achieved and assessed in formal education. Skilled writers are supposed to be able to produce on average long t-units, possibly punctuating them as one sentence, with some subordinate or embedded clauses and especially using long clauses. However, although this does not necessarily imply that a skilled writer always writes using this pattern for any genre or for any rhetorical purpose, research predicts that they would be competent and comfortable in doing so when this is needed.

As Hunt suggests, however, the average length of t-units is mostly influenced by the amount of long or short t-units produced. Therefore, rather than simply examining the average score, it is also important to notice the quantity of long and short t-units produced. Furthermore, Hunt reports that some studies also showed that these differences are influenced by the IQ of the students, with high IQ students more likely to approach skilled adult's competence sooner. This effect is noted in particular regarding clause length. Individuals with higher IQ tend to produced longer clauses and less subordinate or embedded clauses than their respective average IQ peers.

A problem with Hunt's two studies is that not all the linguistic productions require high clause length or high t-unit length and therefore in not all the linguistic situations the linguistic maturity of an individual should be measured by how long their clauses or t-units are. Since the average scores for both clause length and t-unit length are unavoidably going to vary depending on the register, it is up to the researchers' judgment not to generalise erroneously for different registers.

Regarding the measurement of t-unit length or other measures relative to the t-units, it is worth noting the research of Witte & Davis (1980). Noticing how t-units as a linguistic tool was often used by researchers, the authors set an experiment to test whether there is intra-author stability in the production of t-units. In a limited study of 43 college students tested on three texts per student in the descriptive and narrative discourse modes, the authors found no statistical support of the hypothesis that t-unit length is constant within authors or within discourse modes. Although the study was preliminary

in nature, the researchers concluded that much care should be expressed when adopting t-units as there is no evidence that this measure is constant within authors. Witte & Davis' (1980) research was limited not only by the small sample size but also by the fact that the first year college students were not enough familiar with the registers to represent a real sample from the population of the two discourse modes, as the two authors admit. Although not many other researchers seem to have addressed the problem of the stability of certain measures such as t-unit length, it is important, as Witte & Davis (1980) suggest, to take into account register variation and the standard deviation of the variables tested when dealing with multiple samples from different authors.

Within the section on teaching of English it is also possible to locate the studies conducted on readability, of which Dubay (2004) offers a comprehensive historical review. The importance of the research on readability should be traced to research carried out by the US army on assessing intelligence, job success and literacy with a view to structure their texts according to their typical readership. In their research the US army found that measures of literacy and readability correlated with intelligence, with knowledge and with years of education achieved. Almost simultaneously to this strand of research, Kitson (1921) noticed that significant differences in terms of average sentence length and average word length between magazines influenced the kind of readership and therefore the kind of social groups that would read a certain magazine. These findings suggested that there was a correlation between the average sentence length and word length encountered by a social group and its social status.

Both the U.S. army research strand and the psychology research strand gave a start to a series of studies aimed at finding a readability formula that could predict the readability of a text. After years of research, the proposed formulae grew in complexity, although the addition of more variables only slightly increased their efficiency. In general, however, researchers found that the number of morphemes and the number of syntactic branches in 100 words were the useful variables to measure, respectively, semantic (lexical) complexity and syntactic complexity and to give an estimate as to the overall complexity of a text. Since the number of morphemes is highly correlated with the number of syllables, and since the number of syntactic branches is highly correlated to sentence length, the most successful readability formula ended up being just a combination of syllable counts and word counts. The most famous of these formulas is probably the Flesch Reading Ease score, calculated as a function of word and sentence length (Flesch, 1949).

Most of the literature reviewed in this section and in the section on ‘age’ seems to confirm at least theoretically that word length and sentence length are two measures of, respectively, lexical and syntactic complexity. Since it is possible to assume that a person can produce only language that they can read, then the scales provided by Flesch (1949) can be added to the list of variables that are tested for the present study regarding level of education.

This section ends with the review of a multi-linguistic study conducted by Berman (2008), who reports on a series of experiments carried out across countries and languages on the processes

underlying text construction across age and education phases. The sample considered by Berman consisted in 80 participants with a middle class background divided between school children aged 10, 13, and 17 and graduate school university students aged between 20 and 30. These subjects were asked to produce four texts: a narrative and an expository spoken text and a narrative and an expository written text. All the texts dealt with the same topic: problems between people. The analyses of these samples were divided in several layers, from the assessment of the quality of writing or appropriateness of the text type to the analysis of grammatical and lexical features. Many of the features were found to correlate with level of education. However, the majority of these features were not calculated in a way that is possible to replicate. Some of the variables, such as the scale of discourse construction, rely on subjective assessments of the texts. For the present work, therefore, only the variables that can be replicated are taken in consideration. These are the variables that account for the lexical and grammatical levels.

For the lexical level, Berman looked at three variables that are possible to replicate in the present work: word length in syllables; lexical density, defined in the Hallidayian sense as the proportion of content words in the text; and the proportion of words from Romance and Germanic origin. The author indicated that there is a relationship between scores on these variables and the level of literacy of the subjects, although no exact statistics are reported on the magnitude of the effects or the level of significance obtained. These lexical variables were also correlated with each other both in English and in the other language tested, Hebrew. The author suggest that a hidden factor, perhaps lexical knowledge, can be regarded as underlying those variables and correlating with education and cognitive development (Berman, 2008: 755).

For the syntactic level, Berman examined five variables: clause density, defined as the mean number of words per clause; noun phrase heaviness, defined as the combination of four different variables: average length of noun phrases in words, average number of dependent nodes, depth of noun modifiers, and number of types of noun modifiers; the proportion of relative clauses; the frequency of use of the passive voice; and the frequency of non-finite subordinations. As for the lexical variables, these variables correlated with literacy development.

In conclusion, Berman proposed that several linguistic variables, including the lexical and syntactic variables listed above, are an approximation to cognitive development and literacy level. The correlation of these variables with literacy seems to indicate that with education and with age individuals shift their focus from material, specific and relationship-based concerns to more abstract, general and entity-based concerns (Berman, 2008: 736, 755). This type of mind attitude not only improves the socio-cognitive faculty of individuals but requires a kind of language that is more elaborate both lexically and syntactically and therefore more oriented towards nouns and other nominal parts of speech or grammatical patterns.

2.3.2 Psycholinguistics

The Section on psycholinguistics studies on language and level of education includes only a series of studies already covered in the section on age in which some collateral findings on level of education were noted when examining the relationship between language and age.

The first of those studies is the work on language change across age and genre carried out in Kemper *et al.* (1989). In the section of age it was reported that these researchers found a correlation between decrease in working memory capacity in old age and decrease of number of clauses per utterance and the frequency of left-branching clauses. This effect was found in a sample of 30 young adults and 78 elderly adults who provided both spoken and written samples. The authors, however, also divided this sample according to level of education in order to avoid the influence of this variable when analysing age. Even though studying level of education and language use was not their goal, the authors found that better educated adults were more likely to produce longer utterances in terms of words, more right branching clauses and fewer main clauses. As reported previously, the essays written for this experiment were submitted to a series of judges for evaluations of interestingness and clarity. Quite surprisingly, level of education resulted to be a poor predictor of both interestingness and clarity. The most significant predictor of these two variables was, on the other hand, age. The authors do not go to great lengths to explain the effect of level of education, as that was not their goal. However, they mention that people with more years of education tended to score higher on the vocabulary test. Thus, if individuals with higher education have a larger vocabulary, they can also be able to express their ideas with a more packaged and detailed lexis and therefore avoiding a more complex syntax. This explanation is rather tentative but it does fit with the rest of the literature reviewed in the present section.

Just quickly mentioned in the section of age, Bromley's (1991) study incidentally found interesting results related to level of education. The sample that the author took for analysis consisted in 9 brief descriptive essays produced by 240 adults from 20 to 86 years old stratified for gender, level of education and occupational status. The last two variables were assessed in the form of two scores from 1 to 5 and, since they were highly correlated, they then were summarised by one single score named ‘status’. A linguistic analysis of the sample indicated that lexical variables were likely to predict the status of the participants, as opposed to syntactic variables, which, on the other hand, were more likely to predict the age of the participants. Lexical variables such as text length, number of words longer than 10 letters, average word length and the Flesch readability score predicted a combination of status and vocabulary level, as tested with a vocabulary level test. The author argues that this could indicate that the said variables are correlated with a vocabulary factor that is a function of vocabulary knowledge and educational status.

Another study within this series is Byrd's (1993) analysis of written language and ageing. The sample that the author examined consisted in short essays of three kinds (narration, comparing and

argumentation) produced by 200 subjects. These subjects were assessed with cognitive tests as well as stratified for social variables. In terms of occupation and education, the subjects were ranked on a social position scale from 1 to 7. Independently from the discourse mode, the author found that a multiple regression could predict education level and the score on the cognitive test on vocabulary from the type-token ratio and from the mean rarity scores for all the words in a text. Byrd's explanation consists in a tentative proposal that high education facilitates the phase of writing that consists in translating the mental plan of the text in lexical and syntactic forms.

Finally, Mitzner & Kemper's (2003) examined written and oral autobiographic narratives from 118 nuns aged between 78 to 91 that participated in the Nun Study for linguistic markers of working memory decline. While doing so, the authors controlled the sample for level of education measured as the total years of formal education pursued by the subjects. They consequently noticed and reported some correlations between some linguistic markers and level of education. Although the authors note that the magnitudes of the correlations were rather low, subjects with more years of education tended to produce longer utterances, more clauses per utterance, and less fragments per utterance than subjects with less years of education. The authors did not propose any explanation for this finding. However, the fact that well-educated individuals produced syntactically complex utterances seems to support some of the other findings reported in the present survey.

2.3.3 Corpus linguistics

The first study surveyed in this section is Heylighen & Dewaele's (1999) report already presented in the section on gender as being another interpretation of Biber's (1988) Dimension 1. The measure that the two authors propose, called 'deep formality', was described in the section on gender as being a score that determines how deictic a text is. The authors found this score to vary significantly with register as well as with personality and with social variables, such as gender and level of education. For this variable, the authors present limited and provisional evidence from Dutch that educated individuals as well as individuals from a higher social status tend to produce language that score higher for deep formality than other individuals. The authors explain that to produce formal language more cognitive power is needed, as there is more information to be processed. That being the case, the author theorise that educated individuals as well as individuals with more cognitive power at their disposal are likely to employ formal language more often than others. The authors also add that there might be a correlation between high scores on deep formality and the personality dimension that measures intellect in the Big Five personality test, that is, Openness to Experience.

The second and final study reviewed in this Section is Mollet's *et al.* (2010) survey of a number of linguistic variables. Their study was aimed at comparing a set of variables in a fashion similar to the present work with the scope of finding the most reliable tools to quantify linguistic features. In the

process of assessing these variables, the authors found some correlations between certain linguistic variables and the level of education of the subjects that were participating in the study. Their data set consisted of 400 to 1000 word essays not controlled for genre variation written by fifty-five 17 years old students in Australia for the purpose of an exam.

The authors surveyed a number of linguistic markers and tested them on their data to assess whether there is evidence that these markers are successful at measuring the independent variables that they are supposed to measure. The authors also compared the results obtained from these markers to a battery of psycholinguistics tests with which the subjects were assessed. Among the five variables that the authors described in the conclusions, two markers, namely, Advanced Guiraud 1000 and P-Density, presented interesting findings regarding level of education.

P-Density is a measure already encountered in the survey on age, specifically in the research on linguistic markers of Alzheimer's disease carried out by Kemper and her colleagues. The density of propositions in a stretch of language was found to be correlated with its author's working memory capacity and the likelihood for the author to develop Alzheimer's disease later in life. Compatibly with Kemper's results, in their data set Mollet *et al.* (2010: 460) found a correlation between P-Density and serial recall, a measure of working memory capacity. Furthermore, the authors found that P-Density was significantly positively correlated with the mark that the examiners gave to the essay, thus giving some weak evidence that the education system rewards compositions that are rich in propositions and in density of connections between them.

The second variable, Advanced Guiraud 1000, is a variable that approximates the measurement of extrinsic rarity of vocabulary. Mollet *et al.* (2010: 452) define the *extrinsic rarity* of the vocabulary of a text as the rarity of the words of that text in relation to the language as a whole. The best way to calculate extrinsic rarity is to compare each word to a frequency list of words of a representative reference corpus. However, a quicker approximation to this method is achieved with Advanced Guiraud 1000 by subtracting the number of types that occur in the set of the 1000 most frequent words of a reference corpus from the total types of the text and then divide this number by the square root of the number of tokens. This variable not only presented a strong significant correlation with serial recall and the mark given by the examiners but also showed a significant correlation with the verbal IQ of the subjects (Mollet *et al.*, 2010). Extrinsic rarity is contrasted to the *intrinsic rarity* of vocabulary of a text, which is the rarity of words of the text in relation to the text itself. This construct is approximated by the author through the use of Baayen's P, or the number of *hapax legomena* divided by the total number of tokens of the text. This variable was again found to correlate with the verbal IQ of the subjects as well as with their score on the Advanced Guiraud 1000.

Mollet *et al.* (2010) analysed their data sets using these two variables as an attempt to quantify the writer's vocabulary size. However, they admit that 'extrapolating from the text patterns to properties of the writer makes considerable assumptions about the relationship between what we know and what

we write in any given text […] we must be cautious about assuming that any writer displays all his or her wares in every piece of writing' (Mollet *et al.*, 2010: 438). When attempting to quantify the vocabulary of their subjects, they therefore thoroughly examine both of the aspects that can allow such a measurement, accounting for both intrinsic and extrinsic rarity. The link between these two variables and level of education is only hinted at in their research. The authors propose that Advanced Guiraud 1000 correlates with marks on assignments probably because extrinsic rarity of vocabulary is highly rewarded in the education system, either consciously or subconsciously. However, the correlations between measures of vocabulary size and level of education have been also found in other studies. These relationships point to the possibility that a higher education is likely to expose individuals to a larger number of vocabulary items than a lower education, thus giving them the chance to increase their lexicon.

2.3.4 Discussion

The number of studies reviewed for level of education was the lowest among the four social factors considered for the present work. Some of the studies actually were included because of their collateral findings that they obtained while studying other social factors. Nonetheless, the review still produced a number of patterns of linguistic variation associated with level of education that can be explored in forensic texts. The studies reviewed could be grouped in five patterns of linguistic variation, as summarised in Table 2-3 below.

Table 2-3 - Summary of the studies reviewed for level of education. Variables in bold represent variables that increase with level of education whereas underlined variables are variables that decrease with level of education.

Study	Genre of data	N of participants	Average or min-max text length	Year of data	Summary of linguistic variables	Country
Pattern 1: Vocabulary size Higher levels of education correspond to an increase in vocabulary size and lexical sophistication						
Bromley (1991)	Descriptive essays	240	N/A	1991	**average word length; frequency of words longer than 10 letters**	USA
Byrd (1993)	Narrative descriptions, comparison essays and argumentation essays	200	300	1993	**type-token ratio; mean rarity score of all words**	New Zealand
Berman (2008)	Narrative and expository speech samples and texts	80	N/A	N/A	**average word length in syllables; lexical density; proportion of words from Romance and Germanic origins**	USA
Mollet *et al*. (2010)	Essays for final assignment	55	400-1000	2010	**Advanced Guiraud 1000; Baayen's P**	Australia

Pattern 2: Sentence complexity Higher levels of education correspond to an increase of sentential syntactic complexity						
Loban (1967)	Oral and written language samples	N/A	N/A	1967	**number of clauses; frequency of subordinating connectives**	USA
Hunt (1983)	Essays, newspaper articles	18	1000	1983	**average sentence length**	USA
Kemper *et al.* (1989)	Oral interviews, oral open questions and written essays	108	N/A	1989	**utterance length in words; number of right-branching clauses;** number of main clauses	USA
Bromley (1991)	Descriptive essays	240	N/A	1991	Flesch readability score	USA
Mitzner and Kemper (2003)	Written and oral autobiographies	118	N/A	1995	**utterance length; number of clauses per utterance;** number of fragments per utterance	USA
Berman (2008)	Narrative and expository speech samples and texts	80	N/A	N/A	**proportion of relative clauses; frequency of passive clauses; frequency of non-finite subordinations**	USA

Mollet *et al*. (2010)	Essays for final assignment	55	400-1000	2010	**P-Density**	Australia
Pattern 3: T-unit complexity Higher levels of education correspond to an increase of syntactic complexity within t-units						
Loban (1967)	Oral and written language samples	N/A	N/A	1967	**average t-unit length; long t-units; dependent clauses per t-unit;**	USA
Hunt (1971)	Rewriting of a passage in a better way	300	N/A	1971	**average t-unit length**	USA
Hunt (1983)	Essays, newspaper articles	18	1000	1983	**average t-unit length; clauses per t-units; long t-units;** short t-units	USA
Pattern 4: Nominal elaboration Higher levels of education correspond to an elaboration of information that focuses on nominal devices rather than verbal devices. This translates into more deep formal discourse and a higher average clause length						
Hunt (1971)	Rewriting of a passage in a better way	300	N/A	1971	**average clause length**	USA
Hunt (1983)	Essays, newspaper articles	18	1000	1983	**average clause length**	USA
Heylighen and Dewaele (1999)	Word frequency lists	-	-	-	**Deep formality score**	Netherlands/UK

	with social information					
Berman (2008)	Narrative and expository speech samples and texts	80	N/A	N/A	**clause density; noun phrase heaviness**	USA
Pattern 5: Information distribution The distribution of information in a text is different depending on how much exposure a person had to formal education. Individuals with higher education levels tend to maintain a ratio of one t-unit per sentence.						
Hunt (1983)	Essays, newspaper articles	18	1000	1983	t-units per sentences	USA

Out of the five patterns, four represent complexities of different kinds that increase with level of education. This result is not surprising since it is intuitive that linguistic competence and complexity increase with years of formal education. Vocabulary size and syntactic complexity both at the level of the sentence and of the t-unit are all predicted to increase with level of education. However, the studies that noticed these correlations often underlined the problem of the confounding factor of level of IQ.

In the section on corpus linguistics, Heylighen & Dewaele (1999) proposed that deep formality score is positively correlated with education. Since a high score on this variable corresponds to a higher degree of nominal elaboration, this pattern also corresponds to a higher average clause length, since the size of a clause increases with larger and complex noun phrases (Hunt, 1983).

A final pattern concerns the distribution of information. Hunt (1983) has noted that the strongest distinction between people with different levels of education lies in the way they present the information. Formal education stresses the importance of maintaining a ratio of t-units per sentence almost equal to one since this ratio simplifies the presentation of information and rewards nominal complexity.

In conclusion, all the reviewed studies converge to indicate that higher education corresponds mainly to a larger vocabulary and secondarily to a good command of syntactic patterns that are distributed adequately across the text. However, a problem of previous works is that they consider formal schooling year or job achievements as a direct assessment of the level of education or level of literacy of the individual. Indeed, these two factors do not necessarily coincide, as an individual can achieve a certain level of education or literacy independently from formal education. Even though, typically, the level of education achieved is directly reflected in the language produced, it is expected that these linguistic variable correlating with level of education are indeed associated with the amount of reading and writing that the subject experienced rather than directly with their level of formal education. The correlations examined by the studies in this survey are therefore only secondarily related to level of education, as the number years of formal education is just an easy proxy to a more complex socio-behavioural construct of intellectuality and literacy which can also be self-developed.

In Chapter 5, the presence of these five patterns is tested in fabricated malicious communication to understand to what extent these linguistic patterns are found in texts resembling typical forensic texts. The present Chapter moves to the review of the research of language use variation and social class.

2.4 Literature survey on Social Class

The surveying of the relationships between social class and language is a difficult endeavour since there is no agreement on the definition of 'social class' and, to a certain extent, on the validity of this construct especially in modern society. Indeed, not a single study in the present review used the same classificatory system to rank the subjects of their study in social classes. This diversity of classification is rather alarming, since it also implies that the results of the studies reviewed in this section are not completely comparable with each other. In her review of the literature, Ash (2002) confirms this uncertainty on the definition of social class across many sociolinguistic studies. Her criticism is directed towards the fact that many linguistic studies did not attempt to understand the notion of social class from sociology before embarking in carrying out the experiments. This problem is even more alarming as recent research in sociology points out that individuals rank themselves and others in social classes based on a large number of factors, including the type of family, types of clubs or fraternities, and behaviour in general, and not just the classical indexes of income or occupation used by most studies (Ash, 2002: 404). Furthermore, indexes developed by sociologists based on the perception of education and occupation in society, such as the Werner's Index of Status Characteristics or the Duncan's Socioeconomic Index, are rarely employed by linguists in their work.

In spite of all the differences in social class indexes and the avoidance of more recent research in sociology, as Ash (2002: 402) notes, the irony of social class as a social factor used in linguistic studies lies in the fact that it is one of the social factors that typically presents the strongest effects. In a similar way, Dodsworth (2011: 192) writes that although the division in social class as it if were discrete categories is far from reality, this method has shown that linguistic variables pattern with social classes in many languages. Dodsworth (2011: 205) concludes that a more holistic approach that considers not only the classical approximations to social class but also elements such as the cultural values of the members of the communities or their social network should be considered in the future as a more appropriate social variable to explain certain linguistic variables.

Moves towards these directions were made by a recent wide scale study of social class in Great Britain: The Great British Class Survey, compiled with the help of the BBC (Savage *et al.*, 2013). The authors of this study agreed with previous comments on the inadequacy of current models of social class to account for the phenomenon, and in particular for its cultural dimension. The authors argued that a more accurate theory of social class is one that understands it as a multidimensional construct. For this purpose, the authors used Bourdieu's theory of social capital as a framework for a new exploratory study of social class. This theory proposes that three capitals are available for individuals in society: the economic capital, or the wealth and income; the cultural capital, or the degree to which the individual is engaged with cultural goods; and the social capital, or the connections in the individual's social network (Savage *et al.*, 2013: 5). The result of their analysis consisted in a model

with seven classes: Elite, Established middle class, Technical middle class, New affluent workers, Traditional working class, Emergent service workers and Precariat. The picture that emerges is therefore one in which the famous 'middle class' and 'working class' have disappeared or transformed into a number of sub-classes. In the light of this multidimensional study, the conclusions that can be drawn is that more recent sociological theory is moving towards a well-rounded conceptualisation of social class that includes its cultural values.

Similarly to gender and age, notwithstanding the problem of defining social class and its nature, the present literature review includes studies on the relationship between language and social class independently from the way social class was defined, as this strategy allows the review to be comprehensive. The most significant theoretical contribution to this section comes from sociolinguistics, a discipline that investigated social class quite extensively. A few other studies from corpus linguistics and the study of teaching of English as first language are also presented.

2.4.1 Sociolinguistics

The core of this section consists in the review of Bernstein's theory of code, one of the most important theories on socialisation and its relationship with language (Bernstein, 1962). Bernstein proposed that two ways of coding messages are available in society: a *restricted* code and an *elaborated* code. A code is defined as the set of principles that regulates the processes of a person's verbal planning (Bernstein, 1962: 35). The two codes proposed by Bernstein are different in a number of ways and this difference depends on the purposes for which they are used in society.

The restricted code is the coding modality used when there is a high degree of shared background between the speaker and the listener and where therefore the condition of the listener is assumed to be known. It is a code typically used by speakers with their peers or close family members, for example. For this coding modality, the way something is said is more important than what it is said as the purpose in which this code is used is mostly one of maintaining or building the social relationship in situations in which there is few information to convey.

On the other hand, the elaborated code is a coding modality used when there is greater background distance between the speaker and the listener and where therefore the listener's position or previous knowledge is assumed *not* to be known. In other words, it is a coding modality in which the speakers adapt their speech to the listener and to what they suppose the listener knows. The purpose for which this code is typically used is one of dealing with abstract meanings that have to be made explicit. Thus, for this coding modality the information conveyed is much more important than the task of maintaining or building the social relationship.

Linguistically speaking, the two codes are differentiated by the fact that they have two different levels of predictability. Because of its social purpose, the restricted code is characterised by a limited

range of lexicogrammatical patterns and it is therefore easier to predict than the elaborated code. In its extreme forms, the restricted code would consist in highly ritualised forms that are crystallised for particular meanings.

The two codes are normally used by speakers according to the context and the ability of employing the right one is part of an individual's communicative competence. There is no difference in terms of intelligence or personality in employing one or the other, as empirically verified by Bernstein (1960). In this study, Bernstein looked at the scores produced on IQ and verbal IQ tests from two samples of subjects from middle and working class. The results of the study indicated that a difference between the groups was evident only in terms of verbal IQ but not general IQ. In the lights of this finding Bernstein proposes that 'the *mode* of expression of intelligence […] may well be a matter of learning: in particular, the early learning of speech forms, which create and reinforce in the user different dimensions of significance.' (Bernstein, 1960: 276).

Bernstein suggests that stratification in modern society exists only in terms of social class and not in terms of intelligence. Bernstein proposed that working class individuals are limited in their communicative competence because they rarely have access to the elaborated code. Mastering the elaborated code takes more time than mastering the restricted code and the wealthier and educated classes in our society have therefore an advantage. This advantage is then used to maintain higher positions in society where the elaborated code is more frequently used. By being exposed to the elaborated code earlier, middle class children are more likely to have an advantage in life.

All the above series of hypotheses are clearly theoretical formulations in need of empirical support. Many studies in the second half of the 20th century aimed at confirming or disproving Bernstein's model and some of these studies are reported in this section. Bernstein (1962), for example, tested his model on a data set consisting of interviews produced by 61 working class and 45 middle class male subjects between 15 and 18 years of age. The division in social class was based only on the type of education that the subjects were receiving. The analysis of this data set confirmed Bernstein's initial hypotheses on the type of production that would characterise each social class. The analysis showed that working class subjects used fewer pauses between utterances, thus showing less monitoring and planning of their utterances in the interview setting, and a shorter word length. Statistically controlling for IQ, Bernstein was also able to show that these differences were not accounted for by the IQ of the subjects but by their social class alone. Although the first finding is rather questionable as the correlation between pause length and better planning was not confirmed by psycholinguistic experiments, Bernstein was still able to show that a difference exists at least in word length, and therefore possibly in vocabulary mastering, between the two social classes, as predicted by the code theory.

Hawkins (1969) is a further empirical confirmation of Bernstein's code hypothesis. The author set up a study with the aim of replicating previous findings that showed that working class and middle

class children were different in terms of frequency and ability of producing adjectives, nouns and pronouns. The theoretical contribution of Hawkins on this respect is the understanding that these features are all elements of the noun phrase, also called 'nominal group' in the Scale and Category Grammar, the older version of systemic functional grammar that Hawkins used. With this awareness, the researcher conducted a study to verify whether the analysis of the nominal group can be of help in understanding the phenomenon. The sample gathered by Hawkins consisted in two spoken tasks produced by 124 middle class children and 139 working class children. The social class of the children was inferred using the level of education and occupation of the parents. The children were asked to narrate a story based on a series of pictures and then asked to describe a painting.

When analysing the nominal groups of these spoken texts, Hawkins noticed that working class children were more likely than middle class children to use exophoric reference than anaphoric reference. This phenomenon was explained by the author to be a further confirmation of the code hypothesis, as the working class children, who were more likely not to be exposed to the elaborated code, were coding their language in a way oriented to take for granted that the listener was aware of the pictures in the narrative text and of the painting in the descriptive text. The middle class children, on the other hand, being more expert with the elaborated code, were producing texts that were not placing too much burden on the listener by assuming that the listener was not aware of the context and therefore producing less exophoric pronouns.

Eight years later, Hawkins (1977) published a more thorough study of the same data set where he analysed not only the types of references but also the nominal group as a whole. In general, his finding consisted in the conclusion that there was a greater tendency for working class subjects to use pronouns as opposed to nouns. Consistently with Bernstein's theories of the role of women in working and middle class, working class girls were found to use more hypocoristic adjectives, that is, adjectives to indicate a diminutive or affective meaning. Working class girls were also more likely to use more possessive determiners and rankshifted nominal groups as genitives. Both Bernstein and Hawkins explain this finding as being a consequence of the working class women's greater concern in describing and dealing with relationships. More specifically related to the structure of the nominal group is the finding that uncommon adjectives, Qualifiers and Classifiers were used more often by middle class. The category of Classifiers showed also a significant correlation with the verbal IQ of the subjects. Finally, first person pronouns linked with expressions of tentativeness were a characteristic of middle class subjects.

Hawkins (1977) concluded that the findings indicate a difference in terms of verbal strategies between the two social backgrounds. He defines as verbal strategy the employment of particular sets of meanings to respond to a particular context. In his study, therefore: 'in a situation where children were required to narrate a story from a set of pictures, the working-class children's strategy was oriented to the use of all categories of exophoric reference, and of anaphoric third-person pronouns, at the expense

of nouns. The middle-class strategy was oriented towards the noun' (Hawkins, 1977: 196). This difference in verbal strategies consists in a difference of use of language rather than competence of language. Contrary to some misinterpretations of Bernstein's theories, Hawkins specifies that working class subjects were indeed able to produce nouns or complex nominal groups yet chose not to do so in the context that he analysed. Bernstein's notion of code predicts this behaviour by postulating that social background influences the way individuals perceive the context, in turn then influencing them in producing a particular set of meanings. In Hawkins' data, the difference in terms of social class did not derive from a deficit but from a different interpretation of the context. Whereas middle class children thought they were required to show their knowledge or respond in a way that is objectively clear, the working class children interpreted the task as happening within the relationship between them and the interviewer, and thus influencing them in producing more exophoric references and pronouns. This interpretation of Hawkins' findings was supported by the work of Hasan (1990) who showed that middle and working class mothers in Australia presented different ways of interpreting the context of controlling their child's behaviour at home, with working class mother being more controlling and less explicative than middle class mothers. For this phenomenon, Hasan proposes the term *codal variation* and theorises that this concept can be generalised to other contexts and to other social variables.

In general, Hawkins' study was reproduced by other researchers with conflicting results. Among the successful replications, Johnston's (1977) experiment considered a sample of 18 five year old children for both middle class and working class backgrounds, with the social class being calculated on the basis of the father's occupation (semi-skilled or unskilled jobs opposed to managerial or professional jobs). Johnston also controlled for verbal IQ of the subjects in order to test Bernstein's hypothesis that codes are not a function of intelligence. The narratives produced by the two groups differed significantly in a direction predicted by the literature: working class children used more pronouns and verbs whereas middle class children used more subordinate clauses and higher total number of words for the narrative. A more accurate investigation of the data revealed that middle class children were more likely overall to use noun or noun phrases as Subjects of clauses as opposed to working class children who employed pronouns in this position more often. Johnston (1977: 322) explains that some research on language development provides evidence that noun phrases are initially generated for Objects and only later used as Subjects. Although this could suggest that working class children are a step behind in terms of language development, the author is cautious in clarifying that there is not enough research to confirm this claim. On the other hand, Johnston suggests the possibility supported by Bernstein's ideas that working class children have reached the same developmental level but that they choose to use pronouns as Subjects because of their coding orientation. Since there were not significant findings for verbal variables, which are the loci of propositional meaning, the author concludes by hypothesising that it is perhaps in the referential meanings, whose loci are the noun phrases, that social class differences can be found.

A similar version of his study conducted by Francis (1974) failed to find compatible results. Two groups of subjects from different social classes consisting of 12 boys and 12 girls between 6 and 7 years old did not show any difference in frequency of exophoric references when asked to retell a story originally heard from the experimenter. Francis (1974) concluded that some differences in terms of linguistic variables were found between the two groups, especially in terms of Standard *vs.* Non-standard English forms and syntactic elaboration but no quantification of those is provided in the paper. Similarly, Jenkinson & Weymouth's (1976) replicated Hawkins' experiment in a sample of 30 working class subjects' oral narration of a story. The authors criticised the analytical choices of Hawkins for not considering to ignore certain exophoric pronouns that subjects would have had to almost inevitably produce given the context. Although this criticisms seems to be unfounded, as Hawkins' study involved a high number of subjects from both classes in equal conditions, the researchers' results of the analysis of their small sample 'do not indicate the high level of exophora found by Hawkins in younger children' (Jenkinson & Weymouth, 1976: 109).

Among the empirical tests of Bernstein's theory, Poole (1976) is probably the most thorough empirical verification. After formulating a series of hypotheses on the expected findings predicted by the theory, the author collected spoken and written samples of language from 80 first year university students of an Australian university divided in middle and working class according to the father's occupation and level of education. The researcher gathered spoken data through interview and written data by asking the students to write a 'life-forecast' essay in which the students imagined and described their life in the future. All the samples were examined through a battery of linguistic variables that were proposed in previous literature to account for the difference in restricted and elaborated code. Poole (1976: 84) used 28 variables divided in five categories: structural complexity, language elaboration, verb complexity, personal reference, and linguistic ineptitude (this category only applied to spoken language).

A univariate statistical analysis of all the single variables for spoken language revealed that all the categories presented significant differences in the predicted direction, empirically confirming for this data set that middle class students presented a higher degree of elaboration than working class students. The first function of a discriminant function analysis of the variables divided the two social classes in the direction predicted by Bernstein's hypothesis.

For written language, less unanimous results were obtained. In the univariate analysis most of the variables did not result in significant differences among the groups as happened for the spoken language analysis. However, some of them presented significant results, namely: the Loban weighted index of subordination for the 'structural complexity' category; the ratio uncommon adjectives-adjectives, the frequency of adverbs, and the frequency of unusual adverbs for the category 'language elaboration'. No differences were significant for other categories. These results suggest that written compositions by middle class students seem to present slightly more syntactic elaboration but

significantly more variation in terms of modifiers used. This finding is in line with Bernstein's prediction that the elaborated code is more likely to be less predictable than the restricted code. A discriminant function analysis of the data set resulted in one significant function that clustered the middle and working classes using the following variables: frequency of adverbs, Loban weighted index of subordination, the ratio uncommon adjectives-adjectives, and the ratio I-total personal pronouns. The author concluded that both results give strong support to the code theory proposed by Bernstein, although the differences in production of spoken language were more evident and significant than the differences in production of written language. As some criticisms reported below argue, this difference between the results in the analysis of code for the two modalities could be generated by the fact that written language in general is a contextual configuration that requires to a certain extent the employment of an elaborated code.

It is again Poole that contributes to provide empirical evidence of Bernstein's hypotheses in a study already reviewed in the section on gender. Poole (1979) investigated the relationship between sex, social class and linguistic coding. As shown in the section on gender, the researcher was successful in finding linguistic variables that provided evidence of a code differentiation between the genders. Similarly, Poole (1979) also finds support for the theory of codes for social class. Her sample consisted in interviews structured in several tasks, from questions with images as stimuli to open questions gathered from 96 secondary school Australian students drawn randomly from schools differentiated for indices of social status. The results that the researcher obtained were slightly different and yet compatible with her study conducted three years before. The researcher found that middle class students had a higher score on: mean sentence length, the ratio subordinate clauses-finite verbs, mean pre-verb length, the ratio I-total words and I-total personal pronouns and the ratio Ah-disturbances-verbal tics. Comparing these results to Poole's (1976) previous ones it seems evident that although the variables are different, the basic concept is similar, as the middle class students presented more elaboration. In Poole (1976), however, this elaboration consisted mostly on specifications of nouns or clauses, whereas in Poole (1979) this elaboration is more evident in the syntax. This difference is probably due to the register differences between the two data sets. Another possible explanation for the diverging findings could be the different grades from which the samples were gathered. If that is the case, social class differences would therefore consist not only in one and one only form of elaboration but in the concept itself of elaboration, which is then realised in various ways according to the linguistic maturity of the subject.

Another empirical exploration of Bernstein's theory is Plum & Cowling's (1987) study of social class through the lenses of systemic functional linguistics. The researchers analysed fragments of sociolinguistic interviews gathered in Australia for the Sydney Social Dialect Survey produced by a sample of 24 subjects divided in two age categories (adults and teenagers), three social classes (MC, UWC and LWC) and gender. Plum & Cowling analysed the verbal groups observed in these interviews

with a view to exploring modality as well as to what extent choices such as the selection of present or past tense can be said to be influnced by social background. The authors were clear in stating that the data was not collected with the aim of answering this particular research question and that therefore some of these results could be influenced by the way the interview was conducted. Plum & Cowling (1987) found that social class was stratified according to Halliday's classification of modality in: "always" – "usually" – "sometimes". In turn, these three forms of modality were typically realised in the verbal groups of their sample by, respectively: present tense – modality – past tense. In accordance with Bernstein's theories of the different types of meanings expressed by the social classes, it can be argued that the context of recalling past events and narrating them orally is an occasion for the subjects not to express the same meanings. For lower working class subjects, the researchers noticed the expression of universality realised through a narration using the present tense. Middle class subjects, on the other hand, tried to narrate in past tense, thus possibly conveying the meaning that the narrated events do not necessarily happen for everyone in the way they were narrated. Finally, the upper working class situated itself in the middle of the two by using modality more often to express that the events only "usually" happens in the way they were narrated.

An indirect verification of Bernstein's theories is the work of Labov & Auger (1993) already reviewed in the section on age. Aware of Kemper's research on the syntactic competence decrease in older adults, the authors devised an experiment aimed at determining its causes. Using sociolinguistic interviews of subjects participating in a longitudinal study, Labov & Auger concluded that no difference can be observed in terms of reduction of syntactic complexity in their sample of 12 subjects at different stages of their lives. However, at least for a the sub-sample of 10 subjects from Montreal, the researchers note that 'the social class of the speaker proved to be a consistent and major determinant of complexity' (Labov & Auger, 1993: 121). Judging social class by the occupation of the subjects, the authors noted that there was no overlap of scores between the two sub-groups when counting how many dependent clauses are used per t-unit, with professional workers producing more than lower middle or working class subjects. A slightly less significant result was obtained when counting how many left-branching clauses were produced per t-unit. Although the authors do not attempt to propose an explanation for this finding as their experiment was designed to study language and ageing, this finding indirectly confirms the proposal of Bernstein as it shows a middle class group of speakers producing more elaborate language in a fashion similar to the experiment conducted by Loban 26 years before.

Some evidence for Bernstein's hypotheses being valid also in Hebrew is presented by Berman *et al.* (2011). Although the authors do not cite Bernstein directly, their findings that 80 children of different socioeconomic classes differ significantly in the way they develop their writing style is evidence that Bernstein's theory might also apply to other cultures. The authors found that 'high SES children produce more nouns in general, while low SES children produce more verbs in general'

(Berman *et al.*, 2011: 180). This finding is compatible with Bernstein's predictions and with a number of studies reviewed in this section.

Bernstein's theories mainly influenced the research on language and social class in the sixties. Apart from sociolinguistic research on sociolinguistic alternations, especially for phonetic variables, research on discourse variables or discourse styles has been almost abandoned lately. However, a recent investigation of Bernstein's hypotheses is Macaulay's (2002) analysis of adverb usage and social class. The sample gathered in this study consisted in Scottish subjects from two studies. The first sample consisted in sociolinguistic interviews produced by 12 speakers whereas the second sample consisted in recorded conversations produced by 33 speakers. Both samples were divided in social class on the basis of occupation, education and residence. Macaulay (2002) investigated the use of adverbs in these two samples with the aim of confirming or disproving one of Bernstein's claims that a component of the elaborated code as opposed to the restricted code is a more frequent use of uncommon adverbs. The analysis of the two samples revealed that in general middle class subjects uttered more adverbs and more adjectives than working class subjects.

In the light of these findings the author examined the functions that adverbs had in his data set in order to verify whether Bernstein's hypothesis was valid. The author's conclusion was that differences in discourse styles are evident in the data but that these differences seem to reside in the way stance is expressed. On one hand, the middle class speakers make clear their stance using adverbs and evaluative adjectives in a way that Biber & Finegan (1989) named 'involved, intense conversational style'. On the other hand, working class speakers express their stance by letting the hearer infer it from the details of the content of their message. Macaulay (2002) concludes that these differences do not support Bernstein's theory on how adverbs are used by social classes. However, looking at the findings from a different perspective, this emphasis found by Macaulay on having the hearer infer meaning from the wordings of the speaker is a feature that Bernstein mentioned as being part of the restricted code. Although it is not clear what definite conclusions can be drawn from Macaulay's study, it seems possible to argue that Bernstein's code theory cannot be excluded entirely as an explanation to account for the findings.

Another study carried out by Cheshire (2005) seems to confirm Macaulay's (2002) hypotheses. Cheshire's (2005) work consisted in an analysis of the information structure of a series of sociolinguistic interviews gathered from working and middle class fifteen years old children in three English towns. The researcher analysed the discourse-new entities introduced by the speakers and noted down how these were introduced in the discourse. The main finding of her research consisted in noticing that individuals from middle class, and in particular males, were differentiated by individuals from working class, and in particular females, from the amount of new information introduced in the form of bare noun phrases used in unmarked canonical clauses, with the first group of individuals producing fewer instances. This finding was interpreted by the author as indicating a tendency for male middle class

individuals to be more explicit in terms of reference and for having an orientation to discourse that is more oriented on the hearer, rather than on the speaker, as opposed to working class females. This finding mirrors Macaulay's (2002) conclusions that the typical discourse style of working class individuals consists in presenting the facts to the hearer without making stance explicitly. Cheshire (2005: 498) adds that this difference between orientations is related to the degree of collaboration that is expected by the receiver of the communication, with working class individuals expecting the hearer to draw their own conclusions. Cheshire (2005: 498) mentions research pointing to a similar orientation being present in cultures that have received little exposition to written genres. In English culture, middle class individuals are the social group that is most exposed to written genres and this group therefore is the one that is more likely to present a compatible discourse orientation.

Not all the empirical verifications of Bernstein's code hypothesis were successful in finding an effect. Poole (1973: 108), for example, concluded that 'the factorial organization of linguistic coding abilities for the middle-class group was not more differentiated than that of the working-class group' after comparing the result of two factor analyses on two sets of written data produced by middle class and working class children.

Likewise, Rushton & Young (1975) in a pilot study of fifty 17 years old students' writings found evidence of differences in style but could not confirm the hypothesis that middle class students and working class students were actually distinguished by two types of codes, as the working class subjects could switch from register to register as easily as the middle class subjects. The authors conclude by citing a strand of research that proposes that perhaps the modality of writing is in itself a contextual configuration that demands an elaborated code, and that therefore the coding differences can be only found in speech.

This same opinion seems to be endorsed by Poole (1983) in her meta-analysis of previous studies regarding social class differences in written language production. After thoroughly reviewing thirteen studies for the effect size found and the analytical choices taken, the author concluded that drawing any conclusions as to whether a real effect exists is difficult because of the different methodologies employed by every researcher. Although she noted a small effect size at least in the UK, the author stated that 'there is no clear evidence for a strong association between socioeconomic status and written language' and that there is only 'limited support for a weak relationship between socioeconomic status and written language' (Poole, 1983: 370).

In conclusion, there is some indication that coding differences between classes present a strong effect only for spoken language, as the contextual configuration of spoken language offers the possibility to the speaker to choose between the two codes. In written language, however, the typical contextual configuration already requires being explicit regarding meanings and the effect of the differences between the two classes tends to be small.

2.4.2 Teaching of English

In this section Loban's (1967) survey is reported. As previously seen in the section on level of education, Loban's survey was a significant milestone for quantified studies on English as a first language acquisition. Loban examined the written and spoken language produced by 211 school children for the whole time they spent in formal education from kindergarten to age 18. The sample was stratified for social class using the Minnesota Scale of Parental Occupations, a scale based on the average of the occupation category of the subject's parents. In general, Loban reports

> 'the subjects' socio-economic status to be clearly related to the ratings of their written compositions. […] In every year studied, those in socio-economic groups I, II and III always receive higher ratings on their written compositions than do the subjects in socio-economic V, VI, and VII. Thus the evidence on mean scores makes quite obvious a clear relationship between socio-economic status and proficiency with written language' (Loban, 1967: 102; emphasis in the original)

In terms of linguistic variables, the study shows that the high proficiency group tends to be constituted mostly by children having a high social class background. In other words, the same markers that according to Loban predict level of education are also very good predictors of social class. As shown in the level of education section, these variables consist of the length of communication unit, the elaboration of syntax and the use of subordinating connectives. Loban proposes that this finding might support Bernstein's code hypothesis.

2.4.3 Corpus linguistics

Already reviewed in both the section on gender and age, Rayson *et al.* (1997) analysis of the spoken subsection of the BNC reveals interesting differences between social classes. To differentiate between social classes, the authors adopted the scheme used in BNC based on the occupation of the subjects. The authors then grouped the first three categories' speech and compared it to the other three categories using a key word analysis. The list of significantly different words obtained points to a picture compatible with previously reviewed studies. Lower social classes tended to prefer pronouns in general and in particular third person pronouns, as proposed by Hawkins (1977). The middle class speakers, on the other hand, used more adverbs, possibly used for empathic reasons or for expression of modality, as found by Macaulay (2002). Low social class speakers showed a tendency to report speech by using the verb *say* as well as to use swear words significantly more often than higher social classes. Unfortunately, the authors do not attempt to propose an explanation for these findings.

The final study surveyed in this section is the already reviewed analysis of swear words conducted by McEnery (2006) also reviewed for gender and age. His analysis of the BNC spoken corpus

revealed that there was a significant difference between individuals from different social class backgrounds as categorised within the BNC and their use of swear words. In terms of simple frequency, the number of swear words increased as the ranking of social class decreased. In terms of strength of swearing, a slight alteration to this pattern was noted by McEnery (2006), with speakers of classes A and B, or upper/upper-middle classes, using stronger swear words than the speakers of the class just below, C1, the lower middle class. This phenomenon was explained by McEnery as being a result of hypercorrection. The analysis of the grammatical use of swear words used pointed to the frequent use by lower social classes D and E of all the types except for Predicative Negative Adjectives ('this film is shit'), Literal ('we fucked') and Pronominal ('we got shit to do') type, which were more typical of A and B classes. As explained in the previous surveys, the explanation suggested by McEnery, supported by Bourdieu's 'theory of distinction', is that swear words are a symbol that conveys social information. The classes are therefore indexed by the kinds of swear words they used and the social stratification arises as the individuals in the classes want to convey their membership.

2.4.4 Discussion

This literature survey has pointed out several perspectives on the relationship between language production and social class. In general, however, all the studies considered for empirical exploration for the present work can be summarised under four main linguistic patterns of variation. This classification is presented in Table 2-4 below.

Table 2-4 - Summary of the studies reviewed for social class. Variables in bold represent variables that increase with social classes whereas underlined variables are variables that decrease with social classes.

Study	**Genre of data**	**N of participants**	**Average or min-max text length**	**Year of data**	**Summary of linguistic variables**	**Country**
Pattern 1: Syntactic complexity Higher social classes use more complex syntax. This is likely to be caused by their greater familiarity with complex grammar						
Loban (1967)	Oral and written language samples	N/A	N/A	1967	**average t-unit length; clauses per t-units; frequency of subordinating connectives**	USA
Poole (1979)	Structured interviews	96	N/A	1979	**average sentence length; ratio subordinate clauses per finite verbs**	Australia
Labov and Auger (1993)	Sociolinguistic interviews	10	N/A	1984	**dependents clauses per t-units; left-branching clauses per t-units**	USA
Poole (1976)	Life-forecast essays	80	N/A	1976	**Loban weighted index of subordination**	Australia
Johnston (1977)	Spoken narration with eliciting pictures	36	N/A	1977	frequencies of: **subordinate clauses**	England

Pattern 2: Referential precision Higher social classes show higher precision in referencing entity in discourse than lower social classes. This end is achieved through the use of complex noun phrases. Conversely, lower social classes are more likely to use pronominal forms and exophoric references.						
Hawkins (1977)	Spoken narration with eliciting pictures	263	N/A	1977	frequencies of: **nouns; uncommon adjectives; classifiers; qualifiers;** pronouns; exophoric references	England
Poole (1979)	Structured interviews	96	N/A	1979	**mean pre-verb length; frequency of *I*; ratio *I* per total personal pronouns**	Australia
Macaulay (2002)	Sociolinguistic interviews	45	N/A	1997	frequency of **adjectives**	Scotland
Johnston (1977)	Spoken narration with eliciting pictures	36	N/A	1977	**text length;** frequencies of: **noun phrases as Subject;** pronouns; verbs; pronouns as Subjects	England
Poole (1976)	Life-forecast essays	80	N/A	1976	**ratio of uncommon adjectives per adjectives**	Australia
Rayson, Leech and Hodges (1997)	Casual conversations	N/A	N/A	1990	frequencies of: **second person**	UK

					pronouns; personal pronouns; third person pronouns	
Pattern 3: Use of expletives The use of expletives and their strength varies with social class. Higher social classes are less likely to swear often and/or use strong expletives.						
McEnery (2006)	Conversation	N/A	N/A	1990	frequency of swear words; strength of swear words	UK
Rayson, Leech and Hodges (1997)	Casual conversations	N/A	N/A	1990	frequency of swear words	UK
Pattern 4: Stance types Social classes are different in the types of stance that they select. Lower classes prefer to anchor their statements to the present time and are less likely to express stance overtly. On the other hand, higher classes tend to anchor their statements in the past and to express stance overtly						
Rayson, Leech and Hodges (1997)	Casual conversations	N/A	N/A	1990	frequencies of: **adverbs;** verb *say*	UK
Plum and Cowling (1987)	Sociolinguistic interviews	24	N/A	N/A	frequencies of: **past tenses;** present tenses; modal verbs (upper working class)	Australia
Macaulay (2002)	Sociolinguistic interviews	45	N/A	1997	frequencies of: **adverbs**	Scotland

Poole (1976)	Life-forecast essays	80	N/A	1976	frequencies of: **adverbs; unusual adverbs**	Australia

A first pattern of variation that characterizes social class is the different ability to produce complex syntax. In the studies reviewed in this sociolinguistics Section it has been found that the most promising strand of research comes from the paradigm of codal variation proposed by Bernstein and the difference between elaborated and restricted codes. The empirical findings related to the differences in syntactic complexity seem to be compatible with Bernstein's theory. Many studies were seen to be able to replicate the claims put forward by Bernstein with different degrees of accuracy thus suggesting that there is some truth in the claim that the frequencies in which elaborated and the restricted codes are used is different between the classes.

However, although first studies on Bernstein's codal variation focused on grammatical complexity, the most striking differences in terms of code elaboration and social class were observed in relation to the use of the noun phrase. Similarly to gender, a correspondence was noted between nominal and pronominal styles on one hand and higher social classes and lower social classes on the other hand. This correspondence is noted in the second pattern of Table 2-4, the pattern of referential precision. Indeed, rather than the noun phrase as a whole, it was most commonly found that it is the type of phoric reference commonly used by the two social classes that is seen to vary consistently, with lower social classes using exophoric references more often.

The last two patterns consist of studies conducted within corpus linguistics. On one hand, it was found that a low frequency of adverbs that express stance is a characteristic of lower social classes. Higher social classes are more likely to express stance directly. However, the opposite is true in terms of swear words. When negative stances are present, lower social classes seem to more likely to employ expletives whereas higher social classes tend to avoid these means, as per Bourdieu' theory of distinction.

Even though social class is most of the times poorly defined, the review still suggests that differences between social classes can typically be found, especially when occupation is taken into account. In Chapter 5, the presence of these four patterns is tested in fabricated malicious communication to understand to what extent these linguistic patterns are found in texts simulating typical forensic scenarios. The literature review on social class is the last review of Chapter 2. The next Chapter focuses on the methodologies employed for the collection of the data sets.

3 The data sets

After having presented the literature review of Chapter 2, before applying the knowledge gathered from this review to the analysis of the data sets considered for this study, Chapter 3 explains the methodologies adopted for data collection and subject selection. As described in Section 1.2, this study aims at testing the variables collected through the review of the literature of Chapter 2 on a data set of fabricated texts that simulate typical malicious texts. In order to validate the data set of fabricated malicious text, a second data set consisting of real malicious text was collected. The first two Sections of this Chapter describe these two data sets. For the authentic malicious texts corpus, Section 3.1 describes the method of collection and the basic descriptive statistics of the corpus. In Section 3.2, the fabricated malicious texts corpus and the experiment carried out to create it are described. This same Section also presents the basic descriptive statistics of the set of participants and the definitions of the social factors. The last Section of this Chapter describes some procedures of data manipulation that were performed on both data sets before the analyses.

3.1 The Authentic Malicious Texts (AMT) corpus

The corpus of Authentic Malicious Texts (AMT from now on) is the corpus of authentic forensic texts compiled for the purposes of the present study. This corpus contains **malicious texts**, which were defined in Section 1.1 as *those texts that are a piece of evidence in a forensic case that involves threat, abuse, spread of malicious information or a combination of the above*.

The data collection for this data set was carried out using several resources. The 132 texts of the AMT corpus were collected from (a) printed books in which forensic texts were reported, (b) the FBI Vault, a repository of old case files that the FBI has digitalised in pdf format or image format and made public, (c) private collections of letters gathered by forensic linguists and (d) web searches. A more comprehensive list of all the AMT texts is presented in Appendix 9.1.

While these resources were explored for data, the collection was guided using the knowledge available from previous studies, especially regarding threatening texts. The works of Fraser (1998), Solan and Tiersma (2005) and Shuy (1996) was used to determine whether a text contained a threat. All these linguists who worked on the pragmatic characterisation of threats agree on the fact that that even though there are several syntactic patterns common to threatening utterances, it is difficult or probably impossible to determine whether a text is threatening by the language alone, since it is mostly the extra-linguistic context of the utterance that makes a text threatening. Solan and Tiersma (2005: 204), for example, cite several real cases in which a gesture or act with no language uttered were considered as threats. The linguistic vagueness typical of threats and their extensive linguistic overlap with other speech acts probably arises from the intended vagueness of the threatener who can in this way in a second moment retract the threat. Fraser (1998) described three conditions for the felicitous realisation of a threat that were later expanded by Solan and Tiersma (2005: 198), who also added a fourth condition:

(C1) the speaker has the intention to commit (or have someone commit) an act;

(C2) the speaker believes that the outcome of this act is unfavourable for the addressee;

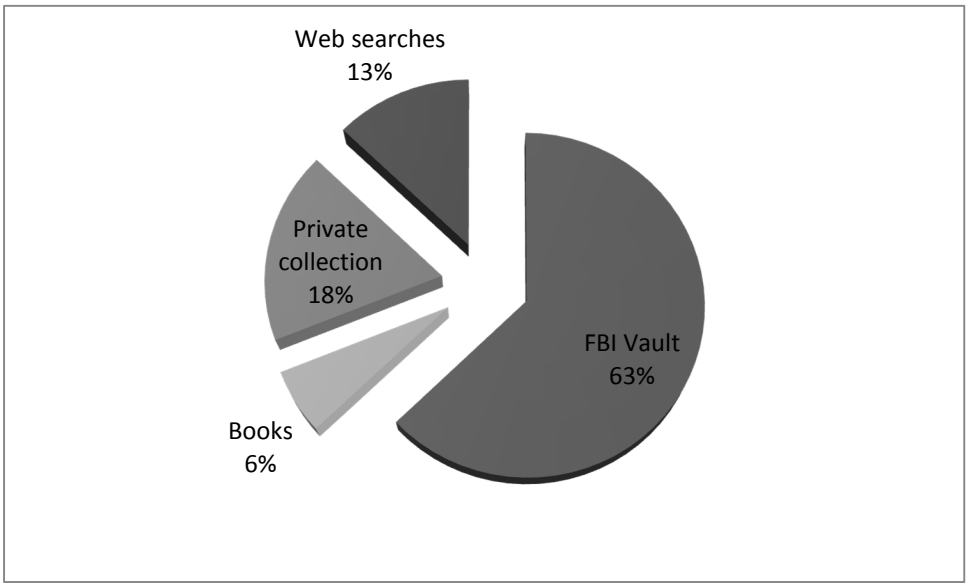
(C3) the speaker intends to intimidate the addressee.

(C4) the act is intended to be taken seriously

These four conditions above were used as guidelines to judge whether a text was threatening, since, even though it is impossible to know with certainty the speaker's intentions or beliefs, it is still possible to speculate on these four conditions given the context of the case. Ultimately, however, it was the context of the case in which the text was involved that determined whether the text was included to the corpus. In general, a text was included in the corpus if it was part of an investigation and/or of a criminal

or civil case that reached the court, and if this text was also abusive, threatening or spreading malicious information (or a combination of these three malicious acts). The breakdown of the contribution of each source to the corpus is represented by a pie chart in Figure 3.1.

Figure 3.1 – Pie chart representing the sources of texts for the AMT corpus



As Figure 3.1 shows, the majority of the texts were collected from the FBI Vault. This repository of file was manually scanned in search of malicious texts and every time a text was found this was copied into a plain text file and its details recorded on a spreadsheet. In order not to skew the corpus, only a maximum of three texts per case were chosen so as to avoid the influence of a single author's style to the overall corpus.

The second most common source of texts was the private collections of a number of forensic linguists operating in the UK and in the USA who were contacted and asked to provide malicious texts for the present research. The forensic linguists that responded provided several texts with different levels of confidentiality that were anonymised and then copied from their original format into plain text files.

Web searches were also carried out with the aim of finding publicly available authentic malicious texts. Using a search engine the following key words were inserted: "threatening letter", "abusive letter", "threatening email". Although more key words could have been added, such as "ransom letter" or "poison pen letter", for reasons of time this was not possible. The 'Images' section of the results of the search engine was also searched for scans of original letters or emails. Every time a malicious text was found, this was copied into a plain text file and its details recorded on a spreadsheet

Finally, the least common category consisted of those texts publicly available from books. Forensic linguistics textbooks such as Olsson (2003) were searched for malicious texts and every time a text was encountered this was copied into a plain text file and its details recorded on a spreadsheet.

The corpus consisted in total of 38,994 tokens, with an average text length of 295 tokens (min: 28; max: 1602; SD: 272.2). However, 27 texts were shorter than 100 tokens and they were therefore dropped from the analysis. This choice was taken as, although most of the features are normalised by text length, the calculation of the relative frequencies is not always accurate if the text is not relatively long. The required length for a variable to be accurately measured varies depending on the rarity of the feature in the language as a whole (Biber, 1993). For very common linguistic items, such as nouns or verbs, even samples of 100 words can be enough to calculate a reliable estimate of their frequency. However, for rarer features such as sentence relatives, it is likely that samples of thousands or millions of words are necessary. Since the average text length for this corpus is already low, a threshold of 100 tokens was selected, as previous empirical studies has shown that multidimensional analysis can be carried out for such short texts (Biber & Jones, 2005) (cf. Section 4.1 below).

The data collection for the AMT corpus focused on English texts only. This corpus was created for the purpose of comparison with the corpus of fabricated texts. Since the subjects who produced texts for the latter were all native speakers of an English variety (cf. Section 3.2 below), an ideal comparison corpus should contain only texts produced by native speakers of an English variety. However, since for the majority of the texts in the AMT corpus the information about the real author is unknown, this aspect of the corpus could not be controlled using reliable methods. In the absence of any reliable system to determine the native language of an unknown writer and in the absence of information on the author, the approach taken for the selection of texts for the AMT corpus was therefore an inclusive one. Of the 105 texts that met the condition of having at least 100 tokens, however, one text presented enough non-standard linguistic features to indicate that it was not produced by a native speaker of any English variety. Even though it was not possible to determine the degree of certainty for this conclusion, this text was nonetheless conservatively discarded.

The final number of texts considered in the AMT corpus was therefore 104, for a total of 36,792 tokens and an average text length of 354 tokens (min: 100; max: 1596; SD: 278.5). The date range of the texts spanned from 1937 to 2013.

3.2 The Fabricated Malicious Texts (FMT) corpus

The corpus of Fabricated Malicious Texts (FMT from now on) is a corpus of experimentally generated forensic texts compiled for the purposes of studying social variation for the present

dissertation. In its final version, the corpus is made up of 287 texts produced by 96 subjects for a total of approximately 87,000 word tokens (average: 302; min.: 97, max.: 994, SD: 108.96). The FMT corpus is sub-divided in three sub-corpora, one for each Task that the subjects recruited for the experiment were asked to write. The experiment task was designed to simulate three scenarios that resemble three malicious forensic texts: Task 1 simulated a formal text of general complaint for a holiday that did not go as expected and it contained a threat of suing the holiday company if a compensation was not received; Task 2 simulated a text addressed to the Prime Minister of the United Kingdom in which complaints about the economic crisis were communicated as well as a threat of not voting for the Prime Minister's party again if nothing was done to change the situation; Task 3 simulated a threat with possibility for abuse from an anonymous employee towards their newly appointed abusive boss. A copy of the Tasks that were given to the subjects is in Appendix 9.2.

When recruited, the subjects of the present experiments were informed that the participation involved filling in a questionnaire with basic information about themselves and a one-hour-and-a-half writing task conducted in an experimental settings in a University room. The subjects were also informed that their time and travel expenses would be compensated with a participation fee of £10. The questionnaire that the subjects filled in is reproduced in Appendix 9.3 and was presented to them at first. The subjects were then presented with the experiment tasks and asked whether they wanted to handwrite or type on a computer. Only 26 of the participants chose to handwrite as opposed to type. The subjects were then told that they could draw from their past experiences for their writing and were reminded that they could take as much time as they wanted. The participants also signed a consent form that explained that the data that they provided to the researchers would be treated confidentially. This consent form is reproduced in Appendix 9.4. In gathering participants, it was attempted to recruit subjects by paying attention to the social factors considered for the study and described in Section 1.2: gender, age, level of education and social class. To achieve this aim, many social groups were sampled, such as: students from a British University, Police Officers in training, homeless individuals supported by a charity organisation, and members of one of the writing groups arranged by a recreational organisation for retired and semi-retired people. Section 3.2.1 below describes the methodology used to measure the social factors in which the FMT corpus is stratified.

3.2.1 The social factors

This Section explains the methodology used to determine the values of these social factors for the subjects that participated in the experiment that generated the FMTs. Furthermore, in this Section the distribution of subjects by social factor is examined in order to test for skeweness of the sample.

The discussion presented in Section 2.1 points out that an important distinction is present between *sex* and *gender*. For the present study it was however chosen to consider only the subjects' biological gender that was reported by the subjects themselves in the questionnaire. This choice was taken for two reasons. Firstly, all the studies considered for replication in Chapter 2 used the biological distinction in sexes. Secondly, for reasons of space and time it was not possible to elaborate a system to identify and classify the social gender of the author to use together with the biological classification. In the rest of the present work, for simplicity and consistency with the previous studies reviewed, the term 'gender' is used. The distinction between social and biological gender is retained theoretically and discussed in the light of the findings of the linguistic analysis.

For the two same reasons for which only biological gender was taken into account, notwithstanding the importance of distinguishing biological/chronological age from social age, only the former was taken into account for the present study. The social factor 'age' simply corresponds to the biological age of the authors self-reported in the questionnaire and measured in years. The reason for this methodological choice is two-fold and consistent with the similar decision for gender: firstly, all the studies reviewed in Chapter 2 only accounted for biological age and, secondly, the scope of the present work did not allow the development of a scale of social age that could be reliably used for the analysis.

Level of education was treated as a three-group categorical variable. In total, three solutions were trialled. The first solution consisted in a five-group categorical variable using the same categories that appear in the questionnaire used with the participants (cf. Appendix 9.3). However, this solution suffered from a lack of cases in the highest and lowest groups. Another solution using a two-group categorical variable was trialled, using on the one hand subjects with a degree and on the other hand subjects without a degree. However, this solution was abandoned since it obscured the differences between undergraduate students or individuals with just an undergraduate degree against individuals with postgraduate degrees. Finally, a three-group solution was adopted and considered satisfying as it allowed to maintain the difference between subjects with a postgraduate degree and subjects with an undergraduate degree only while also maintaining a considerable number of subjects per category. The final categorical variable consisted therefore in three groups: 'below undergraduate', including subjects whose highest education level achieved is below an undergraduate degree (N = 50); 'undergraduate', including subjects whose highest education level achieved is an undergraduate degree (N = 15); and 'above undergraduate', including subjects whose highest education level achieved is higher than an undergraduate degree (N = 28). For three subjects it was not possible to determine the education level as they did not provide this piece of information. These subjects were excluded for the analysis of linguistic variation and level of education.

The last social factor considered for the study is social class. Many ways of describing social class have been employed in several disciplines and Section 2.4 describes at length the complexity of measuring this social factor. For the present work, the categorisation in classes adopted was borrowed from the studies reviewed in the literature surveys. Most of the studies in the survey considered the subject's occupation and/or the subject's parents' occupation as a proxy to their social class. To account for social class in the present study, a Social Class Index (SCI) was calculated using the same class categories adopted in the classification of the data of the British National Corpus and outlined in McEnery (2006: 27). Each of the BNC classes was given a score 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

A special class with score 0 was added for students, since it was not possible to categorise students in any of the categories used for the BNC. The final SCI, however, was calculated by averaging out the score of the parents of the subjects and then averaging out this average with the subject's score as per the formula below:

$$SCI = \frac{\left(\frac{Father\ BNC\ score + Mother\ BNC\ score}{2}\right) + Subject\ BNC\ score}{2}$$

For the students, only the average SCI score of the parents was used. An example of a calculation is displayed in Table 3-1 below.

Table 3-1 – Example of calculation of SCI

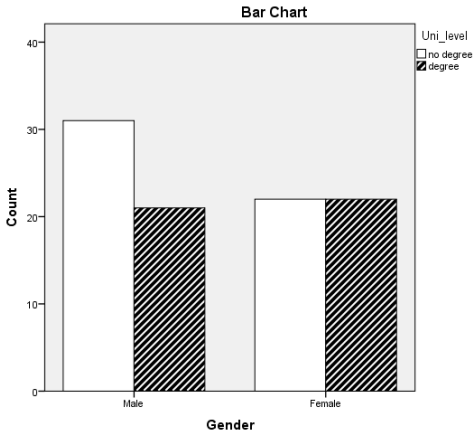
Subject	Occupation	BNC score	Mother's occupation	BNC score	Father's occupation	BNC score	Average family score	SCI
Author1	Factory worker	2	Housewife	1	Driver	2	1.5	1.75

Although many other possibilities could have been selected to measure social class, the SCI adopted for the present study was judged to be the best replication of the classic sociolinguistic measures in the light of the studies surveyed in Section 2.4. Furthermore, since the students do not have any other occupation, by taking into account the family scores it was also possible to include them in the analysis. Although other combinations of occupation or occupation and education could have been trialled, for reasons of space and time this was not possible. Two subjects did not reveal information regarding one or both of their parents' occupation and their social class index was therefore not calculated. These subjects were excluded from the analysis of linguistic variation and social class.

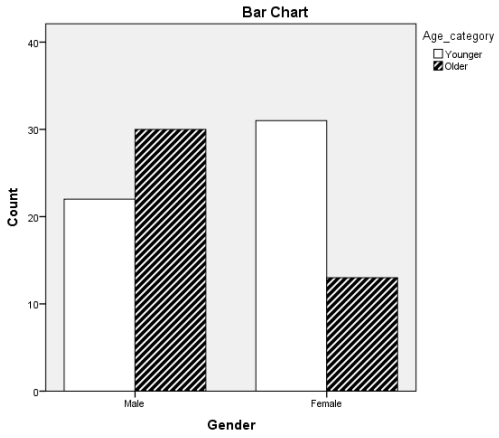
In order to check for skewness in the sample, all the social factors were treated as binary categorical variables and cross-tabulated against each other. Gender is a binary categorical variable by default in this study and was therefore not manipulated for this test. Education was divided in a binary categorical variable as explained above by separating subjects with a degree from subjects without a degree; the two continuous variables, age and SCI, were divided in two categorical variables by splitting the subjects at the median for each variable (age = 38; SCI =3.7). The cross tabulations between these variables are shown in Table 3-2 below.

Table 3-2 – Cross tabulations for all the combinations of the social factors analysed in the present study. A p-value is indicated only for those cross tabulations that presented a significant difference ($p < 0.05$) after a Chi-square test

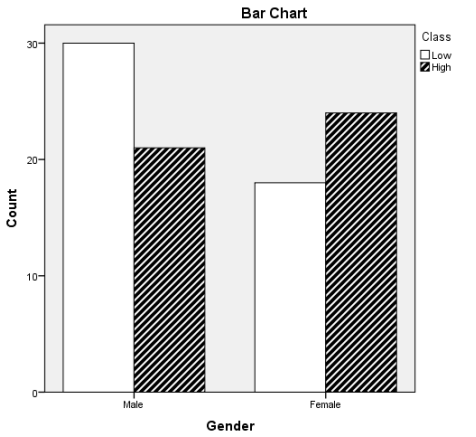
Gender*Education



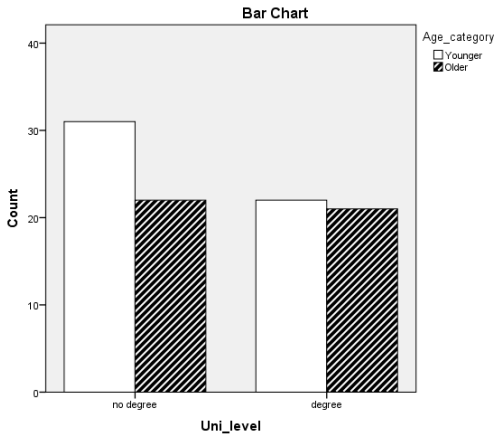
Gender*Age (p = 0.006)



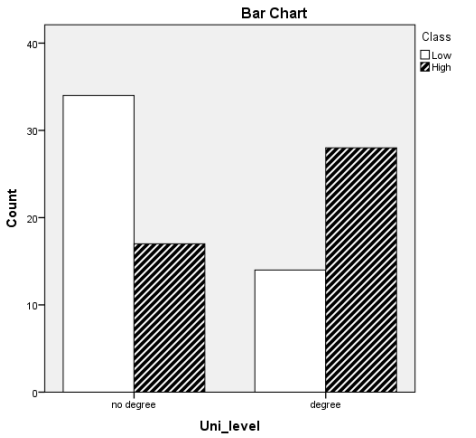
Gender*Social Class



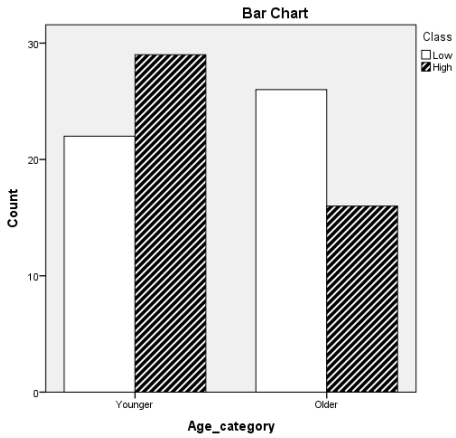
Education*Age



Education*Social Class (p = 0.001)



Age*Social Class



For the six cross tabulations above a Chi-square test was performed. As shown in Table 3-2, the only two skewed social factors in the corpus were age in relation to gender and social class in relation to level of education. This result indicates that the sample suffers a bias in the distribution of authors of different ages across the gender categories and a bias in the distribution of authors of different level of education across social class. The first sample bias constitutes a problem for the study, since most of the young subjects around the age of 20 were females. This bias in the sample is problematic because as the literature review in Chapter 2 has revealed, some variables such as Dimension 1 are affected both by gender and by age at the same time. The limitations caused by this bias are taken into account when the results of the study are interpreted. The second sample bias that can be noticed above is consistent with the nature of SCI. Since SCI was calculated on the basis of occupation, it follows that subjects with a higher SCI are also more likely to have a higher education than subjects with a lower SCI. The test therefore shows that to some extents SCI also takes into account the education of the subjects even though their level of education was not included in the formula for SCI. In conclusions, the analysis shows that a bias in the data was observed only for age and gender, with younger subjects being mostly females. This bias is considered in the analysis and its implications for the results are discussed in details in Chapter 5 below.

3.3 Standardization of the data sets

Before the analysis took place, both the FMT and the AMT corpora were manually scanned for typos and standardised. Although manual intervention is generally undesirable, it was nonetheless judged to be important in the present study since the whole analysis is fundamentally based on automatic processing of data for which any spelling mistake, such as the confusion between *to* and *too*, would result in a tagging error. The process of standardisation was led using a conservative approach and did not affect any case in which the typo was not unequivocally identified as such (e.g. *has* spelled as *as*, *its* spelled as *it's*). A set of further changes was applied to the AMT and FMT texts in order to avoid any problems with the automatic analyses:

1) Formulaic salutations or closings were removed (e.g. *to whom it may concern, kind regards*);
2) Anything written entirely in upper case was transformed to lower case;
3) Elements that were clearly omitted by mistake or distraction were inserted in uppercase (e.g. *it would good - it would BE good*);
4) Any mistaken repetition was deleted (e.g. *you can you can find…*);
5) Any multiple emphatic punctuation was transformed into one punctuation mark (e.g. *you have been warned!!! - you have been warned!*);

6) Single inverted commas were changed to double inverted commas;
7) Lower case first person pronoun *I* was capitalised;
8) In those cases in which a text was handwritten and a word was impossible to decipher, if possible this word was substituted with another one of the same part of speech;
9) The data was fully anonymised;
10) Anything that the author of the text marked as a quotation was removed;
11) Continuous emphatic repetitions of a word were eliminated in order not to alter the counts (e.g. a letter in the AMT corpus repeated the word *mine* more than ten times);
12) The first letters of proper nouns were capitalised if the author did not do so;
13) When the symbol "+" was used by an author in the same way as the coordinating conjunction *and* this symbol was substituted with the word *and*.

Chapter 4 below starts the analytical part of the present work by describing the comparison between the AMT and the FMT corpus with the purpose of verifying their compatibility. If the two data sets are linguistically similar, then the findings of the sociolinguistic analysis of the FMT corpus can be extended to real malicious texts.

4 A comparison between the AMT and FMT corpora

The aim of the present Chapter is to compare the language used in the AMT and FMT corpora with the aim of establishing to what extent fabricated malicious texts are different from authentic malicious texts. The need for comparison between the two data sets arises from the fact that the FMT corpus is made up of fabricated texts, and therefore any finding that is obtained from the sociolinguistic analysis of Chapter 5 is valid and generalizable to real malicious forensic texts only if there is evidence that the experimental conditions have not influenced significantly the language of the texts. The validation of the FMT corpus is carried out in two steps. As a first step, the two corpora are compared to each other using Biber's (1988; 1989) multidimensional analysis framework. Using this methodology it is possible to understand how these two corpora relate to other important genres of the English language and, at the same time, how these two corpora compare to each other linguistically. After this analysis, the two corpora are also compared to each other in order to spot significant differences for the linguistic variables gathered during the literature review presented in Chapter 2. This step is useful as it provides an understanding of which variables are significantly different across the corpora and why. This step can help to reach conclusions regarding why and how linguistic variables vary in one direction or another so that the findings of the sociolinguistic analysis of the FMT corpus can be extended to real forensic data.

4.1 A multidimensional analysis of the AMT and FMT corpora

In the present Section, the multidimensional analysis approach to linguistic analysis is applied to both the AMT and the FMT corpora. Before describing the analysis and presenting the results of this study, the basic concepts of a multidimensional analysis are introduced in this Section. The multidimensional approach is a methodology based on a specific type of multivariate statistics called factor analysis that was introduced by Biber (1988) to study the most important registers of the English language. In this work, Biber (1988) pioneered the use of factor analysis to examine how a number of linguistic variables extrapolated from a general corpus of English co-vary in order to create dimensions of linguistic variations. Biber (1988) found that a combination of six dimensions of variations significantly separate the 23 genres that were considered for his study. These dimensions are:

1. **Dimension 1:** the opposition between Involved and Informational discourse. Low scores on this dimension indicate that a text is informationally dense, as for example academic prose, whereas high scores indicate that a text is affective and interactional, as for example a casual conversation.
2. **Dimension 2:** the opposition between Narrative and Non-Narrative Concerns. Low scores on this dimension indicate that a text is non-narrative whereas high scores indicate that a text is narrative, as for example a novel.
3. **Dimension 3:** the opposition between Context-Independent Discourse and Context-Dependent Discourse. Low scores on this variable indicate that a text is dependent on the context, as in the case of a sport broadcast, whereas a high score indicate that a text is not dependent on the context, as for example academic prose.
4. **Dimension 4:** Overt Expression of Persuasion. High scores on this variable indicate that a text explicitly marks the author's point of view as well as their assessment of likelihood and/or certainty, as for example in professional letters.
5. **Dimension 5:** the opposition between Abstract and Non-Abstract Information. High scores on this variable indicate that a text provides information in a technical, abstract and formal way, as for example in scientific discourse.
6. **Dimension 6:** On-line Informational Elaboration. High scores on this variable indicate that a text is informational in nature but produced under certain time constraints, as for example in speeches.

It is therefore possible, having a text, to determine the scores for this text for each Dimension and thus locate the text in a six-dimensional space made up of these six Dimensions. In this way, texts can be located and compared to each other as well as to the other genres of English that Biber (1988) considered. After constructing this multidimensional space in 1988, Biber (1989) followed up this research by using these same six Dimensions to find out the main text types of the English language, where the term *text type* indicates texts that are maximally similar in terms of their linguistic features.

After applying a statistical technique called cluster analysis, the six dimensions described above were found by Biber (1989) to cluster in eight text types, which therefore represent the main patterns of linguistic variation in his general corpus of the English language. These text types are summarised in Table 4-1 below.

Table 4-1 – A summary of Biber's (1989) text types

Text type	Characterising Genres	Characterising Dimensions	Description
Intimate Interpersonal Interaction	telephone conversations between personal friends	high score on D1, low score on D3, low score on D5, unmarked scores for the other Dimensions	Texts belonging to this text type are typically interactions that have an interpersonal concern and that happen between close acquaintances
Informational Interaction	face-to-face interactions, telephone conversations, spontaneous speeches, personal letters	high score on D1, low score on D3, low score on D5, unmarked scores for the other Dimensions	Texts belonging to this text type are typically personal spoken interactions that are focused on informational concerns
Scientific Exposition	academic prose, official documents	low score on D1, high score on D3, high score on D5, unmarked scores for the other Dimensions	Texts belonging to this text type are typically very technical informational expositions that are formal and focused on conveying information
Learned Exposition	official documents, press reviews, academic prose	low score on D1, high score on D3, high score on D5, unmarked scores for the other Dimensions	Texts belonging to this text type are typically informational expositions that are formal and focused on conveying information
Imaginative Narrative	romance fiction, general fiction, prepared speeches	high score on D2, low score on D3, unmarked scores for the other Dimensions	Texts belonging to this text type are typically texts that present an extreme narrative concern
General Narrative Exposition	press reportage, press editorials, biographies, non-sports broadcasts, science fiction	low score on D1, high score on D2, unmarked scores for the other Dimensions	Texts belonging to this text type are typically texts that use narration to convey information
Situated Reportage	sports broadcasts	low score on D3, low score on D4, unmarked scores for the other Dimensions	Texts belonging to this text type are typically on-line commentaries of events that are in progress
Involved Persuasion	spontaneous speeches, professional letters, interviews	high score on D4, unmarked scores for the other Dimensions	Texts belonging to this text type are typically persuasive and/or argumentative

Using the knowledge presented in Biber (1988) and Biber (1989) it is therefore possible to (1) determine the Dimension scores of a new text or corpus; (2) plot this text or corpus on to Biber's (1988) multidimensional space in order to compare it to the other genres of the English language; (3) assign to

this text or corpus a text type so that it can be compared to other texts or corpora for their general pattern of linguistic variation. This approach is taken in the present work in order to compare the AMT and FMT corpora both to each other and to other genres of the English language.

In order to plot the two corpora on to Biber's (1988; 1989) Dimensions and to assign them a text type, a computer program for linguistic analysis called Multidimensional Analysis Tagger (MAT) was used (Nini, 2014). This piece of software replicates the analysis of Biber (1988; 1989) by using the Stanford Tagger (Toutanova *et al*., 2003) followed up by the application of the algorithms presented in Biber's (1988) appendix to calculate the frequency of the same 68 features used in Biber's (1988) study. The program then plots the analysed text or corpus on to the six Dimensions that Biber (1988) proposed and it assigns one of Biber's (1989) text types to the analysed text or corpus. The reliability of MAT for the present texts was tested before the analysis was carried out. After tagging the AMT corpus, a manual check of a random 20% of the data was performed and the reliability for each text was scored as the number of tagging mistakes divided by the total number of tags. The only mistakes accounted for were those ones that could clearly be identified as such. On average, MAT performed well, achieving an average of 99% correct tags.

The multidimensional analysis of the AMT and FMT corpus is carried out in the present work in two stages: firstly, the two corpora are compared to each other for all of the six Dimensions, with a particular focus on Dimension 1, which is the most important Dimension of variation in the English language (Biber, 1988; 1995); secondly, the two corpora are compared to each other using Biber's (1989) text types.

4.1.1 Biber's (1988) Dimensions of variation in the AMT and FMT corpora

The comparison between the Dimension scores of the AMT and FMT corpora begins with the assessment of the Dimension 1 score. A comparison of the two corpora for Dimension 1 is represented in Figure 4.1 and Figure 4.2, where the two graphs show the comparison between AMT/FMT texts on one hand and other genres of the English language on the other. Rather than presenting all the 23 genres considered in Biber (1988), the present study compares the AMT and FMT corpora only to a subset of seven genres: Conversations, Prepared Speeches, Personal Letters, Professional Letters, General Fiction, Academic Prose, and Official Documents. The choice of these genres for comparison was taken for several reasons: Conversations, Academic Prose and Official Documents were selected as they represent, respectively, the upper and lower bounds of Dimension 1; Prepared Speeches and Personal and Professional Letters were chosen as they are the most comparable genres to the AMT and FMT corpora; finally, General Fiction was chosen as this genre is the most general of the fiction genres considered by Biber (1988) and in this way it is also possible to compare the AMT and FMT corpora to a fiction genre.

Figure 4.1 – Graphs presenting the means and ranges for the AMT corpus compared with the means and range of some of Biber's (1988) genres. The genres, from the left to the right, are: *Conversations, Prepared Speeches, Personal Letters, Professional Letters, General Fiction, Academic Prose,* and *Official Documents*

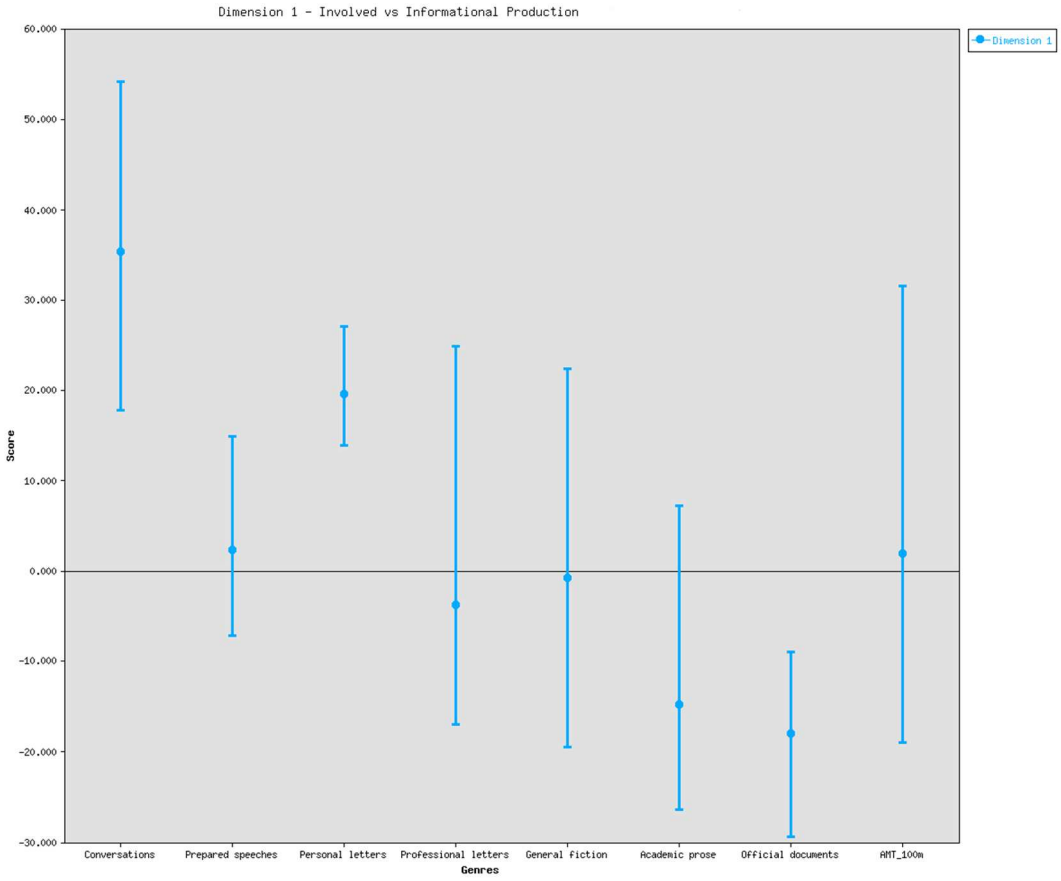
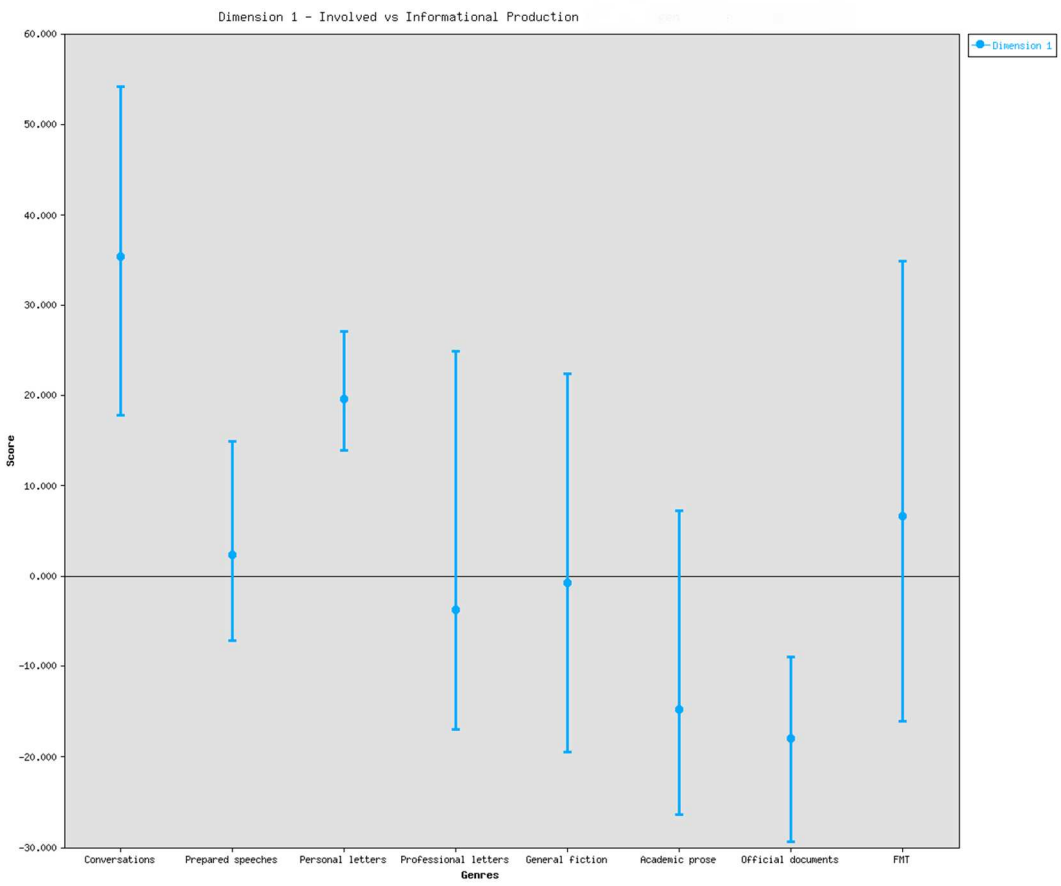
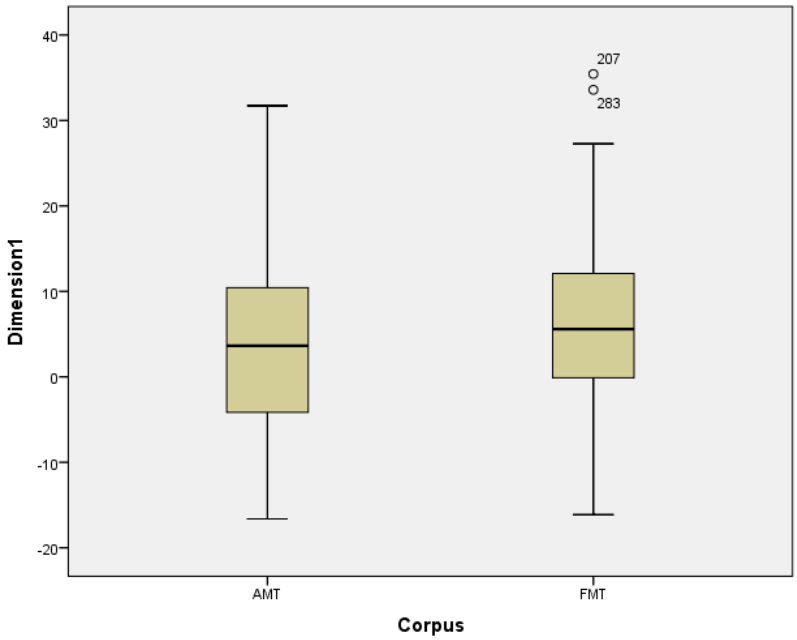


Figure 4.2 – Graphs presenting the means and ranges for the FMT corpus compared with the means and range of some of Biber's (1988) genres. The genres, from the left to the right, are: *Conversations, Prepared Speeches, Personal Letters, Professional Letters, General Fiction, Academic Prose,* and *Official Documents*



The spread of scores for both the AMT and the FMT corpora is larger than for any other genre displayed in Figure 4.1 and Figure 4.2, thus possibly reflecting the fact that the corpora might include one or more sub-genres. The analysis also suggests that both corpora are more Involved than most of the traditional written genres. However, even though most of these texts are abusive and show highly emotional content, it seems that these texts are still more Informational than the typical spoken conversational text. When the AMT and FMT corpora are compared, it appears that the FMT corpus has a greater tendency for texts to appear towards the Involved end of the cline rather than towards the Informational end. A comparison between the two corpora is presented in Figure 4.3 below using boxplots.

Figure 4.3 – Boxplots representing the distribution of Dimension 1 in the AMT corpus



The inspection of the descriptive statistics and the boxplots for Dimension 1 show that although the range is quite high, 50% of scores falls within the area between 10 and -5 for the AMT corpus and within the area between 12 and 0 for the FMT corpus. This difference was not statistically significant using the Mann-Whitney U test. It is possible to conclude that the fabricated data is distributed similarly to the data in the AMT corpus although slightly shifted towards the Involved end of the cline. This shift could however be due to the different communicative situations of the texts. Indeed, the FMT corpus contains an equal number of texts produced in three situations that were controlled in terms of communicative situation whereas the AMT corpus contains a variety of texts that were produced under a large number of different situations.

For further explorations, the FMT corpus was therefore divided in three separate corpora, one for each Task. MAT was run on each separate corpus in order to compare the Dimension 1 score of each Task against the means of the AMT and FMT corpora. The boxplots representing the comparison of the three Tasks to each other for Dimension 1 are reproduced in Figure 4.4 below together with the boxplot for Dimension 1 for the AMT corpus.

Figure 4.4 - Dimension 1 boxplots for the three Tasks of the FMT corpus (left) and the boxplot for Dimension 1 for the AMT corpus (right)

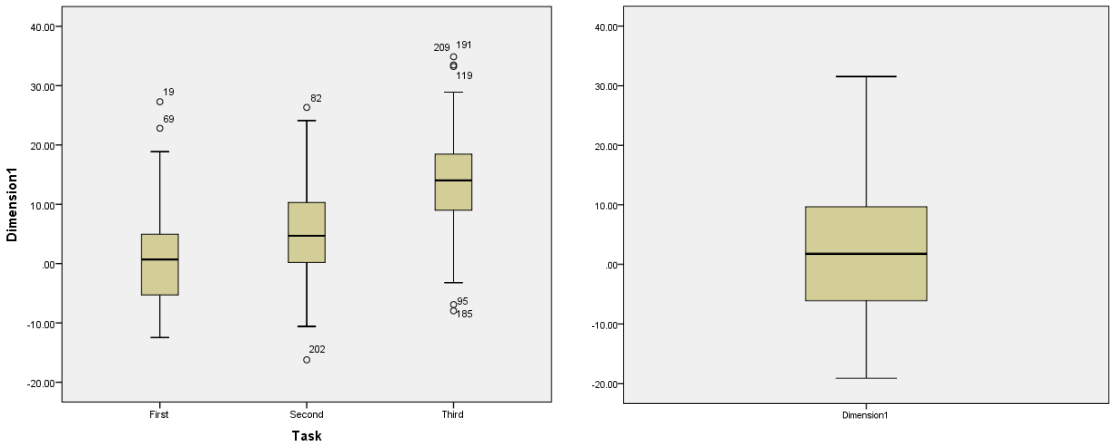
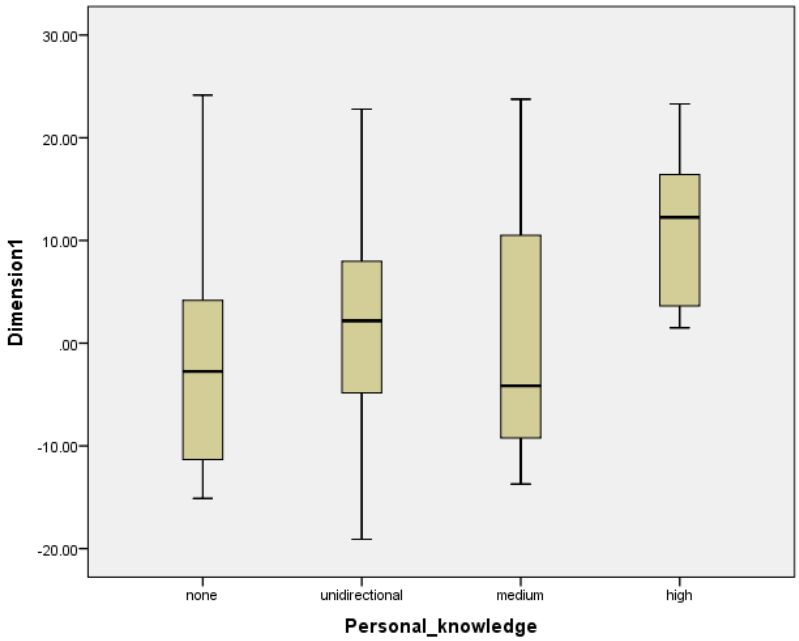


Figure 4.4 does confirm that the slightly more Involved score of the FMT corpus is due to Task 3, the Task that simulates a threatening letter addressed to the boss. Task 1 and Task 2, on the other hand, have very similar Dimension 1 scores that are indeed comparable to the scores of the AMT corpus as a whole. This finding could suggest that in a situation in which the addressee is known it is more likely for a text to be more Involved. To test this additional hypothesis, a further test was carried out: the AMT corpus was manually tagged for the personal knowledge between addressor and addressee and the Dimension 1 scores of these categories were compared to each other. To classify the personal knowledge between addressor and addressee, the extra-linguistic context of the case from which the text was taken was used. The personal knowledge was tagged as: (1) *high* if there was evidence that the addressor and addressee knew each other; (2) *medium* if there was evidence that the addressor and addressee did not personally know each other but nonetheless there was evidence that they were connected to each other by certain people and/or shared a particular environment; (3) *unidirectional* if the addressee was a public figure and there was evidence that the addressor knew them only as such; (4) *none* if there was evidence that the addressor and the addressee did not belong to none of the categories above; and (5) *unknown* if there was no evidence to support any of the above categorisations of relationship from the extra-linguistic context. The AMT corpus presented 23 ‘no personal knowledge’ texts, 52 ‘unidirectional knowledge’ texts, 19 ‘medium’, 7 ‘high personal knowledge’ texts and 3 texts for which the personal knowledge between interactants could not be determined. An independent-samples Kruskal-Wallis test revealed that Dimension 1 was different across these different levels of personal knowledge only when the ‘high personal knowledge’ texts are compared against other texts. This relationship is shown in Figure 4.5 below.

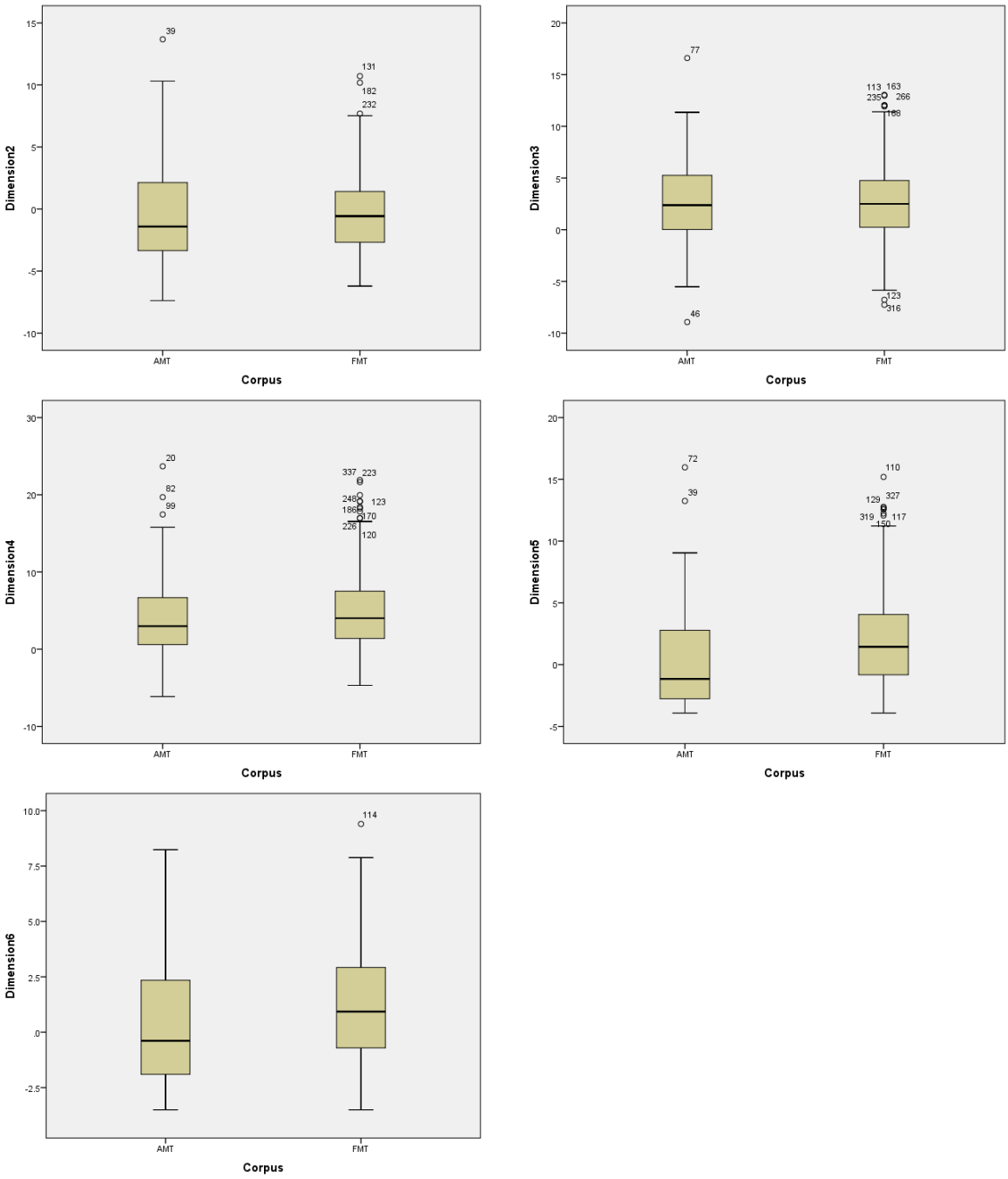
Figure 4.5 – Boxplots representing the distribution of Dimension 1 across the levels of Personal Knowledge in the AMT corpus



This effect is probably due to the fact that less nominal complexity and more pronominal forms are adopted in those texts in which there is a more intimate relationship between interactants and among which therefore there is shared background. The difference between the AMT and the FMT corpus in terms of Dimension 1 is therefore due to the fact that in the AMT corpus there were only seven texts for which there was some knowledge between interactants whereas in the FMT corpus a whole Task, Task 3, was dedicated to this communicative situation. This hypothesis is therefore confirmed by the additional test on the AMT corpus and by the fact that when the personal texts are removed the means and ranges of the two corpora align with each other.

After the analysis of Dimension 1, the most significant dimension of variation in the English language according to Biber (1988; 1989; 1995), the present Section presents a comparison of the Dimension scores for the AMT and FMT corpus for all the other Dimensions. These comparisons are displayed in Figure 4.6 below as a series of boxplots.

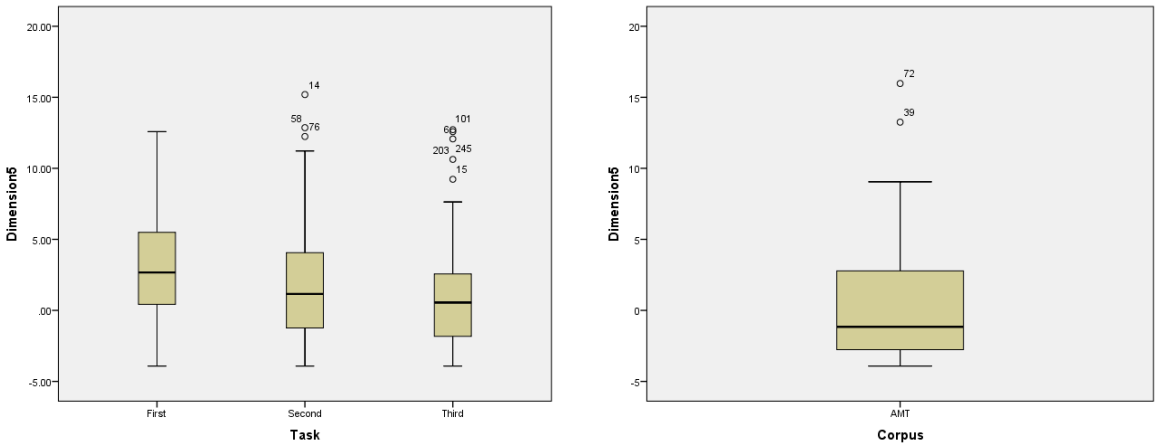
Figure 4.6 – Boxplots representing the distributions of the scores of Dimension 2 (top left), Dimension 3 (top right), Dimension 4 (middle left), Dimension 5 (middle right) and Dimension 6 (bottom left) for the AMT and FMT corpora



For Dimension 2, Dimension 3 and Dimension 4, both the boxplots and a Mann-Whitney U test indicated that there is no significant difference between the two corpora. Linguistically, this means that in terms of narrative discourse, context-oriented discourse and degree of persuasion or modality, there is no difference between the two data sets. Dimension 5 and 6, however, presented a statistically significant difference with the Mann-Whitney U test. Linguistically, this difference indicates that the FMT corpus presented more abstract discourse and more on-line elaboration of information. In order to diagnose the cause of this difference, the FMT corpus was studied independently for these two

Dimensions using the same technique adopted above for Dimension 1, that is, by dividing the FMT corpus in Tasks and studying the score of the Dimensions for each Task compared to the AMT corpus. The results for Dimension 5 are displayed in Figure 4.7 below.

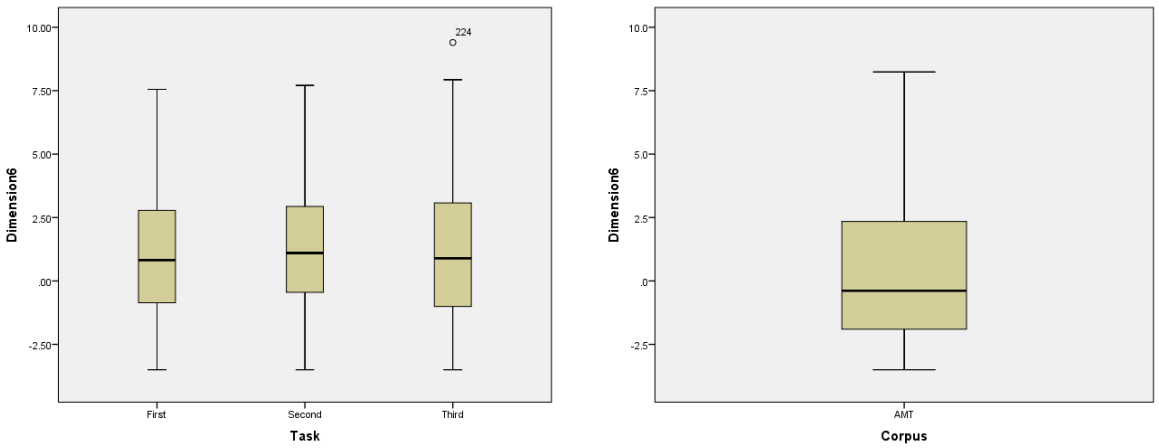
Figure 4.7 – Boxplots representing the distribution of Dimension 5 for the three Tasks of the FMT corpus (left) and for the AMT corpus (right)



The graphs above show that the median score for Dimension 5 for the AMT corpus lies just below zero in a similar fashion to the median scores of Task 2 and 3 of the FMT corpus. An apparently significant difference lies, on the other hand, between the AMT corpus and Task 1. This difference makes linguistic sense since Task 1 is a formal letter of complaint that is more likely to show high scores on Dimension 5 than other texts since Dimension 5 measures the degree of abstract discourse through the use of passive clauses, conjuncts and nominalizations. In conclusions, therefore, the significant difference between the two corpora for Dimension 5 is due to a higher number of formal texts such as Task 1 in the FMT corpus.

A similar comparison is carried out for Dimension 6 in Figure 4.8 below.

Figure 4.8 – Boxplots representing the distribution of Dimension 6 for the three Tasks of the FMT corpus (left) and for the AMT corpus (right)



For Dimension 6, the difference between the corpora cannot be explained by reference to an unequal distribution of situational types as it was shown for Dimension 1 and 5. In the AMT corpus, the median for Dimension 6 is just below zero whereas the medians for all the Tasks of the FMT corpus are above zero. Similarly, the interquartile range for the AMT corpus reaches -2 whereas this is never the case for any Task of the FMT corpus. As such, the cause of this difference is not immediately obvious. However, it should be noted that Dimension 6 was a weak dimension of variation in Biber's (1988) study and it was later abandoned in further studies because the variance that this Dimension explained in Biber's (1988) study was low. This lack of power is probably due to the fact that the variables that load on this Dimension are generally rare and they therefore need large data sets to be studied accurately. The Dimension 6 variables are all post-modifiers of noun phrases that tend to occur rarely in the AMT corpus (the median of all of these variables, that is, *that clauses as adjective complements, that clauses as verb complements* and *that relative clauses on object positions* was zero). Given their rarity, the discussion of this Dimension is abandoned here as the importance is likely to be very limited. In the future, if more data is available, a better assessment of Dimension 6 can be carried out.

The analysis of the Dimension scores therefore shows that the two corpora are similar to each other. Differences were noted in Dimension 1 and Dimension 5 but they could both be explained by the fact that the FMT corpus contains three times the number of texts of the AMT corpus while, at the same time, the AMT corpus contains a greater variety of communicative situations than the FMT corpus. As it was seen for Dimension 1, however, once these parameters are controlled, the results are compatible and they confirm that no important linguistic differences are present in the fabricated texts when they are compared to authentic texts. Further evidence towards this conclusion is given by an exploration of text types contained by the two corpora.

4.1.2 Biber's (1989) text types in the AMT and FMT corpora

In this Section the application of Biber's (1989) text type classification to the AMT and FMT corpora is reported. Using MAT, each text of both corpora was assigned to one of Biber's (1989) text types on the basis of the scores that they presented for the six Dimensions examined above. Similarly, a text type was assigned by MAT to the whole corpus by using the averages of the Dimension scores.

For the AMT corpus, the automatic classification provided by MAT points to the Involved Persuasion type as being the text type of the corpus as a whole. The distribution of text types resulted from the analysis with MAT was: 47% Involved Persuasion, 25% General Narrative Exposition, 9% Informational Interaction, 8% Imaginative Narrative, 7% Scientific Exposition, 4% Learned Exposition, 1% Situated Reportage. The first most common text type is therefore the same text type assigned to the mean dimension scores of the corpus, the Involved Persuasion text type. The Involved Persuasion text type was found by Biber (1989) to be a typical text type for professional letters. This

finding thus implies that almost half of the AMT corpus behaves linguistically as a typical professional letter. The General Narrative Exposition text type is the second most frequent text type of the AMT corpus, even though it is almost half as frequent as the first text type. Biber (1989) explains that the text type General Narrative Exposition is a frequent text type of the English language that involves texts that present unmarked scores for almost all the Dimensions. Finally, 30% of the texts, that is, almost as many as in the second most common category, were classified as belonging to other text types.

Similarly to the AMT corpus, for the FMT corpus the automatic MAT classification also points to the Involved Persuasion type as being the text type of the corpus as a whole. The distribution of text types resulted from the analysis with MAT were: 69% Involved Persuasion; 13% General Narrative Exposition; 11% Informational Interaction; 3% Imaginative Narrative; 3% Scientific Exposition; 1% Learned Exposition. The distribution of text types of the FMT corpus is therefore extremely similar to the distribution of text types found for the AMT corpus. The fact that in FMT corpus 70% of the texts could be classified as Involved Persuasion as opposed to the 50% of the AMT corpus suggests that there is more internal consistency for the corpus. This result reflects the fact that the FMT corpus was deliberately constructed to have a controlled distribution of communicative situations.

Given that the Tasks of the FMT corpus are rather different between each other, the FMT corpus was divided in one corpus for each Task and the text type classification performed by MAT was run for each sub-corpus. For all of the Tasks, MAT assigned the same text type of the whole FMT corpus, the text type Involved Persuasion. The distributions of the text types for all of the Tasks confirm this conclusion and their compatibility with both the FMT corpus as a whole and the AMT corpus. Even though the distributions are very similar, however, it is clear that Task 3 stands out as being different from the other two Tasks. In order to visualise this difference, all the distributions seen so far are displayed in Table 4-2 below, together with Biber's (1988; 1989) Personal Letters and Professional Letters.

Table 4-2 – Distribution of text types for the AMT and the FMT corpora as well as for each Task of the FMT corpus and for Biber's (1989) genres Personal Letters and Professional Letters.

Rank	AMT (N = 104)	FMT (N = 287)	FMT Task 1 (N = 96)	FMT Task 2 (N = 96)	FMT Task 3 (N = 95)	Personal letters	Professional letters
Informational Interaction	9%	11%	10.4%	18.7%	26%	50%	10%
Involved Persuasion	47%	69%	47.9%	55.2%	67%	33%	40%
Imaginative narrative	8%	3%	15.6%	17.7%	4%	17%	0%
General Narrative Exposition	25%	13%	9.4%	0%	2%	0%	20%
Learned Exposition	4%	1%	10.4%	5.2%	0%	0%	30%
Scientific Exposition	7%	3%	6.2%	3.1%	1%	0%	0%
Situated Reportage	1%	0%	0%	0%	0%	0%	0%

As Table 4-2 indicates, the AMT and the FMT corpora are extremely compatible with each other in terms of text types, as they present almost exactly the same distribution. However, when the FMT corpus is divided into its three Tasks, it is possible to notice that Task 3 is different from all the other Tasks and from the AMT corpus. For this Task, the virtual absence of text types that typically belong to formal written genres as well as the strong presence of Informational Interaction as a common text type indicate that Task 3 is linguistically different from the other two Tasks. By comparing the data sets explored in the present work with Biber's (1988; 1989) Personal Letters and Professional Letters it becomes clear that whereas the AMT corpus and Task 1 and 2 of the FMT corpus are more similar to Professional Letters, Task 3 of the FMT corpus is closer to approaching a Personal Letter.

The text type analysis therefore confirms that at a general level the AMT and the FMT corpora are not different from each other. The most significant difference between the two data sets arises from the fact that Task 3 is more similar to a Personal Letter than the majority of the texts of the AMT corpus. At this general level, however, no linguistic difference seems to be due to the fact that the FMT texts are fabricated as opposed to genuine. In order to perform the most accurate test as possible, in the next Section all the variables gathered from the literature review of Chapter 2 are tested for significant difference between the authentic and fabricated data sets.

4.2 The variation of the linguistic variables across the AMT and FMT corpora

After confirming that at a general level there is no important linguistic difference between the fabricated and the authentic texts, the investigation moves to a more fine-grained analysis that concerns the variables that are considered for the sociolinguistic study. The aim of this analysis is to uncover and explain any difference between the two data sets and to test whether even at this level of detail there still are no important differences between the two corpora. All the 141 variables summarised in Appendix 9.6 (excluding the six Dimension scores already explored above), including the ones considered for the sociolinguistic analysis (cf. Introduction to Chapter 5 below for more details on the variables used for the sociolinguistic analysis), were tested for significant difference across the two corpora using an independent-samples Mann-Whitney U test. Even though a variable is traditionally considered to have a statistically significant difference between two groups when the p-value associated with the statistic is lower or equal to 0.05, given the fact that 135 comparisons have to be carried out, this p-value was corrected using the Bonferroni correction to avoid Type I errors. Since the p-value of 0.05 divided by 135 results in an incredibly small number, for the present analysis a less small but still conservative p-value of 0.001 was set as threshold for statistical significance. The results of the statistical tests are displayed in Table 4-3 below, where the variables are organised by linguistic category.

Table 4-3 - Variables for which a significant *Corpus* effect was observed using an independent samples Mann-Whitney U test. The corpus with a higher score for the variable is identified within parentheses

Forms correlated with Involvedness (Biber 1988)
contractions [AMT]
amplifiers [FMT]

Pronominal variables
first person pronouns [FMT]
third person pronouns [AMT]

Nominal variables
singular proper nouns [AMT]
plural proper nouns [AMT]
predicative adjectives [FMT]
genitives [AMT]

Past tense variables
past participles [FMT]
perfect aspects [FMT]

Other
conjuncts [FMT]

Verbal variables
that as verb complement [FMT]
suasive verbs [FMT]

Out of the 135 linguistic variables tested, only 13 were significantly different across the two corpora. The rest of this Section explores these 13 variables more thoroughly to find out explanations as to why these differences were present.

For the first group of variables, contractions and amplifiers, the distribution is displayed in Figure 4.9 below.

Figure 4.9 – Boxplots representing the distribution of amplifiers (left) and contractions (right) in the AMT and FMT corpora

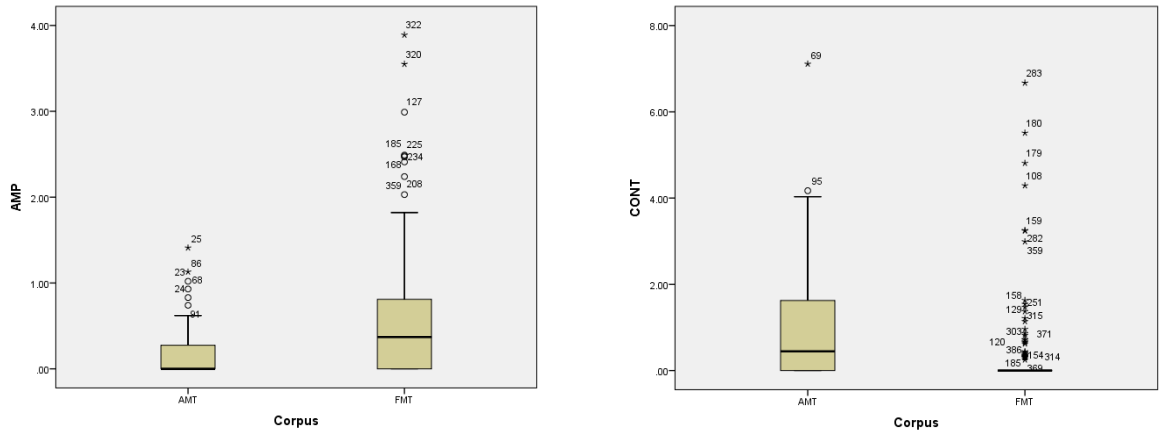
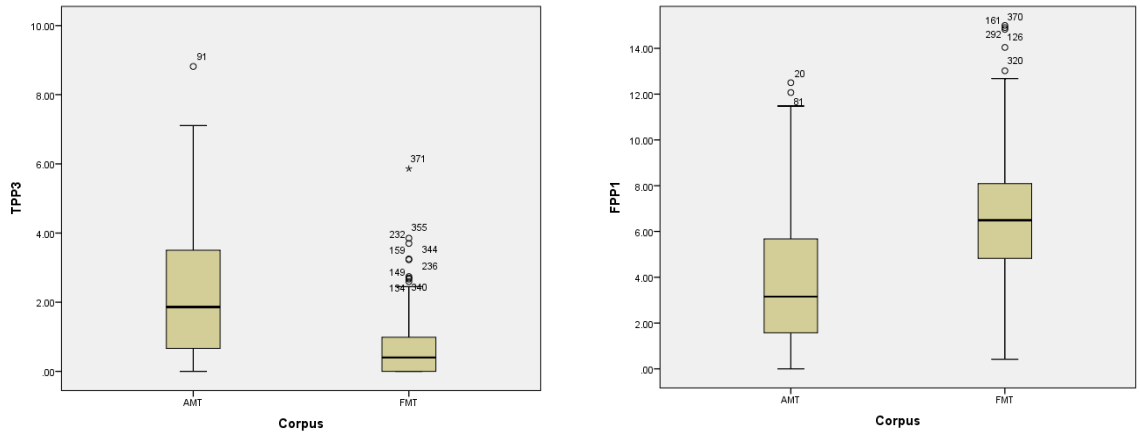


Figure 4.9 above suggests that contractions are more common in the AMT corpus whereas amplifiers are more common in the FMT corpus. However, in both cases one of the corpora presented a median of zero, thus indicating that these features are rather infrequent. Given this limitation, it is difficult to understand why there is difference between the two corpora. However, few hypotheses can be proposed for these differences. In terms of the difference in contraction patterns, it is likely that the experimental condition of the FMT texts is responsible for the difference. In AMTs, the data suggests that we can expect roughly one contraction every two hundred words. In the FMTs, however, even though more texts were available, there were almost no contractions. It is possible that the experimental conditions of the FMT Tasks prompted the subjects to avoid contractions as their language would have been under analysis. In terms of the difference in frequency of amplifiers, the explanation for the difference could be the same reported in Section 4.1.1 for Dimension 1, since this feature contributes to Involved Dimension 1 scores. A higher incidence of amplifiers in the FMT corpus would therefore be related to the fact that the FMT corpus contains more personal texts than the AMT corpus.

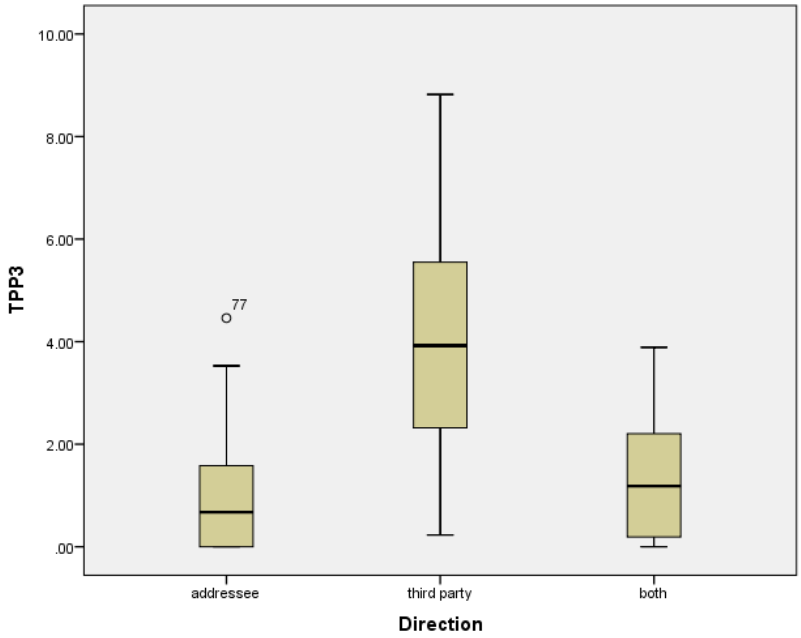
For the pronominal variables, Figure 4.10 shows the distribution of first and third person pronouns for the AMT and FMT corpora.

Figure 4.10 - Boxplots representing the distribution of third person pronouns (left) and first person pronouns (right) in the AMT and FMT corpora



The boxplots above confirm that AMTs were more likely to employ third person pronouns than FMTs. This characteristic of the AMTs is likely to be due to the number of texts that had harmful content directed to a third party rather than to the addressee, as for example, in texts sent to the boss of a company to spread rumours about an employee of a company. To test this hypothesis, the AMTs were manually tagged for the direction of harmful content. The texts in the AMT corpus were assigned to three categories depending on the direction of the threat, abuse or malicious content: *towards the addressee* of the letter (N = 26), *towards a third party* (N = 40) or *towards addressee and third party* (N = 28), such as in cases in which the violent action involved the addressee and their family or in cases in which the threat was directed to the addressee and the violent act to a third party (e.g. “give me the money or I will kill your daughter”). For 10 texts of the AMT corpus it was not possible to classify the direction of harmful content for lack of context. After the manual tagging, the distribution of third person pronouns was checked for statistically significant differences using the Kruskal-Wallis test. The distribution of third person pronouns was indeed significantly different across the categories of direction ($p < 0.001$), as the boxplots in Figure 4.11 confirm.

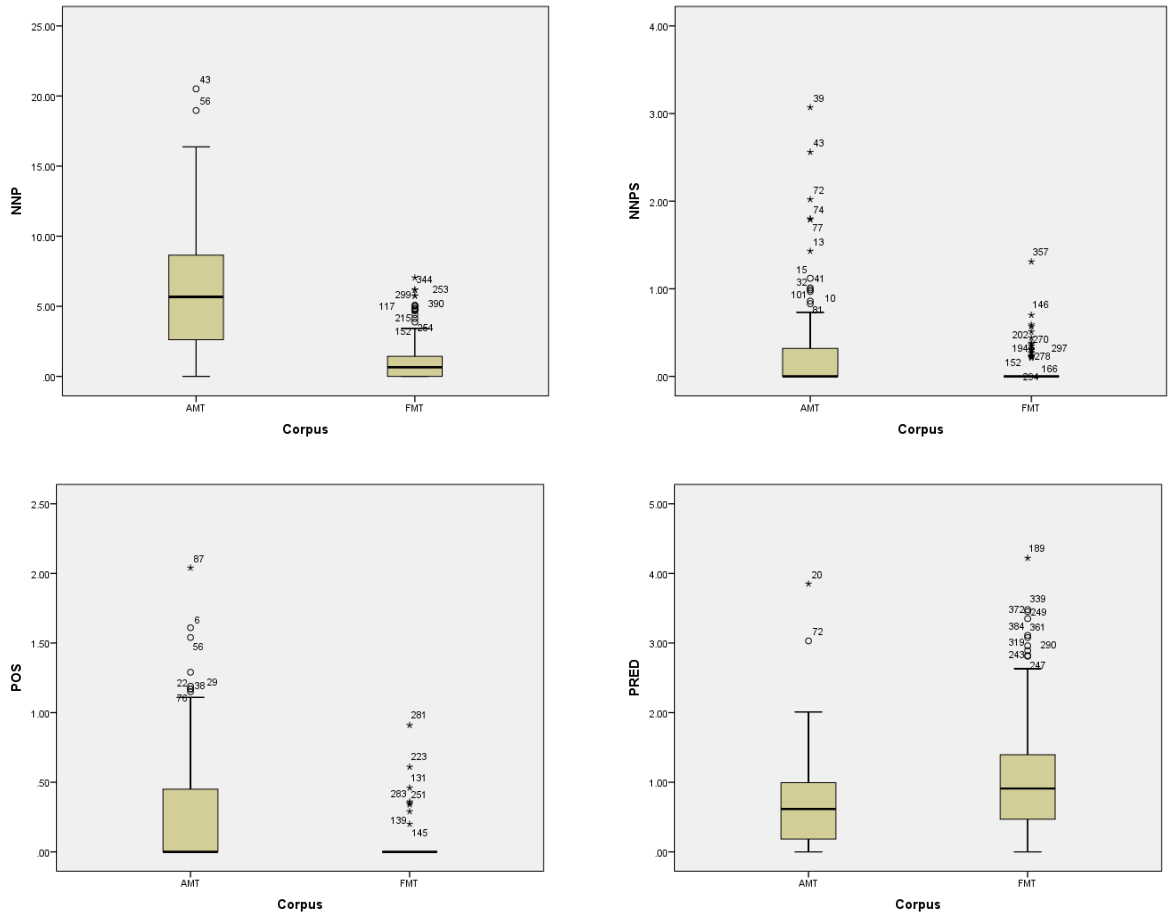
Figure 4.11 - Boxplots describing the distribution of third person pronouns (right) for the direction categories in the AMT corpus



Since almost half of the texts in the AMT corpus were indeed directed to a third party and since no text of the FMT corpus contained harm directed to a third party, it seems reasonable to assume that the difference in terms of pronominal distributions for the two corpora depend on the difference in communicative situations. Indeed, the medians for the AMT texts classified as directed to addressee or to both addressee and third party are compatible with the median for the FMT corpus. The different distribution of first person pronouns can similarly be attributed to the different communicative situations. In Task 1 of the FMT corpus, many subjects chose to recount the holiday from their personal perspective. Furthermore, the higher incidence of first person pronouns is compatible with the fact that the FMT corpus contains Task 3 texts, since first person pronouns is a variable that increases the Dimension 1 Involved score.

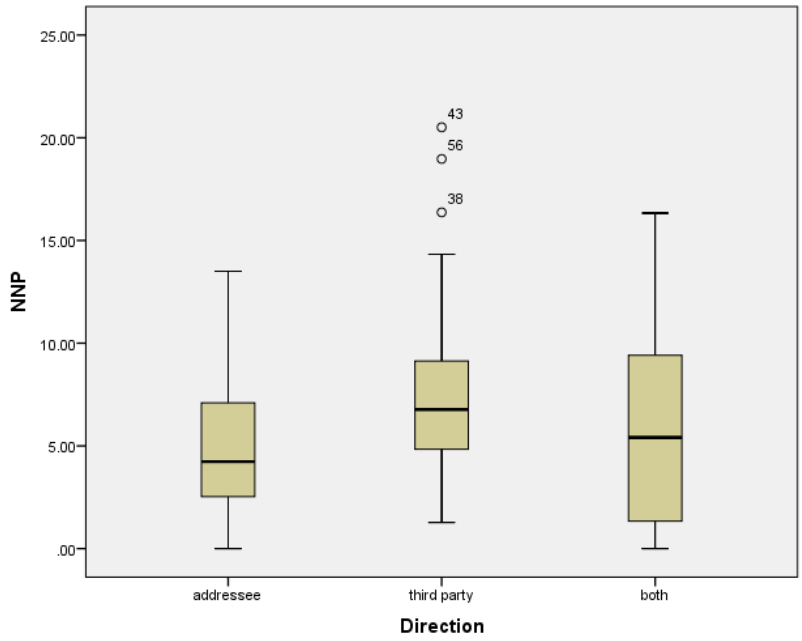
For the nominal variables, the AMT corpus was found to use more proper nouns and more genitives, as shown in the boxplots in Figure 4.12, whereas in the FMT corpus there was a higher incidence of predicative adjectives.

Figure 4.12 - Boxplots representing the distribution of singular proper nouns (top left), plural proper nouns (top right), genitives (bottom left) and predicative adjectives (bottom right) in the AMT and FMT corpora



In terms of proper nouns and, consequently, genitives, the difference between the corpora could be attributed to the experimental character of the FMTs. The higher incidence of references to proper nouns is a strategy that is often used in the authentic texts, especially in the ones in which the harmful content is directed to a third party. Indeed, the boxplots in Figure 4.13 confirm this hypothesis by showing that the texts addressed to a third party are more likely to use proper nouns.

Figure 4.13 – Boxplots representing the distribution of singular proper nouns across the direction of harm categories of the AMT corpus

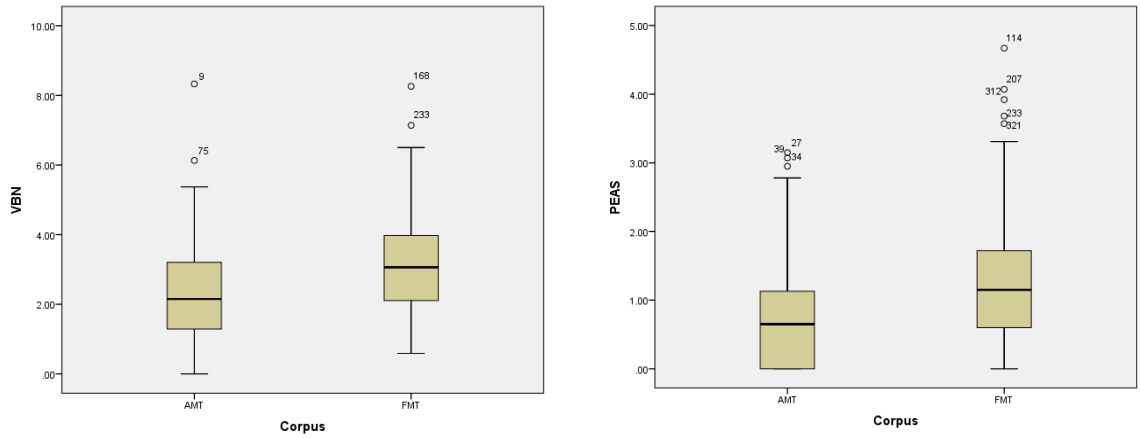


When comparing the medians between the FMT corpus and the AMTs with harmful content not addressed to a third party, however, a difference is still evident since in the FMT corpus the median frequency of proper nouns was almost zero. In summary, it cannot be excluded that the difference in frequency of proper nouns is the result of the experimental situations of the FMT texts.

Compared to the difference in distribution of proper nouns, the different distributions of predicative adjectives is less easy to explain. The FMT corpus had significantly more predicative adjectives than the AMT corpus. Compared to other variables, however, the magnitude of the difference is rather small and therefore the investigation of this feature is not pursued in this work.

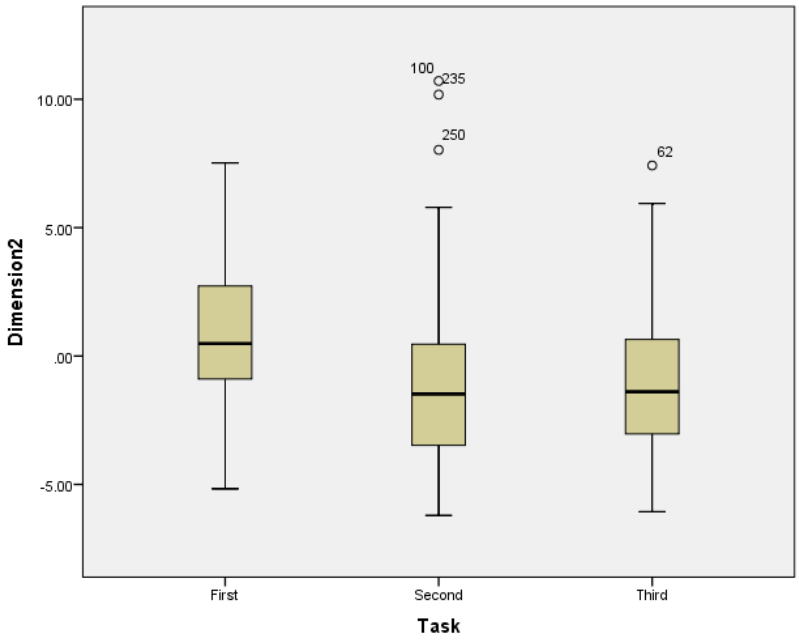
Finally, the difference in distribution of past tense forms shows that for both variables FMTs tended to be slightly more oriented towards the past, as can be seen in Figure 4.14 below.

Figure 4.14 - Boxplots representing the distribution of past participles (left) and perfect aspects (right) in the AMT and FMT corpora



The higher incidence of past forms in the FMT corpus is in all likelihood attributable to Task 1 texts, as they all showed a certain degree of narrative discourse. This hypothesis can be tested by looking at the distribution of Dimension 2, the Dimension of narrative discourse, for the FMT Tasks.

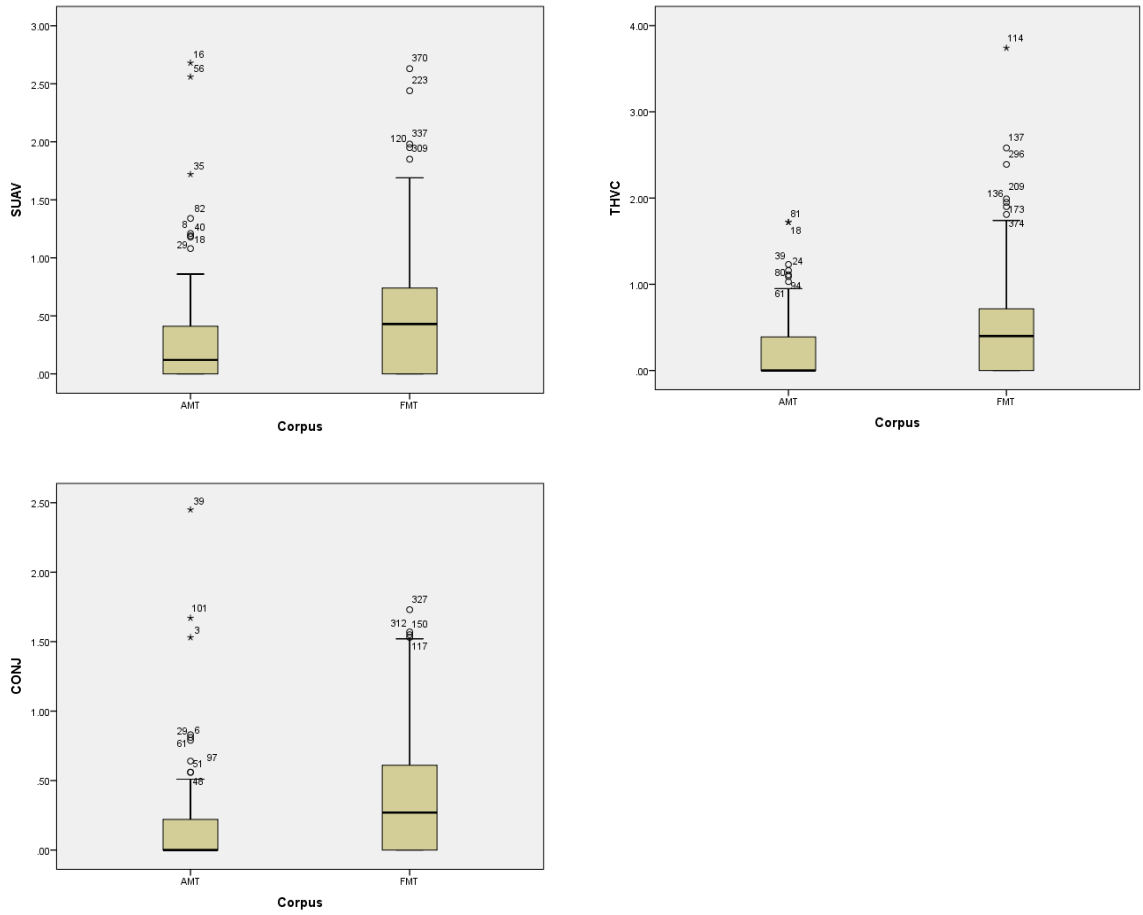
Figure 4.15 – Boxplots representing the distribution of Dimension 2 across the Tasks of the FMT corpus



The boxplots confirms that Task 1 texts tend to have a more narrative discourse than texts for the other two Tasks. This effect is due to the fact that in Task 1 many subjects decided to recount the events that happened in the holiday while complaining about it. As such, the difference between the two corpora is attributable to the different distributions of communicative situations.

The last two categories of variables were joined together for the last discussion and the boxplots of their distributions are visible in Figure 4.16 below.

Figure 4.16 - Boxplots representing the distribution of suasive verbs (top left), *that* as verb complement (top right), and conjuncts (bottom left) in the AMT and FMT corpora



The higher frequency of suasive verbs is correlated with the higher frequency of *that as verb complements* since suasive verbs (e.g. *ask, arrange, command, decide, demand*) are very likely to be followed by a complement clause introduced by *that*. Since these two variables show the same patterning in the boxplots, it is likely that they are connected in such a way. The reason why FMTs used more suasive verbs might connect to the fact that all the three Tasks of the FMT corpus included a request of some sort whereas not all the AMT texts presented requests. In general, however, the rarity of these variables and the small magnitude of the significance do not require further investigations for the scope of the present work.

Finally, the last feature, the frequency of conjuncts, is distributed differently across the corpora since it is a feature that contributes to a high score on Dimension 5. The significant difference between the two corpora for Dimension 5 was already explored in Section 4.1.1 above.

In conclusion, the analysis of the linguistic variable suggests that no important differences are present between the two corpora. Only 13 out of 135 variables presented significantly different distributions between the corpora and most of these variables were rather infrequent. For some of them, such as first and third person pronouns or for the past tense variables, the explanation for the differences

lies in the difference of range of communicative situations between the AMT and the FMT corpora. Nonetheless, for two variables, that is, the frequency of proper nouns and the frequency of contractions, it cannot be ruled out that the experimental settings are responsible for the difference.

4.3 Discussion and conclusions

This section presented the results of the analysis aimed at exploring the linguistic differences between the AMT and FMT corpora to assess whether there is any significant difference between fabricated and non-fabricated data. The aim of this comparison is the validation of the FMT corpus for further analysis of social variation. For this purpose, two analyses were carried out: the application of the multidimensional analysis originally carried out by Biber (1988; 1989) and a comparison of the two corpora for all of the linguistic variables found in the literature review of Chapter 2. The results of both analyses suggest that the fabricated texts are not significantly different from the authentic texts and that therefore the results of the analysis of the FMTs for social variation can be extended to authentic forensic texts.

The first analysis of this Chapter concerned the comparison of FMTs and AMTs on the Dimensions and text types proposed by Biber (1988; 1989). The analyses indicated that both fabricated and authentic malicious texts fell within the same area of Dimension 1 and could be classified within the Involved Persuasion text type, the text type commonly adopted by professional and personal letters. However, further explorations of the AMT corpus indicated that the personal knowledge between the writer and the recipient is a significant factor that affects Dimension 1 and, consequently, the text type of the texts. When the personal knowledge between the interactants decreased, a highly significant difference in Dimension 1 was noted. This effect was the most significant source of variation between the two corpora. Indeed, the FMT corpus is almost three times bigger than the AMT corpus and it contains almost 100 texts that simulate a communicative situation in which interactants personally know each other. Apart from this difference in distribution of communicative situations, no differences were found that could be due to the experimental conditions of the FMT corpus at this level of generality.

The second analysis went further down the level of generality by looking at the differences in distribution for all the linguistic variables. Of the 135 comparisons performed, only 13 were found to produce highly significant results. An in-depth analysis of these 13 significant variables suggested that only two variables might present differences that can be connected to the experimental conditions of the FMTs: the frequency of contractions and the frequency of proper nouns. For these two variables, therefore, it might be not possible to generalise any social variation found to real forensic texts. However, more generally, for all the other variables no evidence has been found that the experimental conditions of the FMT corpus has influenced the subjects to produce language that is significantly different from the language encountered in a typical authentic malicious forensic text. In conclusions,

therefore, any result of the analysis of sociolinguistic variation for the fabricated data can be extended to real malicious forensic texts. This sociolinguistic analysis is described in Chapter 5 below.

5 Sociolinguistic analysis of the FMT corpus

Since Chapter 4 has shown that the language of the fabricated malicious texts is not drastically different from the language of the authentic malicious texts, this Chapter can now move the focus of the dissertation to the FMT corpus and to the four social factors gender, age, level of education and social class and on how these social factors affect the linguistic variables gathered from the literature reviews of Chapter 2 and listed in Appendix 9.6. Before describing the analysis, a few remarks on the methodology of the analyses reported in this Chapter are discussed.

A first point concerns the linguistic variables considered for analysis. The literature review of Chapter 2 generated 132 linguistic variables that are summarised in Appendix 9.5. However, of these 132 only 67 variables could be analysed in the present study for two reasons: firstly, some variables required considerable manual intervention and would have therefore required an extensive work of formalisation of objective rules and application of those rules to the texts that is outside the scope of the present work (e.g. identification of left-branching clauses; abstractness of nouns; measures of cohesion); secondly, some of the automatic variables were excluded because they were rarely occurring in the corpora (e.g. frequency of *its*, uncommon adjectives, types of expletives). Appendix 9.5 below lists which variables were excluded and which variables were kept for the final study.

Another remark is related to the design of the analysis. The aim of the study is to test to what extent the linguistic patterns found in Chapter 2 are also found in the FMT corpus. To reach this aim, a simple strategy would be to take each linguistic pattern listed in Chapter 2 and test whether it is found or not in the FMT corpus. However, this strategy can be strengthened using a more thorough approach that consists in testing all of the variables gathered for all of the social factors. The reason why this strategy is a stronger test is that this method allows a more comprehensive assessment of the presence/absence of the patterns. If one linguistic pattern observed in the literature review for one social factor turns out to be present also in the FMT data set while, at the same time, a large number of other linguistic variables do not present significant differences, the validity of said pattern is greatly reinforced. Furthermore, this more comprehensive analysis allows the discovery of other similar linguistic patterns that might be related to the ones observed in previous literature even though manifested through different linguistic variables. In Appendix 9.6 it is possible to find a description of all the variables used for the analyses reported in this Chapter. This list includes (a) the 67 variables gathered from the literature review of Chapter 2 and (b) the variables calculated by MAT for the comparison analysis of the AMT and FMT corpora. These 141 variables were analysed using parametric or non-parametric tests of significance, depending on whether a variable was, respectively, normally or non-normally distributed. Depending on the social factor, different types of tests were carried out: for the categorical variables gender and level of education an ANOVA was performed whereas for the continuous variables age and social class a correlation test was performed. A significant effect was

noted when the two-tailed p-value resulting from the test was equal or lower than 0.05. However, in those case in which the effect was predicted by the literature a one-tailed p-value was considered instead. Even though for each social factor a total number of 141 comparisons was performed, for this analysis a Bonferroni correction was not applied. The reason for this choice lies in the fact that the present study is interested in verifying whether it is the patterns of variation that are present rather than whether any one of the single variables presents a significant effect. The exclusion of those variables that do not reach a conservative Bonferroni-corrected threshold of 0.001 limits the conclusions that can be drawn in the light of the literature review. Given the limited sample size, if, for example, only very few variables turn out to have a significant effect at this conservative p-value, then it is very difficult to discern any pattern of variation and to conclude whether it matches or not previous findings. It is for this reason that an inclusive approach was chosen for the present work.

The analysis for each of the social factors was carried out both for the whole FMT corpus and for each Task treated separately. This method was employed as it allowed to isolate register variation from social variation as well as to determine which variables still retain a social effect even when register variation is not controlled. The analyses of this Chapter are divided in Sections, with one Section dedicated to one of the four social factors. These results in turn lead towards the conclusions that can be drawn regarding the application of this work to forensic purposes in Chapter 6.

5.1 Gender

Table 5-1 below summarises all the variables that had a gender effect by showing the variable significance levels as well as their effect size, in case of normally distributed variables. The presentation of the variables is organised according to the three major patterns introduced in the literature review of gender in Section 2.1. All the variables that presented a statistically significant difference were organised in these categories, even in the cases of variables that were not gathered from the literature review of gender. The variables that did not fall in any of the patterns for gender are categorised in the general category 'Other variables'. Because gender was treated as a binary categorical variable, the test used was a t-test for normally distributed linguistic variables and a Mann-Whitney U test for non-normally distributed linguistic variables.

Table 5-1 - Linguistic variables that presented a significant effect for gender, showing: p-value ('1-t' indicates a one-tailed value); Cohen's *d* for the normally distributed variables only; the gender for which the variable has an advantage.

Variable	Task 1	Task 2	Task 3	Whole corpus
Pattern 1: Rapport/report orientations On average, males prefer a nominal *report* discourse orientation whereas females prefer a clausal/deictical *rapport* discourse orientation. These two orientations could be due to socialisation effects or average biological differences in brain organisation or a combination of both effects.				
deep formality 2	p = 0.023 *d* = 0.478 M			p = 0.049 *d* = 0.235 M
first person pronouns	p = 0.025 (1-t) *d* = -0.411 F			
deep formality	p = 0.026 *d* = 0.467 M			p = 0.049 *d* = 0.049 M
total personal pronouns	p = 0.030 (1-t) *d* = -0.39 F			
common nouns	p = 0.034 *d* = 0.441 M			p = 0.009 *d* = 0.314 M
nouns followed by *of*	p = 0.037 *d* = 0.37 M			

total nouns	p = 0.049 *d* = 0.411 M	p = 0.027 *d* = 0.402 M		p = 0.008 *d* = 0.318 M
genitives		p = 0.011 F		
pronoun *it*		p = 0.030 F		
predicative adjectives		p = 0.035 F	p = 0.025 M	
third person pronouns		p = 0.030 F		
indefinite pronouns			p = 0.006 F	
average word length in syllables			p = 0.015 *d* = -0.50 F	
average word length			p = 0.022 *d* = -0.466 F	
words longer than six letters			p = 0.023 *d* = -0.477 F	
prepositions				p = 0.045 (1-t) *d* = 0.201 M
downtoners				p = 0.010 F
Pattern 2: Distribution of expletives Males and females on average tend to produce language in different ways in order to show their affiliation and/or their detachment with certain values or social groups. For gender, a point of distinction lies in the use of swear words.				
swear words			p = 0.001 M	p = 0.004 M
Other variables				
social words	p = 0.032 (1-t) *d* = -0.378 F		p = 0.029 *d* = -0.457 F	p = 0.012 *d* = -0.298 F
negative emotion words		p = 0.020 F		

positive emotion words			p = 0.0003 *d* = -0.761 F	p = 0.012 F
present participial WHIZ deletion relatives	p = 0.037 M			
Dimension 3	p = 0.043 *d* = 0.419 M			
time adverbials		p = 0.049 M		
past participles	p = 0.032 *d* = -0.45 F			
suasive verbs			p = 0.010 F	
lemma SAY			p = 0.028 M	
past tenses			p = 0.049 F	

The table above largely confirms that the linguistic patterns observed previously in other studies also apply to the FMT corpus. Even though all the variables gathered from all the literature surveys were tested, the ones that presented an effect for gender were almost only the ones for which the literature would have predicted a gender effect in other registers.

In general, it can be noticed how the features that distinguish gender in Task 1 are different from the features that distinguish gender in Task 2 and 3 and that therefore even though most of the findings of the studies reviewed in the literature are largely confirmed, they are not valid in all registers. In Task 1, deep formality is the variable that has the greatest gender effect. On the other hand, in Task 2 and 3 emotional language and swear words are more useful to distinguish the genders. As predicted by many studies reviewed in the literature survey on gender, the effects obtained were small when the Tasks are all combined together, as it is the case in the last column of Table 5-1. However, it is possible to observe from this study that when register variation is controlled the effect sizes increase considerably. Under such conditions, higher effect sizes can be observed, thus suggesting that the smaller effect sizes registered so far in previous studies can be a consequence of lack of control of register variation. The highest effect size noticed appears in Task 2 for positive emotion words, where the Cohen's *d* of 0.761 found indicates that the two genders are separated by almost one standard deviation. In the other Tasks, Cohen's *d* scores approaching 0.5 indicated that the difference can be approximated to half of a standard deviation. Gender differences of this size have been rarely found in previous studies.

The findings of this empirical study confirm that nominal forms are more typical on average of male writing whereas pronominal and verbal forms are more typical on average of female writings. Although this finding is largely valid for the whole corpus, it is significantly stronger for Task 1. The highest effect size for this pattern is the one of Deep Formality 2, which summarises the opposition between deictic and formal linguistic features. For this variable, only in Task 1, the two genders are separated by almost half of a standard deviation. Although other nominal or pronominal and verbal variables do not show such strong effects, the predicted direction is found. In the *Other variables* category, the higher values of Dimension 3 for male subjects confirm that on average in Task 1 males used a more refined nominal elaboration than females. Since high scores on Dimension 3 reveal a discourse orientation that focuses on refined and context-independent reference, this group of variable could also be in principle grouped with the more general rapport/report pattern. The only exception to the general pattern that shows females and *rapport* features on one hand and males and *report* features on the other hand is a higher average word length for females in Task 3. An example of the opposition between *rapport/report* discourse found in Task 1 can be observed in Table 5-2 below, where the highest and the lowest scoring texts for Deep Formality (DF) in Task 1 are shown.

Table 5-2 - The highest and the lowest scoring texts for deep formality in Task 1. The features contributing to a high score in deep formality are underlined whereas the features contributing to a low score on deep formality are in bold

MAPO – 71, above undergraduate, male, SCI = 4.8, DF = 67.62
In December 2011 **I travelled** to Malta for one week on one of **your** "Super Deluxe" holidays, at a cost of £1500. The brochure description of the holiday **specified** that **it** would **be** all-inclusive, in a 5 star hotel, and that there would **be** no extra costs **whatsoever**. The Malta Palace Hotel **was well below** 5 star standard. None of the staff in the hotel **spoke** English (unusual in Malta), and all **were extremely** surly and unhelpful. The room, on the 14th floor, **was** dirty, smelly, and noisy. None of the lifts **worked**. The restaurant **was far too** small for the number of occupants in the hotel, and on five days out of seven **it was closed** by the Maltese authorities because of rat infestation, as a result of which all meals for those days **had** to **be taken** in a restaurant. The nearest restaurant **was** 6 miles **away**, **requiring** taxi travel at a cost of €30 for each meal, a total of €450. Expenditure on food for five days **totalled** €750. **Your** local representative, Mr S. Berlusconi, **evaded** all attempts at contact, and **was utterly** useless. **I require** full reimbursement, within seven days, of the £1500 **paid** for the holiday, plus €1000 for the expenses **detailed above**, plus £1000 compensation for inconvenience and distress. In addition, **I require** compensation for physical and mental trauma, full details of which will **be provided** after **my** discharge from hospital, and **when** medical assessment of **my** condition **has taken** place.
ANMA2 – 41, below undergraduate, male, SCI = 2.3, DF = 44.44
I am writing to **you today** to **ask you why my** holiday on **your** travel package **was** not satisfactory in the aspects of food and **how it was served** because **it was sometimes** cold and not edible and the whole of the meals **sometimes was** not at all enjoyable but some of the time **it was** and **then I** will **speak** about the **sleeping** arrangements because **they were** not at all what **I expected** from the package **I paid** for and **then** there **was** the entertainment which **was** not **entertaining** at all, **I wish** to **ask you why you** would **advertise** a holiday that **seemed** to **be really** nice but **when I went** to the place **even** the travel **was** not acceptable because of the way **we were left** not **knowing where** or **when we were meant** to **get** off coaches and **where** to **go** because the staff **you had** to **show us were** not knowledgeable about what **we were doing** and **where we were going** and **I am severely disappointed** in this holiday which **I have saved my** hard **earned** money to **have** and **I am going** to **be asking** for a refund because of the way not **just I was treated** but the way **we were** all **treated** because **it became** a **very** horrible experience not a holiday that **I was expecting** and due to this **I wish** to **make** a formal complaint due to the stress and anguish **I experienced**.

The examples above show how in Task 1 male authors on average tended to be more detached from the shared context than female authors and that, on the other hand, female authors on average tended to be more focused on the personalisation of the discourse and/or on moving the focus of the text towards the individuals taking part in the interaction. The difference between the genders found in Task 1 is therefore compatible with the difference observed by Heylighen and Dewaele (1999). From the texts in Table 5-2 it is clear how DF is highly influenced by the complexity and quantity of noun phrases that do not consist of pronouns. Halliday (2004) identified the use of complex noun phrases and their ability to create detailed taxonomies as one of the most typical strategy used in the language of science. However, the fact that pronouns are included in DF complicates the interpretation of what DF is measuring. DF could in fact distinguish both person-centred discourse against object-centred discourse (when the incidence of nouns against pronouns is dominant) and complex load of information

against interpersonal concerns (when the incidence of complex noun phrases against verb phrases is dominant). Indeed, these two aspects are intrinsically linked, since a higher usage of pronouns has to be present in texts produced by subjects that struggle with producing highly informational discourse. For example, in ANMA2's text, the low DF score is probably caused by their low level of education and possibly familiarity with the use of written language and therefore their reduced capacity of producing informationally loaded discourse. In ANMA2's text a lack of nominal elaboration is noticeable together with a high frequency of pronouns. On the other hand, MAPO's text shows both the tendency to elaborate nominally and the tendency to simply refer to objects rather than to people.

As mentioned above, the gender effect for the *rapport/report* pattern is strongly present only for Task 1. The fact that there is no effect for Task 3, the most Involved of the three Tasks (cf. Section 4.1) could indicate that the more personal a text becomes the less likely it is to show a gender pattern of the *rapport/report* type. In other words, in a register in which individuals are already pressed to be Involved and person-centred then there is no room for variation between *rapport* and *report* discourse, thus blocking the gender pattern from emerging. However, in Task 3 other patterns of linguistic variation were found to be gendered.

In the FMT corpus, and especially in Task 3, on average males produced more swear words than females. An example of use of expletives in Task 3 is shown in Table 5-3 below, where the two highest scoring texts for swear words in Task 3 are reproduced.

Table 5-3 – The two highest scoring texts for swear words in Task 3

TYJO – 30, undergraduate, male, SCI = 4.8, swear = 3.51
We're all extremely angry and disappointed with you. We feel that you do not value the work we do for you and the company and we're sick to our teeth with you and your abuse. We have worked in the firm for a very long time and you should start to think of changing your ways otherwise things might start to happen to your property. Your behaviour has **pissed** many of us off and we're going to **fuck** your car up if you don't stop. The way you make me feel is like **shit** and I will not take your bullshit anymore. Someone will **fuck** your car up if you don't stop. The damage to your car will start with a warning like an early morning spray paint on your bodywork. Someone will put a nail in your tyre. It could happen in the lunch break and at night, you just don't **fucking** know. Someone will take a jack and smash your lights and then if you carry on it will get worse. How can you be so **fucking** horrible? Were you bullied as a kid and you're now in a bit of power you think that you can put your **shit** in everyone else? Well, were not having it and someone will put your mirrors and bodywork in with a **fucking** baseball bat if you don't stop.
MIBO – 40, below undergraduate, male, SCI = 4.5, swear = 3.04
Alright son listen up, I'm only doing this to give you fair warning. Your **fucking** attitude towards the people who work in the office is appalling, you are a bully. Well I have been dealing with bullies my whole life and have found the easiest and most effective way to sort the problem is to meet it head on. Who do you think you are you jumped up little **prick**! You strut around barking out orders demanding people make you drinks fetch your lunch, were you raised like that, did Daddy spank you and belittle you is that where you've got this enormous chip on your shoulder from. Understand me you **fucking cock** your attitude had better improve dramatically or it could become a very expensive and unpleasant place to work. Allow me to enlighten you in case your pea sized intellect cannot compute what's going on. That **fucking cock** extension that you drive and fawn over, that really is a nice car how much was that 30, 40 k, now imagine paint stripper all over the panels you'll wake up one morning and its gone from red to gun metal grey. So it gets repaired and you don't alter next time me and my mates take a **fucking** baseball bat to it, I will smash every window, the light clusters and batter every panel, and then just for good measure I will re decorate the interior, all that lovely beige leather sprayed neon green. Get that **fucker** repaired you **cunt**! So just so we understand each other and in case you've missed the gist of this little note wind your **fucking** neck in, you'll find that people respond a lot better from not being balled at, people are far more productive if they feel that you value what they do. There is absolutely no need to behave in the ways you are behaving. This is not an idle threat unless there is real change get ready Babbie!

The negativity and aggressiveness surrounding the two sample texts is evident even if the swear words are ignored. It is possible to contrast these two samples above with two highest scoring texts for positive emotion words displayed in Table 5-4 below.

Table 5-4 – The two highest scoring texts for positive emotion words in Task 3

CHHA – 40, below undergraduate, female, SCI = 2.5, posemo = 6.60
Parasite This letter to you is to let you know that your behaviour towards your staff at the office will no longer be tolerated. You constantly belittle your employees and bring them to a desperate wreck. You are a bully who only picks on those weaker than you and we have now decided that enough is enough. Many people have worked so hard at this company even before you joined us. You reap the **rewards** of all our hard work and what **thanks** do we get. NONE! Instead we are victimised, bullied, made to feel we are not **valued**. How about **giving** some **thanks** to those who keep you in your position instead of constantly kicking people when they are down. You are nothing but a coward and you will get what's coming to you very soon. First we will start with something which is **precious** to you. Let say, oh your **lovely** BMW. How would you **like** us to damage some of that **beautifully** black bodywork that you **value** so **dearly** (unlike your hard working staff). I think maybe if we treated your car bodywork **like** how you **treat** your employees - you might get the gist of how we feel at the hands of your temper tantrums. Also, it may also hurt you **dearly**, if we gave your vehicle a new spray job one evening. How about Yellow - for the yellow bellied coward that you are. Would you **like** that, Mr Cannon?? The above is just an idea of the **rewards** you may receive from us disgruntled employees if you do not change the way in which you **treat** us. This is just for starters and it will be in your best **interest** to modify your behaviour immediately and give us the **respect** and **gratitude** we deserve. You have been WARNED!
EMGI – 21, below undergraduate, female, SCI = 4, posemo = 6.44
I have worked at this company for 15 years. I have put an incredible amount of effort into making **sure** our customers get the very **best** and in turn the company earns as much as possible. I have always felt **valued** as an employee here, until 6 months ago when you were elected as our store Manager. The working environment took a sudden plummet with your leadership. You don't **value** your staff; we consistently work hard, yet you **treat** us unfairly. You're rude and obnoxious. You **treat** all the female staff in a vulgar manner that you dismiss when we complain. A particular male member of staff you call very offensive names, due to his weight. You scream at us and call us names for hitting below target, is it any **surprise** we are not doing **well** with your constant abuse. You sit in your office **playing** games and chatting on the phone instead of doing any real work. We constantly hit target with Mr O'Brian in charge, why? Because he worked with us, he lead by example. People work much **better** from **praise** than torment. You are a bully Mr Jones and your behaviour **won**'t be tolerated. We have all had enough, and unless prompt changes are made for the **better** we will take action. **Certain** members of staff have suggested violence, which normally I would be completely against, yet with how you've treated us I could condone such actions. It is known to us how much you **value** your **precious** car, it would be a shame if the disgusting words you shout at us were to be spray painted all over its bodywork late one night. This used to be a place we were **happy** to come to every day, now it is somewhere we dread. **Treat** us with the **respect** and **care** we deserve, or the whole town will see what insults we put up with displayed all over your **pride** and **joy**.

Firstly, the inspection of these samples shows that many classification mistakes were made by Pennebaker's wordlist. For example, the verb *to treat* was mistaken for the positive emotion noun *treat*. Similarly, the construction *would like to* was mistaken for the positive emotion verb *to like*. Even though there were mistakes, it is noticeable how Pennebaker's variable could nonetheless approximate a more positive and polite discourse orientation that is in contrast with the orientation observed in the examples containing swear words and a negative and aggressive discourse orientation above. In the context of Task 3, females tended to focus on the positive aspects that were missing from the job situation and/or on the positive aspects that were requested (e.g. *respect and gratitude we deserve, respect and care we deserve*). Even when describing the car, the two subjects with the top scores for positive emotion used positive attributed, such as *precious* or *lovely*. This focus on the positivity even if in a negatively loaded context is in contrast with the focus on negativity given by the swear words that is explored in the section above. The positive emotion words in Task 3 were often used to stress the mistreatment by focusing on what the boss should have done rather than what the boss was doing. Male authors, on the other hand, were more likely to aggressively attack the boss in relation to what the boss was doing, often using swear words. By avoiding direct confrontation and by dealing with the aggressive context using positive emotion words, it is possible to argue that female authors tended to hedge pragmatically by being less direct than male authors. Even though the powerless register variables were not tested, as these variables were excluded from the study since their calculation would have been too time consuming for the scope of this work (cf. Appendix 9.5) these findings still present some support for the powerless register hypothesis.

The lack of swear words and the presence of positive emotion words used to mitigate the face-threatening situation could be indeed regarded as an act of hedging at the discourse level rather than at the grammatical level. Indeed, it could be argued that the two patterns noted of powerless register features and the distribution of expletives are perhaps part of a bigger underlying pattern of face-threatening management or politeness. This politeness pattern could be also responsible for the higher average word length in the female sample. In the professional-like register of Task 1 there was no significant difference for average word length between males and females. The difference appears only in Task 3 and this could be due to the fact that on average females were trying to be more formal than males, although this hypothesis can only be tested more carefully in the future using more controlled experimental conditions.

In conclusion, the present study suggests that the three general patterns of linguistic variation for gender that were extrapolated from a literature survey in Section 2.1 are also found in the FMT corpus. The review of the findings of the empirical study of the FMT corpus however showed that the powerless register pattern and the distribution of swear words patterns can be conflated in one general pattern of politeness. In sum, therefore, two general patterns of linguistic variation for gender were observed in the FMT corpus: a **rapport/report discourse orientation** pattern that distinguishes, respectively, female gender from male gender on average mainly in Task 1; and a **politeness discourse**

orientation pattern that is more often found in texts produced by females in personal threatening letters such as Task 3.

In terms of the explanation for the patterns observed, the analysis supports the hypothesis that some differences exist on average between the genders, either given by biological, psychological or social effects. Both findings for Task 1 and 3 are compatible with Lakoff's (1973) theory that female gender present a higher degree of powerlessness and lack of commitment. This socialisation or sub-culture hypothesis suggested by Lakoff (1973) and compatible with McEnery's (2006) explanation using Bourdieu's theory of distinction is the most likely hypothesis at the present to explain gender variation in language use, since even neuroscientists and psychologists agree on the significant contribution of socialisation in the creation of a person's gender (Kaiser *et al.*, 2009). Although exposure of the genders to different linguistic varieties can affect their linguistic repertoire, given that the sample in the FMT corpus did not present any skew for gender and social class or gender and level of education, it is unlikely that access to standard language and/or social movement is responsible on the gendered pattern observed. This study thus supports the hypothesis that the most likely explanations for these differences are therefore of a socio-cognitive nature, as suggested by Lakoff (1973).

Although it should be clarified that the explanations provided in the paragraph above are only working hypotheses, the empirical evidence found in the present study as well as in other literature items reviewed in the present work suggest that genders do differ on average in the way they employ language, even though this difference is smaller than for other social factors taken into account in the present study (cf. Section 5.2, 5.3, and 5.4 below). Future work should focus more intensively in understanding the real cause of this variation, rather than speculating on stereotypes. In accordance with Chambers (1992), although it has not been possible in the present study for reasons of space and time, in the future sociological gender as well as biological gender should be considered as the independent variables. Other factors that could contribute to the language variation observed should be measured in future experiments, such as hormone levels, tendency for depression and/or personality. When these components are isolated, if the two-culture hypothesis is true it is predicted that the sociological gender effect would be greater than any other effect. The verification of such fact would provide strong support for the two-culture hypothesis and confirm the nature of gender variation in language use. The present study nonetheless provides evidence of the fact that some kind of gender variation exists and that this variation is found in the interaction between gender and register.

5.2 Age

In this section, Table 5-5 below summarises all the variables that presented an age effect by showing the variable significance levels as well as the magnitude of their correlation. The presentation of the variables is organised according to the four major patterns introduced in the literature review of age in Section 2.2. All the variables that presented a statistically significant difference were organised in these categories, even in the cases of variables that were not gathered from the literature review of age. The variables that did not fall in any of the patterns for age are categorised in the general category 'Other variables'. Because age was treated as a continuous numeric variable, the test used was a Pearson's *r* correlation test for normally distributed linguistic variables and a Spearman's *rho* correlation test for non-normally distributed linguistic variables.

Table 5-5 - Linguistic variables that presented a significant effect for age, showing: p-value ('1-t' indicates a one-tailed value) and the correlation coefficient

Variable	Task 1	Task 2	Task 3	Whole corpus
Pattern 1: Syntactic complexity Syntactic complexity at the clausal level decreases with age and this often corresponds to an increase in lexical complexity. This effect could be due to either a decrease in working memory capacity with older age or to a shift in the way information is packaged due to increased experience with language.				
dependent clauses per sentence	p = 0.00001 *r* = -0.391	p = 0.011 *r* = -0.259		p = 0.00001 *r* = -0.229
Fichtner's C	p = 0.0002 *r* = -0.367	p = 0.039 (1-t) *r* = -0.181		p = 0.002 *r* = -0.185
Baayen's P	p = 0.001 *r* = 0.338	p < 0.00001 *r* = 0.462		p = 0.0002 *r* = 0.221
clauses per t-units	p = 0.0004 *r* = -0.351	p = 0.005 *r* = -0.282		p = 0.00002 *r* = -0.251
average sentence length	p = 0.002 *r* = -0.319			
short t-units	p = 0.007 *r* = 0.272			p = 0.002 *r* = 0.179
Flesch-Kincaid score	p = 0.006 *r* = -0.278			p = 0.018 *r* = -0.140
Dimension 5	p = 0.014 *r* = -0.251			

passives	p = 0.017 *r* = -0.244			
by-passives	p = 0.023 *r* = -0.231			p = 0.047 *r* = -0.117
average t-unit length	p = 0.023 *r* = -0.233			p = 0.005 *r* = -0.165
sentence relatives	p = 0.036 *r* = -0.214		p = 0.037 *r* = 0.215	
independent clause coordinations	p = 0.037 *r* = -0.213			
present participial clauses		p = 0.018 *r* = -0.241		
conditionals		p = 0.024 *r* = -0.230		
type-token ratio		p = 0.039 *r* = 0.211		
long t-units				p = 0.031 *r* = -0.127
Pattern 2: Dimension 1 Older age is correlated to a higher use of Informational features and with less frequent Involved features (using Biber's (1988) terminology).				
be as main verb	p = 0.0002 *r* = 0.366			p = 0.009 *r* = 0.153
deep formality	p = 0.008 *r* = 0.269			p = 0.009 *r* = 0.154
predicative adjectives	p = 0.008 *r* = 0.267			
singular proper nouns	p = 0.008 *r* = 0.271			
total proper nouns	p = 0.010 *r* = 0.263			
contractions	p = 0.013 *r* = -0.254			
deep formality 2	p = 0.011 *r* = 0.259			p = 0.012 *r* = 0.148
prepositions	p = 0.025 *r* = 0.228			p = 0.050 *r* = 0.116

determiners	p = 0.029 *r* = 0.223			p = 0.033 *r* = 0.109
total nouns	p = 0.030 *r* = 0.222			p = 0.024 *r* = 0.134
first person pronouns	p = 0.032 *r* = -0.219	p = 0.045 *r* = -0.174	p = 0.030 (1-t) *r* = -0.194	p = 0.011 *r* = -0.149
average clause length	p = 0.041 *r* = 0.209			p = 0.047 *r* = 0.117
synthetic negations	p = 0.046 *r* = 0.205			
private verbs		p = 0.009 *r* = -0.266		p = 0.024 *r* = -0.133
cardinal numbers		p = 0.027 *r* = 0.225		
demonstrative pronouns		p = 0.030 *r* = -0.222		
demonstratives		p = 0.034 *r* = -0.216		p = 0.011 *r* = -0.149
quantifier pronouns		p = 0.046 *r* = -0.204		
Dimension 1		p = 0.043 (1-t) *r* = -0.176		
total personal pronouns			p = 0.022 *r* = -0.235	p = 0.016 *r* = -0.142
plural proper nouns			p = 0.029 *r* = -0.225	
indefinite pronouns			p = 0.033 *r* = -0.219	
WH relative clauses on subject position			p = 0.010 *r* = 0.262	
analytic negations	p = 0.025 *r* = -0.228	p = 0.033 *r* = -0.218		p = 0.003 *r* = -0.176
common nouns				p = 0.026 *r* = 0.131
that deletion				p = 0.018 *r* = -0.140

Pattern 3: Realisation of stance Generally speaking, younger people tend to use stronger linguistic stance than older people. This pattern could be either due to language change or to change in personality, life-goals and attitude with ageing (Bourdieu's theory of distinction).				
Dimension 4	p = 0.007 *r* = -0.271			
verb bases	p = 0.007 *r* = -0.273			
total modals	p = 0.014 *r* = -0.249			
necessity modals	p = 0.034 *r* = -0.217			
interjections		p = 0.018 *r* = -0.242		
innovative stance adverbs		p = 0.048 (1-t) *r* = -0.171		
suasive verbs				p = 0.026 *r* = -0.131
general adverbs			p = 0.042 *r* = -0.209	
Pattern 4: World-view change As people get older their view change towards more positive feelings and towards looking to the future. This effect could be due to an average decrease in neuroticism and depression that naturally happens with age.				
past participles	p = 0.003 *r* = -0.299		p = 0.043 *r* = 0.208	
total emotion words		p = 0.007 *r* = -0.276	p = 0.016 *r* = -0.247	p = 0.006 *r* = -0.163
negative emotion words		p = 0.034 *r* = -0.217		
positive emotion words		p = 0.045 *r* = -0.205	p = 0.002 *r* = -0.318	p = 0.002 *r* = -0.185
past tenses		p = 0.046 *r* = 0.204		
time words			p = 0.027 (1-t) *r* = -0.197	
social words				p = 0.048 *r* = -0.117
Other variables				

tokens	p = 0.012 *r* = -0.257	p = 0.003 *r* = -0.301		p = 0.007 *r* = -0.158
Dimension 6		p = 0.045 *r* = -0.205		p = 0.018 *r* = -0.139

The table above largely confirms that the linguistic patterns observed previously in other studies regarding the linguistic variation associated with age also apply to the FMT corpus. Even though all the variables gathered from all the literature surveys were tested, the ones that presented an effect for age were almost only the ones for which the literature would have predicted an age effect.

Similarly to what was observed with gender, in the analysis of age it is also noticeable that not all the patterns show the same effects for all the Tasks, thus indicating that register variation has a significant confounding effect. Overall, for age the strongest effects noticed were related to syntactic complexity for Task 1 and 2. However, in Task 3 the emotional language variables of the world-view change pattern show greater effects than in other Tasks.

Similarly to what observed for gender, the effects obtained were small when the Tasks were all combined together but they increased considerably when register variation was controlled. The largest correlation observed was 0.462 in Task 2 for Baayen's P, a measure of intrinsic vocabulary rarity. A preliminary conclusion for age is that it seems confirmed that less syntactic complexity at the level of sentence is more often employed by older adults. Likewise, Informational features pattern in a way that corresponds to findings established in many studies reviewed in Chapter 2, with older adults being more Informational and younger adults being more Involved.

The pattern noted in the literature review of the syntactic complexity decrease related to age is confirmed in the FMT corpus and, more specifically, in Task 1 of the FMT corpus. All the variables that represent a highly elaborated clausal syntax show a decrease with age, with the only exception being the frequency of sentence relatives in Task 3. A sample of Task 1 texts for both the high and the low syntactic elaboration patterns is shown in Table 5-6 below, in which the highest and the lowest scoring text for Task 1 for dependent clauses per sentence are displayed.

Table 5-6 - The highest and the lowest scoring texts for dependent clauses per sentence for Task 1. The texts are here divided in sentences and the dependent clauses in each sentence are underlined and in bold. Embedded clauses are marked by angle brackets (<>).

MAPO – 71, above undergraduate, male, SCI= 4.8, DC_S = 0.58
(1) In December 2011 I travelled to Malta for one week on one of your "Super Deluxe" holidays, at a cost of £1500. (2) The brochure description of the holiday specified **that it would be all-inclusive, in a 5 star hotel**, **and that there would be no extra costs whatsoever**. (3) The Malta Palace Hotel was well below 5 star standard. (4) None of the staff in the hotel spoke English (unusual in Malta), and all were extremely surly and unhelpful.

(5) The room, on the 14th floor, was dirty, smelly, and noisy.
(6) None of the lifts worked.
(7) The restaurant was far too small for the number of occupants in the hotel, and on five days out of seven it was closed by the Maltese authorities because of rat infestation, **as a result of which all meals for those days had to be taken in a restaurant**.
(8) The nearest restaurant was 6 miles away, **requiring taxi travel at a cost of €30 for each meal, a total of €450**.
(9) Expenditure on food for five days totalled €750.
(10) Your local representative, Mr S. Berlusconi, evaded all attempts at contact, and was utterly useless.
(11) I require full reimbursement, within seven days, of **the £1500 paid for the holiday**, plus €1000 for the expenses detailed above, plus £1000 compensation for inconvenience and distress.
(12) In addition, I require compensation for physical and mental trauma, **full details of which will be provided after my discharge from hospital**, **and when medical assessment of my condition has taken place**.

ALWH – 20, below undergraduate, female, SCI = 4, DC_S = 3.64

(1) Last year I purchased my summer holiday through FirstHoliday, and the service my family and I received, from beginning to end, was not satisfactory.
(2) **Being a <working Mother>** the two weeks summer break I get **to spend** with my husband and two children is **what makes every long hour at work worthwhile**, but the package **we received** from your company was just awful, and on my return to the busy city of London I felt like **I needed another 2 weeks rest <to recover>**.
(3) On our arrival to the airport, **which was delayed by an hour <because the coach to the airport <you'd provided> was unapologetically late>**, I was appalled **to discover** **that, <having overbooked the business economy seats on the plane>, my family and I had the choice of <waiting 22 hours for the next flight> or <to take seats in economy class>**.
(4) **Although I can't thank you enough for those excellent alternative options**, **once we'd decided <to not waste anymore of our holiday at the shambolic Heathrow airport>**, our seats in economy class were nowhere near each other.
(5) My 8 year old son was seated next to a member of the young Conservative party.
(6) I don't even wish **to think about the harm <that he may have caused>**.
(7) **Once we'd finally got to our hotel**, we discovered **our rooms hadn't been cleaned since <the last occupants had left>**, **the bath wasn't big enough <to drown a mouse>**, **and the radio didn't work**.
(8) We also found **that the all inclusive hotel <you had sold to us, <and that we had paid for>>, only provided breakfast**.
(9) My husband and I work hard all year **to give our children the sort of holidays <we never had growing up>**, and **to have this as a product of that** makes me exceptionally upset.
(10) A holiday is supposed to be relaxing and enjoyable, and this trip left me feeling stressed and angry.
(11) **Having missold our holiday, and provided nothing but disaster**, we wish **to ask for a partial refund of £500**, or we will be taking legal action.

The two samples show a clear contrast. On one hand, the text with a high frequency of dependent clauses per sentence presented a wealth of information, often incidental to the main topic, and did so using syntactic means of expansions, such as infinitive clauses, projected clauses and adverbial clauses. On the other hand, the text with a low frequency of dependent clauses per sentence were more concise and direct and tended to use a nominal rather than syntactical elaboration, using prepositional phrases, attributive adjectives and nominalisations.

Decrease in complexity of the sentence can be observed also from the negative correlation between age and Dimension 5, the degree of Abstract Discourse. This variable was not predicted by the literature review to vary with age but the significant result obtained is compatible with the general decrease in syntactic complexity, since a high score on Dimension 5 corresponds to a high use of passives clauses and conjuncts, which corresponds therefore to a higher syntactic complexity.

Apart from the decrease in syntactic complexity, previous literature noted that the decrease in syntactic complexity is accompanied with an increase in lexical complexity. This pattern is indeed found in the FMT corpus although only for variables that measure intrinsic vocabulary rarity. The literature would have predicted that age is also positively correlated with the number of long words (or average word length) and number of rare words, that is, with variables that measure how many rare words of the English language are used in the text. Instead, the only variables that showed a positive significant increase with age were type-token ratio and Baayen's P. This latter variable was introduced in the study by Mollet *et al.* (2010) in the literature survey on education and it is a proxy to intrinsic vocabulary rarity, that is to say, to how rare words are within the text, rather than within the language as a whole. Since no other variables that measure lexical rarity, such as Advanced Guiraud 1000 or measure of word length were significantly correlated with age, it is possible to advance the hypothesis that the lexical complexity that some studies refer to is indeed intrinsic vocabulary rarity.

A problem of Baayen's P, however, is that it is highly influenced by text length. Baayen's P is, in fact, simply the relative frequency of *hapaxes*. As shown in Table 5-5, the total number of tokens was negatively correlated with age. Even though in the FMT corpus Baayen's P and text length presented only a medium size correlation (r = -0.602), there is indeed a connection between these two variables and their age effect. For example, there is the possibility that older people produced more lexically sophisticated texts just because they produced shorter texts with more unique types. Similarly, it could also be the case that older subjects with more experience of lexis produced texts that were more varied lexically and that conveyed all the necessary information in fewer words. It is impossible to understand the real direction of these effects within the scope of the present study. In order to untangle this relationship, further studies should be carried out in the future using larger samples and more sophisticated statistics that can help to identify a chain of causality. For the present study, however, it is still possible to conclude that there is a negative correlation between clausal syntactic complexity and age and that this decrease corresponds to an increase in the degree of conciseness and intrinsic vocabulary rarity, given the results for text length, Baayen's P and type-token ratio.

The second pattern identified in the literature review, the linguistic pattern concerning the increase of Informational discourse with older age, is also confirmed in the FMT corpus. Similarly to the analogous pattern examined for gender, the *rapport/report discourse orientation* pattern, a difference in terms of Involved and Informational discourse was found across subjects of different ages. With the exception of only one variable, *be as a main verb*, all the variables that are part of the Informational pole of Biber's (1988) Dimension 1 increase in frequency with age whereas all the

variables that are part of the Involved pole of Biber's (1988) Dimension 1 decrease with age. This result means that the pattern previously shown in the sample texts in Table 5-2 not only characterizes different genders but also different ages. The same linguistic features that on average characterize male texts also characterize older subject's texts. On the other hand, the same features that on average characterize female texts also characterize younger subject's texts. The claim that there is a connection between the Dimension 1 poles, gender and age was made by Schler *et al*. (2006) and is therefore confirmed in the FMT corpus. These findings also indicate that it is not possible to use Dimension 1 features (that appear both in the *rapport/report* pattern and in the nominal complexity pattern) to distinguish either gender or age independently from each other, since the same variables show an effect for both the social factors. Future studies should aim at untangling these relationships perhaps using experimental conditions and/or more advanced statistics.

The only exception found among the Dimension 1 variables is the frequency of *be as main verb*. This variable showed a very large effect but in the opposite direction to what was predicted, that is, there was an increase of *be as main verb* with age. This anomaly could be explained if *be as main verb* were part of the pattern examined above, the decrease of syntactic complexity. Even though the negative relationship between the frequency of copular *be* and sentence complexity is not well established, there is reason to believe that there might be a relationship of this kind in the English language. For example, Biber *et al*. (1999: 360) found that registers that use many copula *be* clauses, that is, academic prose and newspaper articles, do so because this grammatical pattern helps them to focus on the relationships between entities rather than on the action and events or mental states that are expressed by other types of clause patterns. That being the case, registers that use more copula clauses tend to shift the complexity weight to the noun phrase. Furthermore, Biber *et al*. (1999: 446) also found that 50% of copula *be* clauses in the English language consist of noun phrases whereas complement clauses are relatively rare after *be* as main verb. This fact is indeed another piece of evidence that might suggest that *be as a main verb* is more common in any text which is concise and which uses more nominal complexity and less sentence complexity, such as the texts typically produced by the older subjects for the FMT corpus.

The third pattern noted from the literature regarded the different use of stance between people from various life stages. The expression of stance characterized by the variables identified by Barbieri (2008) was limited to only the significant effect of innovative stance adverbs that decrease with age as predicted. Frequency of swear words or frequency of traditional stance adverbs did not show any significant relationship with age. However, when Dimension 4 is accounted, then the overall pattern of a general decrease in the expression of stance is confirmed in the FMT corpus. Dimension 4 is the Dimension of Overt Expression of Persuasion which Biber (1988) found to be characteristic of genres that overtly express modality and other stance items in order to persuade the hearers or readers. The significant effect that this and other related variables show for age thus supports the hypothesis that there is a significant decrease in overt expression of stance with age.

Finally, the last pattern of linguistic variation identified in the literature is the change in emotion words from negative emotions of younger age to more positive emotions in older age. This pattern was only partially observed in the FMT corpus. A decrease in negative emotion words was indeed present in Task 2 and Task 3 of the FMT corpus. However, positive emotion words also decreased and therefore, in conclusion, it seems likely that the general pattern consists in a general decrease of expressions of emotionality. The conjunction of this effect to the decrease of Informational discourse could point to a general decrease of Involvedness that corresponds to ageing. Time concerns were also present in a direction consistent with Pennebaker and Stone's (2003) predictions, since less time words were produced by older participants. However, the prediction regarding future and past tenses was not replicated in the FMT corpus since no effect was noted for future tenses and the effects of past tenses observed are not consistent across Tasks.

In conclusion, therefore, the pattern of world-view change is indeed present to a certain extent and it mainly concerns general emotionality rather than just negativity. However, a limitation of this finding is that the sample of subjects considered was not completely free from bias. As described in Section 3.2.1, gender was skewed for age, with more female younger subjects. Therefore, it is not possible to determine at this stage which social factor is influencing the emotional language variables more significantly than the other. Further tests can be conducted in the future to establish the contribution to the explanation of variance of the emotional language variables by age and gender. At this stage, however, it is still possible to conclude that there is a relationship between age and emotionality in the FMT corpus.

In conclusion, three of the four patterns retrieved from the literature review in Section 2.2 were also found in the FMT corpus, with the exception of the pattern related to the change in world view that was only partially replicated. The empirical work on the FMT corpus showed evidence of the presence of another pattern of linguistic variation that characterizes age: the increase of conciseness. This pattern is marked by the decrease of text length with age and the increase in the amount of intrinsic vocabulary rarity. In summary, the main patterns of variation related to age in the FMT corpus were: a **decrease of syntactic complexity** with age mostly observed in Task 1, a **decrease of Involved discourse** with age characterised by both an Involved discourse and a decrease in emotional language and observed in different shapes in all the three Tasks, an increase of **conciseness** with age that is however not found in Task 3, and the presence of different **patterns of stance realisation** that distinguish younger from older writers in Task 1.

The findings of the present study provide some evidence that age differences in the use of language in the FMT corpus exist and that these differences are in line with the findings of previous studies. There is enough evidence to confirm that there is a general tendency for the variables identified in previous literature to pattern in the predicted way. There is however limited theoretical work that can explain and connect these patterns to provide a coherent theory of linguistic variation with age. It could be the case that the majority of these patterns are linked to each other to form more significant latent

patterns. According to Halliday (2004), complex clausal syntax is a form of Grammatical Intricacy and it is characteristic of spoken genres where there is less time for planning and therefore not much opportunity to produce information in the most typical complexity of typical written genres, that is, by using more lexically dense language. This relationship between Grammatical Intricacy on one hand and Lexical Density on the other hand could explain how the decrease of sentence complexity is linked to an increase in Informational variables. However, even if this link is empirically established, the reason why this pattern appears to be connected with age still remains a mystery.

Kemper *et al.* (2001a,b) proposed that the relationship between Grammatical Intricacy and age is given by the loss of working memory of older subjects, since working memory is necessary to produce complex syntax. Although this hypothesis would seem valid, since older subjects were also the subjects that produced the shortest texts, the working memory explanation is not supported by the present work for at least three reasons. Firstly, P-Density, which should also correlate with working memory, was not found to have any relationship with age in any of the Tasks. Secondly, although it is true that older subjects used a less complex syntax, their higher score on Dimension 1 deep formality shows that they were still able to produce complex texts, although using a different form of complexity. Thirdly, the present study did not find a drop in use of grammatical intricate language and/or text length but found a general negative correlation between age and these two linguistic variables. The most likely explanation to account for the present findings in the reduction of grammatical intricacy and text length is therefore that older subjects tried to be concise and direct and that they preferred other forms of complexity, as per the 'style' hypothesis suggested by Kemper *et al.* (1989) as a second explanation.

Although in general the above points indicate that it is unlikely that the effect noticed for grammatical complexity and text length is due to the subject's loss in working memory, it is not possible to prove this claim in the present study because the working memory of the participants was not measured. Kemper's findings were replicated in many studies using valid and reliable measurements of working memory and these studies therefore provide evidence that an effect of some sort exists. It could be the case that in older subjects such as the ones analysed in Kemper's studies stronger effects can be noticed that could not be found in the present study. Working memory loss and age are likely to be positively correlated and therefore it is important in the future to understand the contribution of each of these factors in explaining age-related linguistic variation.

A final remark concerns the validity of these findings in the light of Eckert's (1998) comments discussed in Section 2.2. Whether the findings observed in the present study are correlated or not with ageing can be confirmed only using a longitudinal study, as it is otherwise not possible to understand whether the variation observed is given by language change in general or by the individual's linguistic change. The present work suggests that the effects noticed are not given by language change but by ageing for at least two reasons. Firstly, the studies surveyed in the review included participants from many generations that lived in different years within the last century. It is therefore unlikely that exactly the same linguistic features are found varying in the same way even across generations and samples

that are not connected with each other. Secondly, at least for Dimension 1, a reasonable explanation exists for the ageing hypothesis: Dimension 1 Informational features are linked to literacy and ability to write and these two life-long skills are mastered only with much practice and familiarity with learned genres.

In conclusion, although it seems more likely that the relationship between Dimension 1 and ageing is not caused by general language change, only a longitudinal study could provide definite evidence to confirm this hypothesis. It is therefore extremely important that such a study be conducted in the future. Furthermore, in order to untangle the relationship between working memory and ageing, future studies should add working memory tests for the subjects and use statistical analysis that can help to understand if the cause of the linguistic patterns is indeed working memory or style.

5.3 Level of education

Table 5-7 below summarises all the variables that had an effect for level of education by showing the variable significance levels as well as their effect sizes. Since education was treated as a categorical variable with three categories, an ANOVA was run for the normally distributed linguistic variables whereas its non-parametric equivalent, the Kruskal-Wallis test, was performed on those variables that were not distributed normally. For the analysis of the FMT corpus as a whole, the non-parametric Kruskal-Wallis test was used for all the variables, since the fact that the same subjects produced more texts would have broken the ANOVA's assumption of the independence of cases. The presentation of the variables is organised according to the five major patterns introduced in the literature review of level of education in Section 2.3. All the variables that presented a statistically significant difference were organised in these categories, even in the cases of variables that were not gathered from the literature review of level of education. The variables that did not fall in any of the patterns for level of education are categorised in the general category 'Other variables'.

Table 5-7 - Linguistic variables that presented a significant effect for level of education, showing: p-value ('1-t' indicates a one-tailed value); Eta squared for the normally distributed variables only; the level of education for which the variable had an advantage (BU = below undergraduate; U = undergraduate; AU = above undergraduate; P = variable increased with education level; N = variable decreased with education level)

Variable	Task 1	Task 2	Task 3	Whole corpus
Pattern 1: Vocabulary size Higher levels of education correspond to an increase in vocabulary size and lexical sophistication				
average word length in syllables	p = 0.002 η^2 = 0.13 P			p = 0.011 P
average word length	p = 0.004 η^2 = 0.11 P			p = 0.005 P
words longer than six letters	p = 0.017 η^2 = 0.09 P			p = 0.011 AU
lexical density	p = 0.019 P			p = 0.042 (1-t) P
words longer than ten letters	p = 0.028 (1-t) η^2 = 0.14 P			p = 0.023 AU
mean rarity score		p = 0.028 (1-t) P		
Advanced Guiraud 1000		p = 0.043 (1-t) AU		
Pattern 2: Sentence complexity Higher levels of education correspond to an increase of sentential syntactic complexity				
P-Density	p = 0.001 η^2 = 0.14 U			p = 0.006 U
Dimension 5	p = 0.025 P			p = 0.008 P
coordinating conjunctions	p = 0.042 η^2 = 0.06 U			p = 0.00005 U
relative frequency of *and*		p = 0.010 BU, U		p = 0.005 U
passives		p = 0.047 (1-t) AU		

present participial clauses			p = 0.044 U	
other subordinators				p = 0.046 P
Pattern 3: T-unit complexity Higher levels of education correspond to an increase of syntactic complexity within t-units				
short t-units				p = 0.029 (1-t) N
Pattern 4: Nominal elaboration Higher levels of education correspond to an elaboration of information that focuses on nominal devices rather than verbal devices. This translates into more deep formal discourse and a higher average clause length				
analytic negations	p = 0.010 N			
total personal pronouns	p = 0.017 $\eta^2 = 0.09$ N			
first person pronouns	p = 0.020 N			
intensifiers	p = 0.036 U			
present participial WHIZ deletion relatives		p = 0.040 BU, AU		
indefinite pronouns	p = 0.048 U	p = 0.045 BU, U		p = 0.044 U
cardinal numbers	p = 0.049 AU			
singular proper nouns	p = 0.040 AU			
Dimension 1	p = 0.049 N			
pre-determiners		p = 0.005 U		
stranded prepositions		p = 0.028 U		
that relative clauses on object position		p = 0.043 AU		
downtoners			p = 0.041 BU, AU	

demonstratives				p = 0.049 N
Pattern 5: Information distribution The distribution of information in a text is different depending on how much exposure a person had to formal education. Individuals with higher education levels tend to maintain a ratio of one t-unit per sentence.				
t-units per sentence	p = 0.010 N		p = 0.038 (1-t) N	p = 0.00005 N
Other variables				
split infinitives	p = 0.021 U			p = 0.011 U
split auxiliaries	p = 0.024 U			
time adverbials		p = 0.048 U		
swear words			p = 0.020 U	
innovative stance adverbs			p = 0.042 U, AU	
total emotion words			p = 0.046 $\eta^2 = 0.07$ U	

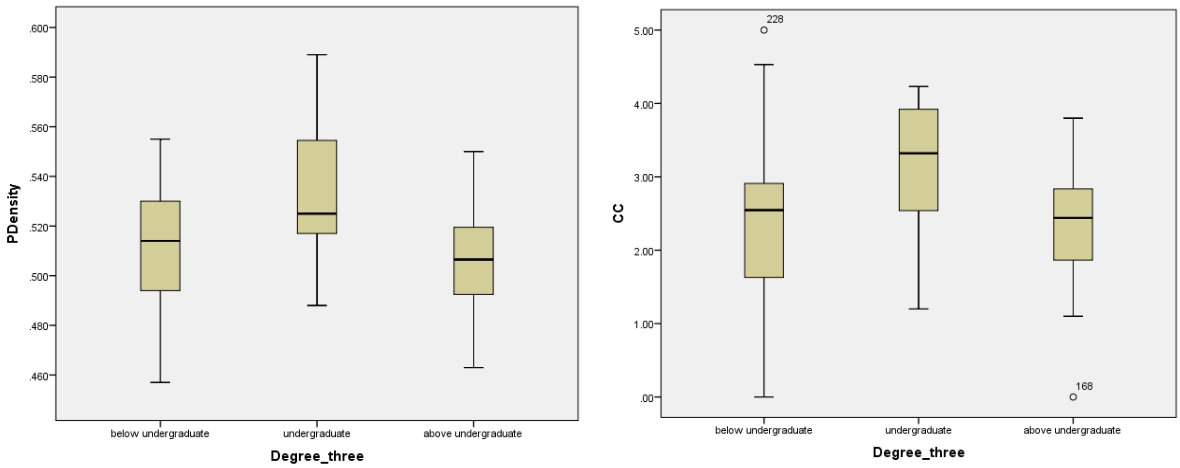
The table above shows that largely most of the linguistic patterns observed previously in other studies regarding the linguistic variation associated with level of education also apply to the FMT corpus. Similarly to the analysis of other social factors above, these effects were not evident in all the Tasks, thus indicating that register variation has a significant confounding effect. The strongest effects noticed were related to t-units per sentence, which showed a highly significant effect even when the corpus was examined on the whole. The following sections discuss the five patterns of variation introduced in Section 2.3.4 in the light of the empirical findings just presented.

The pattern related to vocabulary size is the most consistent pattern for level of education found in the FMT corpus. Table 5-7 shows that all the variables related to average word length or to rarity and sophistication of vocabulary such as Advanced Guiraud 1000 show a significant and consistent increase with level of education. However, it is possible to notice that the variables that measure intrinsic rarity, that is, the rarity of the words within the text as opposed to within the language, did not show any significant effects. This could indicate that higher education does indeed influence the vocabulary size of an individual rather than how many new words are introduced in the text. More controlled experimental work should be carried out to confirm this hypothesis.

Although many variables related to sentence complexity were found to vary with level of education in previous studies, in the FMT corpus this pattern is only partially replicated. No measures related to sentence length or to number of dependent clauses showed a significant effect for level of education. Dimension 5, the Dimension related to the abstractness of discourse characterised by use of passives showed a significant increase with level of education. More specifically, there was a significantly higher frequency of passive clauses for subjects with a postgraduate degree. This effect can be easily explained by the fact that high Dimension 5 scores are common in two genres that are frequently encountered by subjects with a postgraduate degree, that is, academic and scientific prose (Biber, 1988).

P-Density, which Mollet *et al*. (2010) found to correlate with the marks given in the assignments they examined, showed the greatest effect size for level of education in the FMT corpus. However, quite puzzlingly, P-Density was higher for subjects with only an undergraduate degree and very similar between subjects with no degree at all or with a postgraduate degree. A similar pattern was also noted for coordinating conjunctions. The distributions of these two variables can be seen in Figure 5.1 below.

Figure 5.1 - Boxplots describing the relationship between P-Density and level of education (left) and between coordinating conjunctions and level of education (right)



Since P-Density is a count of the proportion of verbs, adjectives, adverbs, prepositional phrases and conjunctions, it is possible that this variable captures non-nominal expansion of information. The analysis would thus suggest that subjects with no higher education are less likely in general to apply any sort of elaboration whereas subjects with a postgraduate degree are more likely to employ nominal elaboration of the kind measured by deep formality or Dimension 1. Therefore, the remaining type of elaboration probably measured by P-Density seems to be a prerogative of subjects with only an undergraduate degree. This same explanation could be applied to coordinating conjunctions. On one hand, subjects with a low education are not likely to expand the informational content of the text whereas on the other hand subjects with a postgraduate degree are more likely to use nominal elaboration to package information rather than simply coordinating sentences. As such, both of these

variables suggest that immature devices of elaborating information are often found in individuals with some education. That P-Density does indeed measure immaturity of information elaboration should be tested more thoroughly in the future. A large body of studies provided evidence that P-Density is linked to working memory capacity. Even though it seems unlikely that only subjects with just an undergraduate degree presented higher levels of working memory capacity, it is clear that future studies should include a measure of working memory to untangle the explaining factors underlying P-Density.

The pattern of t-unit complexity was not replicated in the FMT corpus. The only variable that presented an effect was the frequency of short t-units. Although this variable significantly decreased with education as predicted, the effect was small and no other related variable was significant. As such, it is likely that the effect found is a Type I error rather than a real effect. It is possible to conclude that education level did not affect t-unit complexity in the FMT corpus as found in previous studies.

The pattern related to nominal elaboration and clause length is partially replicated in the FMT corpus. Even though average clause length did not show any significant effect, in general a number of variables that presented an education level effect clustered in the direction predicted by this pattern. Variables such as the frequency of negations or the frequency of pronouns decreased with level of education whereas variables such as the frequency of items that elaborate noun phrases increased with education. However, there are some puzzling cases of features such as pre-determiners or indefinite pronouns in which the highest effect noticed was for subjects with undergraduate degrees only and with the other two education groups being equal. Indeed, the most important effects noted in relation to the variables presented for this pattern is the negative relationship found between personal pronouns and level of education. Since pronominal forms are negatively correlated to complex nominal form, the evidence provided by the analysis of the FMT corpus seems to suggest that increase in levels of education correspond to a decrease in use of personal pronouns. That being so, it is possible to conclude that the nominal elaboration pattern is only slightly replicated in the FMT corpus and that there is some weak evidence to be further explored in the future that it is the presence of pronouns that creates a real effect for level of education rather than an increase in nominal elaboration.

The strongest effect found in the FMT corpus for level of education was related to the distribution of t-units per sentence. Hunt's (1983) finding that the strongest difference between individuals with different education levels is the way the t-units are marked is therefore also found in the FMT corpus. The results of the analysis confirm that subjects with a higher level of education were more likely to follow a ratio of t-units per sentence equal to one, as taught in formal education. Table 5-8 below presents two examples of the t-unit per sentence analysis by showing the lowest and the highest scoring texts for t-units per sentence in Task 1.

Table 5-8 – The lowest and the highest scoring text for t-units per sentence for Task 1. T-units are marked with the hash symbol (#) whereas sentences are displayed in different paragraphs.

GAPR – 28, above undergraduate, male, SCI = 4.8, T_S = 1.07
I am writing to you regarding the travel package (Ref: 1234) I purchased from you in September last year for £999.99. # The travel package included flights, transfers and two week hotel accommodation in the Costa del Sol in Spain. # Your company brochure promised that "this package will provide you with complete confidence that everything is taken care of, allowing you to relax in luxurious surroundings and enjoy the sunshine". # In fairness, the sunshine was extremely enjoyable # but sadly the rest of the holiday failed to meet the expectation your brochure created. # In terms of being able to "relax", I can honestly say that there is nothing relaxing about being forgotten about at the airport, and unable to contact your company, for 7 hours when we arrived. # Furthermore, does your marketing department seriously think that "luxurious surroundings" include sewerage on the bathroom floor and a colony of cockroaches living under the bed? # Now, in my most sympathetic mood, I could probably forgive these as one-off mistakes. # However, what really made me angry was that your organisation's representative at the hotel did not seem to care about the problems we experienced or do anything to help us. # This leads me to believe that your organisation really does not care about your customers or their repeat custom. # However, I would like to offer you the opportunity to redeem yourselves, at least partially. # Enclosed with this letter are ten photographs which provide evidence of the conditions we faced at the hotel. # In light of the problems we experienced with your organisation, I am seeking a £500 refund from you. # I expect you to respond within 14 working days to this request. # Should this not be forthcoming, we will pass this issue on to our legal representative.
KRGA – 39, undergraduate, male, SCI = 4.3, T_S = 2.30
I'm very sorry to say that we (me and my family) is very displeased with our last travel arranged by your agency. # Last year we bought a journey to Spain for the whole family # but nothing in the package lived up to our expectations.

There was no sun for the whole trip, # the barmaid was not pleasant at all # (in fact really ugly and un polite), # the dive instructor grabbed my wives ass on several occasions # and the kids club forgot them in the water 3 times.

Even more problematic was the fact that our water in the room was off # and we have to do our “things” on a bucket at the balcony.

At night this horney man was moaning outside the door # so we could not sleep properly more than a few nights.

The sausages at the buffet looked like somebody has taken a dump on the plate # and the coffee was sweet as in USA, # the beans made me fart all day # and the car we rented leaked gasoline.

The hole fucking country hated us from day one # and all the animals were just trying to hurt us all the time!

I’m turning to the EU for sanctions on the whole island if they don’t apologize in writing!

I can assure you that I never will use your fucking company again, # and that I will start a homepage to smear all of your staff on place in Ibiza, especially Esmeralda, who refused to blow me under the table on several occasions, even thou I asked her genteelly AND offer her a little something for the trouble!

By the way, this journey to Ibiza was not worth the 10000 pounds I paid for it # and I would very much like the company, as a show of faith, refund me 8000 pounds.

The results of the analysis of t-units per sentence confirm that in the FMT corpus there is no gap between the three categories of education since the number of t-units per sentence slightly decreases and it is in a negative relationship with level of education, as predicted. Interestingly, the effect of t-units per sentence is stronger when the whole corpus is considered as a whole. This finding could be a piece of evidence that points to t-units per sentence as not being greatly influenced by register variation.

Finally, it was found in the FMT corpus that the frequency of swear words is higher in subjects with an undergraduate degree only. No literature item presented this finding and there does not seem to be a hypothesis at the moment that could explain this effect. It could be the case that the effect is caused by a series of confounding factors, including the experimental settings of the Tasks. No conclusion can be reached regarding this feature within the scope of the present work.

In conclusion, only two linguistic patterns out of the five patterns retrieved from the literature review of Section 2.3.4 showed consistency with previous literature: vocabulary size and information. Subjects with higher levels of education showed a higher vocabulary sophistication indicated by a higher extrinsic rarity of vocabulary. This lexical elaboration is combined with a smaller ratio of t-units per sentence that indicates that subjects with higher education follow the punctuation conventions of formal education. In summary, level of education showed the following patterns of linguistic variation in the FMT corpus: an increase in **extrinsic vocabulary rarity** with education level in Tasks 1 and 2, a decrease of **t-units per sentence** with education level in all the data sets but in particular in Task 1, a decrease of **personal pronouns/deep formality** with education level mostly in Task 1, and an increase

of **immature devices** in writings produced by subjects with only an undergraduate degree in general across the whole corpus.

The pattern of sentence complexity found in other studies carried out in the past was replicated in the FMT corpus only partially. Overall, the findings of the study do not support the hypothesis that syntactic complexity is higher in writings produced by subjects with a higher level of education. Nonetheless, the examination of the findings indicated that another linguistic pattern could be present in the FMT corpus: the high incidence of immature devices of informational elaboration in writings produced by subjects with only an undergraduate degree. This explanation could be also applied to other variables that resulted to be significant in the 'Other variables' section of Table 5-7. Variables such as the frequencies of split infinitives, split auxiliaries or even stranded prepositions could all be regarded as linguistic items of elaboration that appear in writings of individuals who only have limited familiarity with writing. All these variables could therefore be part of the same underlying pattern that accounts for subjects who have only some experience with literacy.

Future studies concerned on the relationship between language variation and level of education should focus on the untangling of these relationship using more advanced statistical techniques and larger samples. More generally, controlling for IQ and for working memory should help in shedding light on the underlying patterns that explain the effects noticed in the present work.

5.4 Social class

In this section, Table 5-9 below summarises all the variables that had a social class effect by showing the variable significance levels as well as the magnitude of their correlation. The correlation coefficient was calculated using Pearson's *r* for the normally distributed variables and Spearman's *rho* for the non-normally distributed variables. The presentation of the variables is organised according to the four major patterns introduced in the literature review of social class in Section 2.4.4. All the variables that presented a statistically significant difference were organised in these categories, even in the cases of variables that were not gathered from the literature review of social class. The variables that did not fall in any of the patterns for social class are categorised in the general category 'Other variables'.

Table 5-9 - Linguistic variables that presented a significant effect for social class, showing: p-value ('1-t' indicates a one-tailed value) and correlation coefficient

Variable	Task 1	Task 2	Task 3	Whole corpus
Pattern 1: Syntactic complexity Higher social classes use more complex syntax. This is likely to be caused by their greater familiarity with complex grammar				
present participial clauses	p = 0.050 *r* = 0.204			
conditionals	p = 0.006 *r* = -0.284			
concessives	p = 0.036 *r* = 0.218			p = 0.012 *r* = 0.150
average t-unit length		p = 0.0002 *r* = 0.373		p = 0.001 *r* = 0.198
clauses per t-units		p = 0.0002 *r* = 0.372		p = 0.003 *r* = 0.179
subordinating connectives		p = 0.0003 *r* = 0.370		p = 0.046 *r* = 0.120
long t-units		p = 0.001 *r* = 0.348		p = 0.001 *r* = 0.193
short t-units		p = 0.001 *r* = -0.345		p = 0.001 *r* = -0.195
Dimension 5		p = 0.002 *r* = 0.313		p = 0.003 *r* = 0.178

conjuncts		p = 0.005 *r* = 0.291		p = 0.001 *r* = 0.199
tokens		p = 0.008 *r* = 0.272		p = 0.005 *r* = 0.167
Fichtner's C			p = 0.005 *r* = -0.291	p = 0.026 *r* = -0.133
average sentence length			p = 0.006 *r* = -0.285	p = 0.032 *r* = -0.129
coordinating conjunctions			p = 0.006 *r* = -0.285	
by-passives				p = 0.027 *r* = 0.132
Pattern 2: Referential precision Higher social classes show higher precision in referencing entity in discourse than lower social classes. This end is achieved through the use of complex noun phrases. Conversely, lower social classes are more likely to use pronominal forms and exophoric references.				
total personal pronouns	p = 0.0004 *r* = -0.361	p = 0.005 *r* = -0.290		p = 0.001 *r* = -0.196
that relative clauses on subject position	p = 0.001 *r* = 0.333	p = 0.002 *r* = 0.316		p = 0.00002 *r* = 0.254
total adjectives	p = 0.003 *r* = 0.302	p = 0.001 *r* = 0.334		p = 0.00008 *r* = 0.234
deep formality	p = 0.003 *r* = 0.303			p = 0.011 *r* = 0.152
deep formality 2	p = 0.003 *r* = 0.301			p = 0.011 *r* = 0.152
first person pronouns	p = 0.003 *r* = -0.302			
pronoun *it*	p = 0.005 *r* = -0.284			
total nouns	p = 0.009 *r* = 0.268			p = 0.034 (1-t) *r* = 0.110
attributive adjectives	p = 0.008 *r* = 0.275	p = 0.0002 *r* = 0.381		p = 0.00007 *r* = 0.236

WH determiners	p = 0.025 *r* = -0.232		p = 0.040 *r* = -0.214	
Dimension 3	p = 0.033 *r* = -0.221			p = 0.043 *r* = -0.121
WH relative clauses on object position	p = 0.040 *r* = -0.213			p = 0.005 *r* = -0.168
Dimension 1	p = 0.042 *r* = -0.211	p = 0.006 *r* = -0.285		p = 0.002 *r* = -0.184
that relative clauses on object position		p = 0.006 *r* = 0.283		
Pattern 3: Use of expletives The use of expletives and their strength varies with social class. Higher social classes are less likely to swear often and/or use strong expletives.				
-	-	-	-	-
Pattern 4: Stance types Social classes are different in the types of stance that they select. Lower classes prefer to anchor their statements to the present time and are less likely to express stance overtly. On the other hand, higher classes tend to anchor their statements in the past and to express stance overtly				
present tenses		p = 0.033 *r* = -0.221		p = 0.029 (1-t) *r* = -0.114
verb bases		p = 0.045 *r* = 0.209		
evaluative adjectives			p = 0.041 *r* = -0.213	p = 0.043 *r* = -0.122
Other variables				
Advanced Guiraud 1000	p < 0.00001 *r* = 0.453	p = 0.00002 *r* = 0.430	p = 0.001 *r* = 0.346	p < 0.00001 *r* = 0.367
lexical density	p = 0.00001 *r* = 0.433	p = 0.006 *r* = 0.282	p = 0.035 *r* = 0.220	p < 0.00001 *r* = 0.263
average word length	p = 0.0001 *r* = 0.384	p = 0.00003 *r* = 0.422	p = 0.001 *r* = 0.330	p < 0.00001 *r* = 0.349

words longer than six letters	p = 0.003 *r* = 0.308	p = 0.0004 *r* = 0.361	p = 0.013 *r* = 0.257	p < 0.00001 *r* = 0.264
average word length in syllables	p = 0.006 *r* = 0.283	p = 0.00008 *r* = 0.396	p = 0.011 *r* = 0.264	p < 0.00001 *r* = 0.265
mean rarity score	p = 0.013 *r* = 0.256	p = 0.011 *r* = 0.262		p = 0.001 *r* = 0.202
lexical density H	p = 0.032 *r* = 0.222	p = 0.0001 *r* = 0.393		p = 0.00008 *r* = 0.234
words longer than ten letters	p = 0.046 *r* = 0.207	p = 0.018 *r* = 0.246		p = 0.005 *r* = 0.167
public verbs	p = 0.050 *r* = -0.204	p = 0.038 *r* = -0.216		p = 0.005 *r* = -0.167
place adverbials	p = 0.027 *r* = 0.230			p = 0.025 *r* = 0.135
t-units per sentence		p < 0.00001 *r* = -0.503	p = 0.01 *r* = -0.352	p < 0.00001 *r* = -0.332
WH questions		p = 0.025 *r* = -0.232		
positive emotion words		p = 0.040 *r* = 0.213		
time words			p = 0.022 *r* = 0.239	
be as main verb			p = 0.024 *r* = -0.235	p = 0.016 *r* = -0.144
discourse particles				p = 0.037 *r* = 0.125
genitives				p = 0.011 *r* = 0.152

The table clearly indicates that social class has the strongest effect overall among the social factors. The effect of social class is strong enough to be visible even when register variation is not controlled, as the large effects found for the whole FMT corpus suggest. However, these strong effects are not the ones predicted by the literature. The majority of the variables that showed noticeable effects for social class are collected in the 'Other variables' category of Table 5-9. In general, it is possible to notice that these variables that are classified in the 'Other variables' category are indeed the variables that were predicted to vary with level of education. Since level of education and social class are

significantly correlated in the FMT corpus (cf. Section 3.2.1), it seems to be the case that the SCI used for this study is a good proxy of education and perhaps a better predictor of general level of literacy than the education achieved. The variables that appear in the 'Other variables' category are mostly related to vocabulary richness measures, such as the mean rarity score for all the words in the text or Advanced Guiraud 1000. As an example of these measures, the highest and lowest scoring text for Advanced Guiraud 1000 in Task 2 are reproduced in Table 5-10 below.

Table 5-10 - The highest and the lowest scoring texts for Advanced Guiraud 1000 for Task 2. The words that are not present in the first 1000 types of the BNC are highlighted in bold and underlined

ANMA2 – 41, below undergraduate, male, SCI = 2.3, AG1000 = 1.75
I am writing to you as an **avid voter** and a **disappointed** one as you have asked us to **vote** for you and your **policies** and as yet you have not for **filled** them **adequately**, I am in a job that is now under **threat** of being taken away from me because of the way you are **conducting** your **policies** and your words that you have said in all your **manifestoes** are not being **upheld** because you and the **coalition** government are not being **truthful** to the people who **voted** you into power, and I really think you could do a lot more to help I really do think that it is in your power to **represent** us and get this country (England) back on its feet as the job industry is going and stop all the work going to other Countries this country used to be a **nation** of working sorry hard working people and I think that your **voters** will be **disappointed** and **upset** that you can not hold up your part of the things that you mentioned in your **manifesto** to get you **elected**, I now hope you will go back and look at what you said to get **elected** and try and **reverse** the **decisions** you have made that is **keeping** this country in such a **disarray**, and I and a lot of people won't **vote** for your government due to **disappointment**
PAKI – 25, below undergraduate, male, SCI = 4, AG1000 = 5.79 [first 230 tokens]
I write to you as a young **ambitious** Police **Officer** out on the **frontline** trying to make our country a **safer** and **happier** place to be. **Growing** up, it was my **lifelong ambition** to become a man of authority, a **role-model** for the community. To work hard, have a family and provide for them. It is now **sadly apparent** that a very large dark **cloud** is **hovering** over my head and I **fear** the **worst**. I can **assume** you **receive** many **letters** about how the recent **decline** has **affected** the **lives** of many **fellow voters**. But as a **lifelong loyal Conservative** I have always held the thought that my interests would be at the heart of the **party's fundamental belief** that **sectors** such as local **policing** should be held in high **regard**. So it was to my **disbelief** that upon **returning** home the other day, **exhausted** from a **lengthy shift** on the **beat**, I **discover** the news that you will be planning to **reduce** the **grants** given to local **policing significantly** over the next couple of years, which could very **realistically spell** the end of my **career**, one which I could not **bear** to **lose**. When the **infamous** London **Riots broke** out last year I was on first **dispatch** to the **frontline** and without **hesitation** did everything in my power to return order and bring **calm** to the community.

A visual inspection of the text samples in Table 5-10 suggests that subjects with a low Advanced Guiraud 1000 and therefore typically subjects with a low SCI used uncommon words only related to the core topic of the letters (e.g., *policy, coalition, vote*), whereas subjects with a high SCI were more likely to use uncommon words that are not related to the core topic (*disbelief, lengthy,*

infamous, dispatch, hesitation). This confirms that subjects with a high Advanced Guiraud 1000 were providing more informational content than other subjects. The fact that rarer words are typically used by subjects with a higher SCI confirms that these subjects typically present more familiarity with literacy.

However, similarly to what was observed for level of education, it is t-units per sentence that is the most significant of the linguistic variables with an effect for social class. The present experiment showed that subjects with a higher SCI tended to approach a 1:1 ratio t-units per sentence. Since a t-unit approximates a clause complex, it can be proposed that this ratio corresponds to the 1:1 clause complex per sentence ratio that is described in Halliday and Matthiessen (2004) as corresponding to the unmarked punctuation pattern in the English language. It is reasonable to assume therefore that subjects with a higher SCI were more exposed to this pattern and that therefore had more chances to learn it and internalise it.

Overall, however, the studies that claimed a relationship between syntactic complexity and social class were only partially confirmed by the empirical evidence in the FMT corpus. Indeed, even though the measures of complexity relative to t-units were positively correlated with social class, the measures of complexity related to the sentence were not. However, given the fact that a t-units per sentence ratio that approached one was significantly associated with higher SCI, it is likely that longer sentences were used by writers with lower SCI because of lack of competence in punctuating the t-units. This finding indicates that individuals with a higher SCI tend to produce complex syntax within t-units and at the same time are able to segment the t-units in the standard way. The use of passives and the degree of abstractness of discourse measured by Dimension 5 were all positively correlated with social class, as expected. Although Loban's results on length of t-unit, clauses per t-units and subordinating connectives were confirmed, the strongest of the results were obtained for those syntactic variables that elaborated the noun phrase, such as *that relative clauses on subject or object position*.

As opposed to the pattern related to syntactic complexity, the classic result of higher social classes' more frequent use of nouns and less frequent use of pronouns than lower social classes was strongly replicated in the FMT corpus. Pre-modification of nouns was also more common in writers with a higher SCI, since all kinds of adjectives were more frequent in their texts. The present study thus provides consistent and powerful evidence that Hawkins' (1977) findings are valid, even for recent times. As a result of their more precise nominal elaboration, the texts of writers with a higher SCI exhibited a lower, more informational, Dimension 1 score as well as a higher score on deep formality. Since both Dimension 1 and deep formality mostly measure the opposition between nominal and pronominal/verbal discourse orientation, this is further evidence that individuals with a higher SCI presented more nominally rich discourse.

Finally, a series of patterns noted in the previous studies were not found or only partially found in the FMT corpus. The decrease in use of expletives with SCI, for example, was not noticed in the FMT corpus. It is possible that this effect can only be found in speech or that perhaps it affects only

higher SCI bands that were not considered for this study. Similarly, the frequency of adverbs showed no effect for social class, as instead found by Macaulay (2002). It is possible that in conversation, where adverbs are more common, such a difference between classes is more evident than in writing. More research is necessary to confirm this hypothesis. Lastly, no clear pattern related to the use of the tenses was noticed, although a negative correlation between SCI and present tenses was indeed present.

In conclusion, only the difference in terms of reference to entities was consistently replicated out of the four patterns of variation that were retrieved from a literature survey on the effects of social class in linguistic variation. Nonetheless, considering the patterns discussed above, in total in the FMT corpus seven patterns of linguistic variation were noticed for social class: a positive correlation between SCI and the level of **extrinsic vocabulary rarity** that affects the whole FMT corpus, combined with a greater **density of information** in general across the whole corpus, a greater **t-unit complexity** in Task 2, an increase in the number of **noun modifications** with SCI in general across the whole corpus but more specifically in Task 1, a decrease of **t-units per sentence** with SCI across the whole corpus, and an increase of **deep formality** with SCI across the whole corpus but more specifically in Task 1. Since the last two patterns presented stronger and more consistent effects for SCI than for education, it is possible that the family background of an individual might be a more important predictor of language use than the education level they achieved.

In general, the effect of social class on the linguistic variable was the largest social effect found in the present study. The effect was indeed large enough to hide register variation at times, as could be noticed by the considerable effect sizes reached even when the whole FMT corpus was considered. Bearing in mind that social class was measured using a very simple index that only averages occupation status for the family of an individual, this result can be considered even more outstanding and worth of further inspection in the future.

Given the findings of the present study, as Johnston (1977) suggests, it seems that the greatest variation across social classes lies in the way the noun phrase is realised. Once, however, it is established that the present study suggests that nominal complexity is the factor underlying linguistic variation in social class, what is left is to explain the reason why this correlation exists. Bernstein (1960) suggests that different classes choose different codes because they interpret the same context in dissimilar ways. In the present study this claim is verifiable by qualitative exploration of the data. In Task 1 and 2, for example, subjects who used a lower lexical density and who were less Informational in Dimension 1 tended to produce a discourse that focused on the people rather than on the objects and that therefore put emphasis on relationships and on the subjective experience. This description is compatible with Bernstein's (1960) definition of the restricted code, the code used typically in the family or in contexts in which it is expected that the recipient is familiar with the content. The present study thus indicates that in Task 1 subjects differed in their interpretation of the context depending on their SCI. However, in Task 3, these differences almost faded away and this could be caused by the fact that Task 3 was

almost unequivocally interpreted by all subjects as being a context in which the restricted code should have been fully employed.

The qualitative exploration therefore confirms, as proposed by Bernstein, that subjects that lived in an environment in which the elaborated code is used more often than the restricted code were far more likely to employ the elaborated code. The missing link in this logical argument is the explanation as to why higher social classes should be exposed to the elaborated code more often. A theory that explains this link is Halliday's theory of the development and characteristics of the language of science and his theory of ideational metaphor (Halliday, 1999; Halliday and Matthiessen, 2004). Halliday explains that modern western science developed a language style that is suitable for its purposes based on the ideational grammatical metaphor. An explanation of the ideational metaphor in action is given by Halliday's (1999) example reproduced in Table 5-11.

Table 5-11 – Example of ideational metaphor from Halliday (1999). The SFL formalism was changed to traditional formalism

prolonged	*exposure*	*will result in*	*rapid*	*deterioration*	*of the item*	
adjective	**nouns**	**process**	**adjective**	**noun**	**post-modification**	
if	*the item*	*is exposed*	*for long*	*it*	*will deteriorate*	*rapidly*
conjunction	**noun**	**process**	**prepositional phrase**	**pronoun**	**process**	**adverb**
dependent clause				**independent clause**		

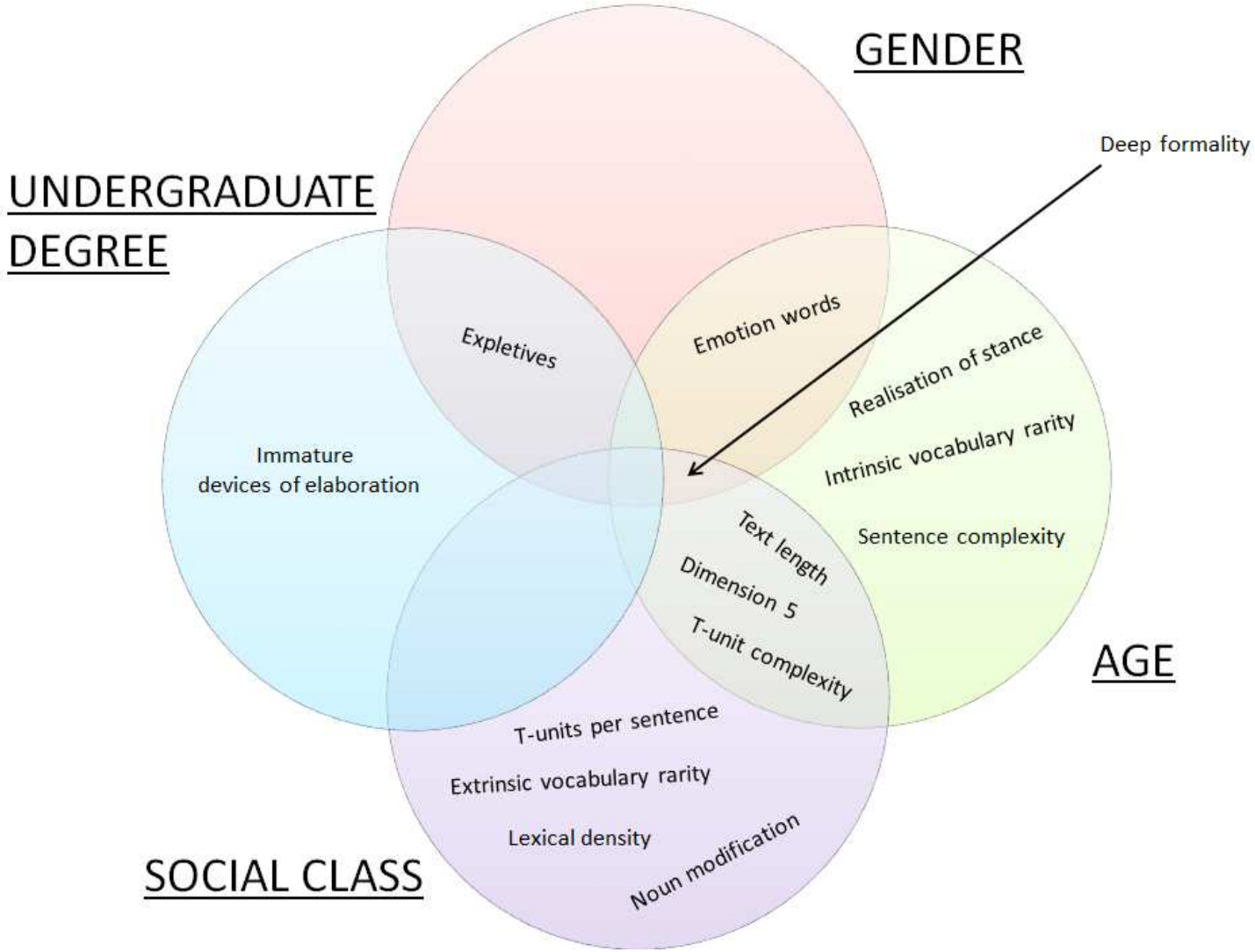
Using ideational metaphors, which correspond to a greater lexical density and to a higher use of Dimension 1 Informational features, modern science can taxonomies and organise knowledge in a way that would otherwise be impossible. Apart from more direct exposure, it could be argued that the form of logic that pervades society and that allows individuals to be successful in the highest paid jobs reflects the logic that is applied in the language of science. This logic or view of the world is not just acquired in education but also passed from generation to generation through the process of socialisation. If the nominal style was just correlated with the individual's exposure to it, then the greatest correlations in this study would have been found between nominal variables and level of education. However, since the occupation of parents is fundamental in explaining the nominal style variance in the sample examined here, it is reasonable to assume that the family environment influenced to some extent the production of ideational metaphors.

Future studies should verify this claim in experimental settings similar to the present experiment, especially controlling for other factors, such as IQ and verbal IQ. Furthermore, given the importance of these findings, it is fundamental to discover any source of variation that could influence these effects and to isolate the cause. Since the most recent and up-to-date study on social class conducted by Savage *et al*. (2013) suggested that occupation was not significantly connected with social class, it is extremely important that future replications of the present study consider these new classes adopted in Savage's *et al*. (2013) work as to allow a more complete understanding of whether the independent variable that correlates with nominal complexity is social class or perhaps occupation or some other latent phenomenon that might or might not have to do with social class.

5.5 Conclusions

With this Chapter, the present work has successfully shown that the major patterns of linguistic variation found in previous research for each of the social factors largely apply to texts of the kinds examined for the present work. The present work has also indirectly highlighted the problem that a number of linguistic variables predict more than one social factor. A summary of the patterns of linguistic variations that are shared across more than one social factor is given in Figure 5.2 below.

Figure 5.2 – A Venn diagram showing the relationship between the major patterns of linguistic variations observed in the FMT corpus and the four social factors



The Venn diagram above summarises the main findings of Chapter 5. In this diagram, the social factor level of education was substituted with just 'undergraduate degree'. This choice reflects one of the findings of this Chapter: the variables that were positively correlated with level of education showed a greater effect in the same direction for social class except for some variables that showed an effect for subjects with an undergraduate degree only.

As the diagram suggests, apart from gender, each social factor has a number of patterns of variation that do not show significant relationships with other social factors. For example, in the analysis above it was found that the number of t-units per sentence, the degree of extrinsic vocabulary rarity, the degree of lexical density and the quantity of noun modification is a characteristic of social class only. Similarly, variables related to stance, sentence complexity and intrinsic vocabulary rarity were only related to age. Finally, the degree of immature devices of elaboration distinguished only subjects with an undergraduate degree from other subjects with less or more education. However, all of the labels above refer to general patterns of variation rather than single variables and therefore it is possible that certain variables that belong to a pattern do show a relationship with more than one social factor. For example, the variable Baayen's P that here would be classified as a variables of the intrinsic vocabulary pattern, is very likely to show an effect for social class as well as age since this variable might present higher scores also in the case in which the writer uses a more extrinsically rare lexis. However, with the tools used in the present work it is not possible to untangle how much variance is accounted for by age and how much variance is accounted for by social class. It can be concluded, nonetheless, that the majority of the variance is explained by age, since only an effect of age was found using statistical tests. Further studies in the future should use more advanced models to address this problem.

Some patterns of variation varied more or less with the same magnitude for two social factors. The emotionality of a text, for example, was found to vary depending on both gender and age, even though it is not possible to assess the validity of this finding given the skewness in the sample as pointed out in Section 3.2.1. Similarly, t-unit complexity, Dimension 5 and text length were all found to vary for both age and social class. For these variables, given a certain value it is generally not possible to predict whether it is one social factor or the other (or, indeed, a combination of the two) that is determining it. As such, it is likely to be difficult to use these variables to predict social factors.

Finally, almost at the centre of the diagram, deep formality was found to vary for gender, age and social class. The reason why this variable is so central is that it summarizes almost all the other variables. A high score on deep formality is likely to correlate with a rarer vocabulary as well as with a more lexically dense text and, possibly, with less emotion words.

Chapter 5 presented the results of simple statistical tests of difference that verified whether the most important patterns of linguistic variation observed in previous studies for the four social factors considered were also present in simulated malicious forensic texts. However, the presence of a difference does not automatically imply that the pattern can be used for profiling purposes. A test of whether profiling is possible can be performed only with the help of more sophisticated statistical

techniques, such as regression analyses. The next Chapter describes such a study in which the patterns of linguistic variation that were observed to correlate with the social factors are inserted in statistical models that predict the social factors.

6 The prediction of the social factors

Chapter 5 above showed that the most important linguistic patterns of variation for the four social factors considered are largely valid for the FMT corpus. However, Chapter 5 did not address the problem of using these patterns for profiling. Chapter 6 provides an answer to this problem by using the patterns of variation found in Chapter 5 to create models that can predict the four social factors. For the sake of producing these models, in the present work a series of regression analysis is applied to the prediction of the social factors. A regression analysis is a statistical tool that allows the calculation of the relationship between some predictor variables and an outcome variable. This statistical tool is effective in determining the effect of a predictor when the other predictors are controlled, therefore indicating the contribution of the predictor independently from the others. Furthermore, a regression analysis results in an equation that can be used to predict the outcome variable using the predictors, thus estimating how powerful the predictors are when they are used to predict the outcome variable.

Regression analysis were often employed in sociolinguistics in the past under the name of *variable rule analysis* (Sankoff and Labov, 1979). However, since the sociolinguists in this paradigm were mostly interested in understanding the linguistic variable rather than in predicting the social factor, their variable rule analysis tried to predict the linguistic variable from other internal characteristic of the variable and from social factors. In the case of the present study, the same methodology is applied but with the aim of predicting the social factor, rather than the linguistic variable. In this way it is possible to check how reliable a model of profiling can be for the data set considered.

The present work uses a specific type of regression: logistic regression. This type of regression is used in cases in which the outcome variable is a binary variable, such as gender. Logistic regression estimates the probability that the outcome variable is, for example, ‘male’ or ‘female’, given the values of the predictors. At the same time, as mentioned above, logistic regression shows the contribution that each predictor makes towards the prediction with all the other predictors accounted for. In the case of the present study, logistic regression is therefore an extremely powerful tool to understand to what extent the linguistic variables can predict the social factors.

Not all the social factors in the present study were binary categorical variables such as gender, however. Age and social class were treated as continuous variables in the present study and it could be argued that they could be divided in categorical variables of three or even more categories (e.g. upper, middle and lower class). However, given that the number of subjects examined in this study is relatively low, it was chosen to adopt a logistic regression for all of the social factors by taking the median values of age and social class and then divide the subjects in two groups, as explained in Section 3.2.1. This option was chosen because the number of subjects in this study is insufficient for performing more advanced types of regression on categorical variables with more than two categories.

Another finding of the present study so far is that the linguistic variables show a highly significant effect for the Task in which they were produced. To address this problem, for each social factor four logistic regressions are carried out: one applied to the complete FMT corpus and one for each Task sub-corpus.

The Sections below present and explain the results of each of the logistic regressions carried out for the present work. Since a logistic regression should contain as few uncorrelated predictors as possible, the regressions below only include those linguistic variables that for each pattern showed the highest effect for the social factor considered.

6.1 The prediction of gender

In Section 5.1 it was concluded that two patterns of variation can be observed for gender in the FMT corpus: a **rapport/report discourse orientation** that is found only in Task 1 and a **politeness discourse orientation** pattern that is more often found in texts produced by females in personal threatening letters such as Task 3. A logistic regression model was therefore fitted to the data to predict gender using as predictors the following linguistic variables that belong to the patterns above: deep formality, swear words, and positive emotion words. The results of the regressions are in Table 6-1 below.

Table 6-1 – Table showing the results of the logistic regressions with outcome variable Gender and predictors *deep formality*, *swear words,* and *positive emotion words*. The table displays the model fit statistics and the coefficient statistics for each logistic regression. * = p ≤ 0.10; ** = p ≤ 0.05; *** = p < 0.01.

	Whole FMT corpus⁺	**Task 1**	**Task 2**	**Task 3**
Model fit	χ^2 = 23.557*** Nag. R^2 = 0.105	χ^2 = 6.136* Nag. R^2 = 0.083	χ^2 = 5.330 Nag. R^2 = 0.072	χ^2 = 23.063*** Nag. R^2 = 0.294
		Coefficients		
Variable	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*
Deep formality	**-0.053 [0.95]***	**-0.11 [0.89]****	-0.06 [0.94]	-0.03 [0.97]
Swear words	**-1.54 [0.21]****	-0.6 [0.55]	-2.45 [0.09]	**-1.51 [0.22]***
Positive emotion words	0.16 [1.17]	-0.18 [0.83]	0.23 [1.25]	**0.58 [1.78]****
Constant	2.28	6.32	2.16	-0.11

⁺ = This regression was run using Robust Standard errors to account for the fact that one person produced more than one Task.

The χ^2 statistics in the first row of the table indicates the ability of the model to explain the variance of the outcome variable given the predictors. As the significance values indicate, three models out of four were successful in predicting gender at a higher rate than chance: the model for the whole FMT corpus, the model for Task 1 and the model for Task 3. This result thus suggests that even though on a large number of cases it is possible to predict gender with a certain degree of confidence, the best results are achieved when texts like Task 3 are considered alone with some success for Task 1-like texts,

while in Task 2 it is not possible to distinguish the genders using linguistic features with any degree of certainty.

In Table 6-1, the statistics associated with the predictors are given for each model. The value expressed under the heading *B [Exp(B)]* in each column expresses the coefficient of the variable in the equation underlying the model. The p-value associated with this coefficient indicates whether the predictor makes a significant contribution towards the prediction of the outcome variable. The coefficients of the logistic regression model in Table 6-1 suggest that deep formality is a variable that contributes to the prediction of gender only in Task 1 and when the corpus is considered as a whole, whereas the variables swear words and positive emotion words contribute significantly to the prediction of gender only in Task 3. In Table 6-2 below, the two classification tables generated by the two significant models above are reproduced.

Table 6-2 – Classification table for the three significant models (whole FMT corpus, Task 1 and Task 3) that predict Gender using the variables: *deep formality, swear words* and *positive emotion words*.

			Predicted		
	Observed		Gender		Percentage correct
			Male	Female	
Whole FMT Corpus	Gender	Male	109	46	70.3
		Female	65	67	50.8
					61.3
			Predicted		
	Observed		Gender		Percentage correct
			Male	Female	
Task 1	Gender	Male	37	15	71.2
		Female	26	18	40.9
					57.3
			Predicted		
	Observed		Gender		Percentage correct
			Male	Female	
Task 3	Gender	Male	36	15	70.6
		Female	13	31	70.5
					70.5

The classification tables indicate that in the whole FMT corpus using the logistic regression above it is possible to classify the cases with a 61.3% accuracy. This accuracy is, however, only achieved because it is relatively easy to spot the male authors, since the female authors are classified with a 50% probability. When isolating Task 3, however, the results are far better, since both genders are classified with the same accuracy of more than 70%. The results of the application of the model to Task 1 are disappointing, thus suggesting that in a text similar to Task 1 it is difficult to deduce the gender of the author from their deep formality, even though this variable does show a significant effect.

These results indicate that even though the genders are different in terms of their rapport/report orientations in the corpus as a whole this difference is not large enough to predict gender. On the other hand, the politeness discourse orientation difference observed in Task 3 is distinctive enough to allow a classification of 70% of the cases but only for personal threatening letters. The findings therefore support the idea that women and men differ in their discourse orientation only in contexts that allow this difference to appear.

This study thus suggests that the profiling of gender in a forensic context is likely to be possible only when an initial register analysis is performed that confirms that the text the analyst is dealing with is a personal threatening letter. In Task 3, where the context pushes both sexes to focus on the 'rapport' rather than the 'report' aspect of the context, the difference between the genders lies in the way emotionality is managed rather than in whether the context is interpreted as being a 'rapport' or 'report' one and this is likely to be the case because both genders understand that context to be fundamentally a 'rapport' one.

6.2 The prediction of age

In Section 5.2 it was concluded that four patterns of variation can be observed for age in the FMT corpus: a **decrease of syntactic complexity** with age mostly observed in Task 1, a **decrease of Involved discourse** with age characterised by both an Involved discourse and a decrease in emotional language and observed in different shapes in all the three Tasks, an increase of **conciseness** with age that is however not found in Task 3, and the presence of different **patterns of stance realisation** that distinguish younger from older writers in Task 1. A logistic regression model was therefore fitted to the data to predict age using as predictors the following linguistic variables that belong to the patterns above: dependent clauses per sentence, *be as a main verb*, average t-unit length, Dimension 5, Baayen's P, tokens, deep formality, Dimension 4, and total emotion words. The results of the regressions are in Table 6-3 below.

Table 6-3 - Table showing the results of the logistic regressions with outcome variable Age and predictors *dependent clauses per sentence, be as a main verb, average t-unit length, Dimension 5, Baayen's P (multiplied by 100), tokens, deep formality, Dimension 4*, and *total emotion words*. The table displays the model fit statistics and the coefficient statistics for each logistic regression. Three outliers who did not use sentence boundaries and for whom the *dependent clauses per sentence* score is therefore skewed were removed from this analysis. * = p ≤ 0.10; ** = p ≤ 0.05; *** = p < 0.01.

	Whole FMT corpus⁺	**Task 1**	**Task 2**	**Task 3**
Model fit	χ^2 = 34.161*** Nag. R^2 = 0.155	χ^2 = 30.482*** Nag. R^2 = 0.374	χ^2 = 25.793** Nag. R^2 = 0.324	χ^2 = 16.840** Nag. R^2 = 0.224
		Coefficients		
Variable	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*
Dependent clauses per sentence	0.14 [1.15]	**-1.43 [0.24]***	-0.63 [0.53]	0.27 [1.3]
***Be* as main verb**	**0.35 [1.4]***	**0.68 [1.98]***	-0.31 [0.74]	0.24 [1.28]
Average t-unit length	-0.16 [0.85]	0.12 [1.13]	-0.02 [0.98]	**-0.28 [0.76]****
Dimension 5	-0.003 [1]	**-0.19 [0.83]****	0.02 [1.02]	0.07 [1.1]
Baayen's P (*100)	0.02.6 [0.13]	**0.136 [1.15]***	**0.18 [1.2]****	0.06 [1.06]
Tokens	-0.004 [1]	0.003 [1]	0.002 [1]	0.007 [1]
Deep formality	**0.07 [1.08]***	0.053 [1.05]	0.08 [1.1]	0.09 [1.1]
Dimension 4	0.03 [1.03]	-0.006 [0.99]	-0.07 [0.94]	0.08 [1.1]
Emotion words	**-0.24 [0.78]*****	-0.25 [0.78]	-0.23 [0.79]	**-0.33 [0.71]****
Constant	-0.41	-8.53	-8.12	-4.4

⁺ = This regression was run using Robust Standard errors to account for the fact that one person produced more than one Task.

The results of the regressions indicate that age can be successfully predicted in the FMT corpus both in the corpus as a whole and in each of the three Tasks. However, similarly to gender, the variables and therefore the linguistic patterns that allow this prediction are different depending on the Task. The

regressions confirm that the degree of syntactic complexity paired with the increase in intrinsic vocabulary rarity are good indicators of age only for formal letters similar to Task 1. As the formality of the letter decreases, these differences between the age groups disappear thus pointing again to the 'style' explanation rather than to the 'working memory' explanation of this effect. In Task 3, the variable that significantly predicts age is the degree of emotionality of the text. Even though it was shown in Section 3.2.1 that the distribution of age is skewed for gender in the FMT corpus, the fact that emotionality in general and not only positive emotion words distinguish age groups is evidence that it is not the skeweness in the data that is responsible for the results. The presence of only one significant variable among the predictors of age in Task 2, Baayen's P, could indicate that as the personal knowledge and/or formality of the text decreases the effect of age on syntactic complexity also decreases and that therefore in texts similar to Task 2 it is more difficult to predict age. When register variation is not controlled for, then these general patterns can still be discerned but less strongly, and this is particularly true for the conciseness pattern. Deep formality turns out to be a significant predictor of age in the whole FMT corpus but this effect disappears when register variation is controlled. An explanation of this phenomenon could be that deep formality summarizes several patterns of variation that do not present themselves if register variation is not controlled. Similarly, another explanation could be that the effect of age on deep formality is small and it can therefore only be captured when a larger data set is considered.

The classification tables generated by using the models above to predict the FMT cases are displayed in Table 6-4 below.

Table 6-4 - Classification table for the four significant models (whole FMT corpus, Task 1, Task 2 and Task 3) that predict Age using the variables: *dependent clauses per sentence, be as a main verb, average t-unit length, Dimension 5, Baayen's P (multiplied by 100), tokens, deep formality, Dimension 4*, and *total emotion words*.

Whole FMT Corpus	Observed		Predicted		
			Age		Percentage correct
			Younger 40	Older 40	
	Age	Younger 40	241	37	86.71
		Older 40	148	130	46.67
					71.6
Task 1	Observed		Predicted		
			Age		Percentage correct
			Younger 40	Older 40	
	Age	Younger 40	38	14	73.1
		Older 40	16	25	61
					67.7
Task 2	Observed		Predicted		
			Age		Percentage correct
			Younger 40	Older 40	
	Age	Younger 40	40	12	76.9
		Older 40	17	24	58.5
					68.8
Task 3	Observed		Predicted		
			Age		Percentage correct
			Younger 40	Older 40	
	Age	Younger 40	39	12	76.5
		Older 40	17	24	58.5
					68.5

The classification tables indicate that overall it is possible to deduce whether somebody is older or younger than 40 with about 70% accuracy in all the FMT texts, independently from the Task. However, as seen in Table 6-4 above, the variables that are used to reach these conclusions are different depending on the Task. In all the models except the one for Task 1, however, the percentage of correct

attributions is mostly increased by the correct classification of the authors younger than 40. In linguistic terms, the findings of these regression analyses suggest that when individuals are faced with the context of writing a formal letter that is informational in nature they tend to use more convoluted syntax if younger than 40. When dealing with a Task 3 type letter, such as a personal threatening letter, then individuals younger than 40 also tend to use more emotional language. However, as the classification tables show, even when register analysis is not accounted for, it is still possible to achieve a similar accuracy rate using a combination of both linguistic patterns.

6.3 The prediction of undergraduate degree only

The prediction of individuals with an undergraduate degree only follows a different methodology than for the other social factors. People with an undergraduate degree only in the FMT dataset were 15 (for a total of 45 cases) compared to 79 subjects with a lower or higher education. The paucity of subjects in this category means that the logistic regression can be carried out only on the complete FMT data set, since otherwise the regressions for each single Task would consist of only 15 cases, which is not enough to be able to produce meaningful results. A new table was therefore created for the purpose of this analysis by using the 45 cases produced by subjects with an undergraduate degree only compared to 22 random texts produced by people without degree and 23 produced by people with a postgraduate degree.

As reminded in Section 6, individuals with an undergraduate degree only are more likely to manifest **immature devices of elaboration**. A logistic regression model was therefore fitted to the data to predict an undergraduate degree only using as predictors the following linguistic variables that belong to the pattern above: P-Density, coordinating conjunctions, split auxiliaries, split infinitives and stranded prepositions. The results of the regressions are in Table 6-5 below.

Table 6-5 - Table showing the results of the logistic regression with outcome variable Undergraduate Degree Only and predictors *P-Density (multiplied by 100), coordinating conjunctions, split auxiliaries, split infinitives (presence/absence),* and *stranded prepositions (presence/absence).* The table displays the model fit statistics and the coefficient statistics for each logistic regression. * = p ≤ 0.10; ** = p ≤ 0.05; *** = p < 0.01.

	Whole FMT corpus[+]
Model fit	$\chi^2 = 9.4$ Nag. $R^2 = 0.07$
	Coefficients
Variable	*B [Exp(B)]*
P-Density	0.07 [1.07]
Coordinating conjunctions	0.41 [1.51]
Split auxiliaries	0.63 [1.87]
Split infinitives	**1.16 [3.2]***
Stranded prepositions	0.57 [1.76]
Constant	-5.69

[+] = This regression was run using Robust Standard errors to account for the fact that one person produced more than one Task.

The logistic regression model fitted to the whole corpus did not produce classificatory results that are higher than chance. Therefore, the model indicates that the prediction of undergraduate degree only is not possible in the FMT corpus. However, in terms of general conclusions, it is worth noting that for other social factors the best results were obtained when register variation was controlled. Given the paucity of data for this social factor it was not possible to control register variation. It is extremely likely that when register variation is accounted this model can become useful and future studies should consider testing this hypothesis.

6.4 The prediction of social class

In Section 5.4 it was concluded that eight patterns of linguistic variation can be observed for social class in the FMT corpus: a positive correlation between SCI and the level of **extrinsic vocabulary rarity** that affects the whole FMT corpus, combined with a greater **density of information** in general across the whole corpus, a greater **t-unit complexity** in Task 2, an increase in the number of **noun modifications** with SCI in general across the whole corpus but more specifically in Task 1, a decrease of **t-units per sentence** with SCI across the whole corpus, and an increase of **deep formality** with SCI

across the whole corpus but more specifically in Task 1. A logistic regression model was therefore fitted to the data to predict social class using as predictors the following linguistic variables that belong to the patterns above: Advanced Guiraud 1000, mean rarity score, average word length, lexical density, deep formality, average t-unit length, total adjectives, *that* relative clauses on subject position, t-units per sentence. The continuous social factor SCI was broken down in two classes for the present analysis that are here named Middle and Working class. These two classes mostly represent the division between managerial, administrative or professional occupations on one hand against skilled or unskilled manual occupations on the other hand. The results of the regressions are in Table 6-5 below.

Table 6-6 - Table showing the results of the logistic regressions with outcome variable Social Class and predictors *Advanced Guiraud 1000, mean rarity score, average word length, lexical density (multiplied by 100), deep formality, average t-unit length, total adjectives, that relative clauses on subject position (presence/absence),* and *t-units per sentence*. The table displays the model fit statistics and the coefficient statistics for each logistic regression. Three outliers who did not use sentence boundaries and for whom the *dependent clauses per sentence* score is therefore skewed were removed from this analysis. * = p ≤ 0.10; ** = p ≤ 0.05; *** = p < 0.01.

	Whole FMT corpus⁺	**Task 1**	**Task 2**	**Task 3**
Model fit	χ^2 = 46.801*** Nag. R^2 = 0.213	χ^2 = 36.669*** Nag. R^2 = 0.446	χ^2 = 26.890*** Nag. R^2 = 0.344	χ^2 = 11.971 Nag. R^2 = 0.168
		Coefficients		
Variable	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*	*B [Exp(B)]*
Advanced Guiraud 1000	**0.54 [1.71]***	-0.08 [0.92]	**0.88 [2.41]***	0.68 [1.98]
Mean rarity score	0.009 [1.01]	0.06 [1.06]	0.04 [1.04]	-0.05 [0.95]
Average word length	0.57 [1.77]	1.38 [3.96]	0.37 [1.45]	0.68 [1.98]
Lexical density	0.04 [1.05]	**0.45 [1.57]*****	-0.06 [0.94]	0.03 [1.03]
Deep formality	**-0.09 [0.91]*****	**-0.14 [0.87]***	-0.11 [0.89]	-0.07 [0.93]
Average t-unit length	-0.01 [0.99]	0.01 [1.01]	0.05 [1.05]	-0.08 [0.92]
Total adjectives	0.11 [1.11]	0.009 [1]	0.2 [1.22]	0.06 [1.06]

***that* relative clauses on subject position**	**0.99 [2.69]*** **	**2.64 [13.98]*** **	0.65 [1.91]	0.3 [1.35]
t-unit per sentence	**-1.21 [0.3]***	1.45 [4.27]	**-2.09 [0.12]***	**-1.67 [0.19]** **
Constant	-0.43	-20.21	3.89	1.46

[+] = This regression was run using Robust Standard errors to account for the fact that one person produced more than one Task.

As the χ^2 tests indicate, out of four logistic regression models, three allowed a prediction significantly higher than chance: the model for the whole corpus and the models for Task 1 and Task 2. The high significance level obtained by these models indicates that among all of the social factors the prediction of social class is the one more likely to be accurate. Even though the variables that significantly contribute to the prediction slightly vary across the different models, the basic pattern that is present in all of the models is the higher density of information especially in the form of complex noun phrases for the individuals classified as middle class. The management of t-units is a significant predictor of class in all the models except for the model for Task 1. Even in Task 3, where the statistics indicate that class could not be predicted with any degree of reliability, the variable t-units per sentence is still a significant predictor of class. This confirms again that the management of the t-units varies greatly by social factors than by register variation, unlike almost all the other variables examined in the present study. The classification tables generated by using the models above to predict the FMT cases are displayed in Table 6-7 below.

Table 6-7 - Classification table for the four significant models (whole FMT corpus, Task 1, and Task 2) that predict Social Class using the variables: *Advanced Guiraud 1000, mean rarity score, average word length, lexical density (multiplied by 100), deep formality, average t-unit length, total adjectives, that relative clauses on subject position (presence/absence),* and *t-units per sentence*.

			Predicted		
Whole FMT Corpus	Observed		Age		Percentage correct
			Working	Middle	
	Age	Working	93	41	69.4
		Middle	44	91	67.4
					68.4
			Predicted		
Task 1	Observed		Class		Percentage correct
			Working	Middle	
	Class	Working	36	9	80
		Middle	11	34	75.6
					77.8
			Predicted		
Task 2	Observed		Class		Percentage correct
			Working	Middle	
	Class	Working	33	12	73.3
		Middle	11	34	75.6
					74.4

The high percentages of re-classification confirm the results of the logistic regression model. The best results for class as well as for any other model presented above are observed for Task 1, where almost 78% of the participants could be correctly attributed to middle or working class just by automatically analysing their language. The reason for this result is likely to lie in the nature of the register of Task 1. Since Task 1 requires a certain level of formality, middle class individuals, who come from managerial or administrative positions, are more used to the kind of language that is appropriate to a formal context. In other words, this analysis confirms Bernstein's (1960) restricted/elaborated code model and it confirms that using this knowledge it is possible to predict the class of an individual using a linguistic analysis. This same theoretical explanation also manages to explain why it is not possible to successfully predict class in Task 3. As Bernstein (1962) explains, the

codes are heavily predicted by the situational aspects of the extra-linguistic context. In a situation such as the one simulated by Task 3 it is likely that even middle class individuals use a restricted code as this is the most appropriate code to use in this situation. The regression models therefore provide further evidence to the usefulness of Bernstein's theory for the forensic profiling of social class.

6.5 Conclusions

Chapter 6 has dealt with the problem of using the patterns of linguistic variation found in Chapter 5 for the profiling of the social factors. The results of the regression models show that the profiling of the social factors can be performed and that the performance of the models is usually increased by controlling for register variation. Table 6-8 below summarises all the results.

Table 6-8 – Summary of the results of Chapter 6. Each cell reports the percentage of reclassification of a regression model for a specific social factor in a specific Task or for the whole FMT corpus. Below the percentage, the variables that contributed to the prediction are reported in order of significance. An "N/A" was used for those models for which the χ^2 test was not significant.

	Whole FMT corpus	**Task 1**	**Task 2**	**Task 3**
Gender	**61.3%** swear words, deep formality	**57.3%** deep formality	N/A	**70.5%** positive emotion words, swear words
Age	**71.6%** emotion words, *be* as main verb, deep formality	**67.7%** Dimension 5, Baayen's P, *be* as main verb, dependent clauses per sentence	**68.8%** Baayen's P	**68.5%** emotion words, average t-unit length
Undergraduate degree only	N/A			
Social class	**68.4%** deep formality, *that* relative clauses on subject position, Advanced Guiraud 1000, t-unit per sentence	**77.8%** lexical density, that relative clauses on subject position, deep formality	**77.4%** t-unit per sentence, Advanced Guiraud 1000	N/A

Table 6-8 summarises all the results of Chapter 6. These results suggest that gender is the most difficult social factor to profile as opposed to social class, which is the easiest social factor to profile. The only social factor that did not report any result was 'undergraduate degree only', although it is very likely that the absence of results is dependent on the paucity of data. In all the cases except for age, the isolation of the register, which was performed by controlling the Task, resulted in better performances

of the models. More importantly, the variables that significantly contributed to the prediction of the social factors always varied from Task to Task. This result indicates that it is essential that the register of the text is identified before the profiling analysis is performed. The results of the models used to predict the social factors suggest that the general patterns of variation identified in Chapter 5 can be used for profiling, even though, generally speaking, the results are valid only for about 70% of the cases. This result is in line with previous literature, although this is the first time these results are achieved on such a small data set made up of short simulated malicious forensic texts.

The results of the analyses of Chapter 6 suggest that when profiling gender, in a text similar to Task 1, the difference between the *report* discourse orientation typically employed by male authors as opposed to the *rapport* discourse orientation typically employed by female authors is successful in distinguishing the genders only with a 61% accuracy. However, for personal texts such as Task 3, a politeness discourse orientation pattern based on a combination of positive emotion words and swear words is useful in predicting gender with a 70% accuracy. For age, the main finding consisted in the fact that as individuals get older, when producing formal texts such as Task 1, the level of syntactic complexity that they produce decreased while their level of conciseness increased. This finding is compatible with previous literature that found similar effects in other registers and in the present data set the knowledge of the individual's performance on these two dimensions allowed the prediction of being older or younger than 40 with about 70% accuracy. In texts such as Task 3, similar level of accuracy in prediction were reached only by looking at the degree of Involved and emotional discourse, which decreased with age as predicted by previous studies. For level of education and social class, similar findings were obtained that generally indicated that individuals with a low social class index or low education are less likely to produce uncommon lexical items and complex grammatical patterns independently from the register. A finding of the present study is that social class is a far better predictor of these two linguistic patterns than education, thus indicating that the family environment of the subjects greatly contributed to their linguistic production. Social class was also the social factor that could be predicted with the greatest accuracy. It was possible to determine whether somebody belonged to middle or working class background with an accuracy of almost 80% on the basis of vocabulary and grammar use, as predicted by the literature and, in particular by Bernstein's (1960) code theory.

In conclusion, Chapter 6 showed how the patterns of linguistic variation observed in previous studies and found to be present also in fabricated malicious texts can be applied to produce models that can profile the social factors. These models produce results that are on average 70% accurate. However, register variation should be accounted for before analysing a text for profiling. The next Chapter concludes the present work with a summary of the dissertation.

7 Final conclusions

The aim of the present work consisted in making a first step towards filling a series of gaps in the present state of the art of forensic authorship profiling. These gaps were listed in Section 1.2 as follows:

1. The lack of a systematic summary of the relationship between linguistic variation and a range of social variables, including gender, age, level of education and social class;
2. The lack of integration of linguistic theory into current research on authorship profiling and, consequently, a general disregard for the importance of register variation;
3. The lack of research in authorship profiling based directly on malicious texts, such as threatening letter, ransom demands, etc.;
4. The lack of an objective methodology or protocol for authorship profiling in the forensic context.

This project intended to make a first step towards filling in these gaps by carrying out a study that involved looking at the already established patterns of sociolinguistic variation in texts that resemble malicious forensic texts produced in an experimental condition that controls for register variation.

To do so, the present work started with an exploration of the literature in Chapter 2 that highlighted a series of previous findings on the correlations between language use and the social factors gender, age, level of education and social class. These findings were organised in patterns of linguistic variation that include several linguistic variables. Even though these patterns were meant to be collected in order to be tested on the sample of fabricated malicious texts produced in an experimental setting, this comprehensive literature review by itself makes a significant step towards filling gap (1). Future researchers can use the literature review as a point of departure for further studies.

Since an experimental situation that involved fabricated texts was chosen, the problem of the validity of the data and of generalisations to authentic data sets was solved by collecting a set of authentic malicious texts and testing for the linguistic comparability between the fabricated and the authentic data set. This comparison is reported in Chapter 4 and consisted in two parts: (1) an analysis of the two corpora using Biber's (1988; 1989) multidimensional analysis framework and (2) a test of statistically significant difference between the corpora for all of the linguistic variables gathered from the literature review of Chapter 2. The analysis revealed that both authentic and fabricated malicious texts were similar to professional or personal letters and that their average Dimensions scores and text type distribution was strikingly similar. Only 13 out of 135 variables tested presented significantly different results between the two data sets and, after a careful examination, only two of those 13 variables were found to present this difference because of the experimental conditions of the fabricated texts: the frequency of contractions and the frequency of proper nouns. For most of the remaining 11 variables, as well as for the slight differences in the Dimensions scores and text types, the explanations

for the difference consisted in the fact that whereas the fabricated texts were equally divided in only three simulated communicative situations, the authentic malicious texts corpus contained a variety of communicative situations. The results indicate that there is no important difference between the fabricated and the authentic data and that therefore the experimental conditions of the fabricated corpus did not affect the linguistic production of the subjects. From these findings it follows that any results of the sociolinguistic analysis of the fabricated data can be extended to real data.

After the comparison between the two data sets allowed the confirmation that the fabricated data set can be used as proxy for a sociolinguistic analysis, Chapter 5 reported the verification of the patterns of variation found in the literature for the fabricated corpus. The aim of this Chapter corresponded to providing an answer to the gaps (2) and (3), that is, the lack of research regarding the profiling of malicious forensic texts and the disregard of register variation. The results of Chapter 5 indicate that previous findings on the relationship between gender, age, level of education and social class and language also apply to malicious forensic texts. However, the present study highlighted that register variation has to be controlled as this increases considerably the power of profiling social factors since the patterns that predict a social factor for one register might be different from the patterns that predict that same social factor for another register. For gender, for example, it was found that whereas in Task 1 it is the opposition between nouns and pronouns that presented a gender effect, in Task 3 it was the opposition between positive emotion words and swear words that presented a gender effect. For deep formality, the score of an uneducated low SCI person for Task 1 was the same as the score of an educated high SCI person for Task 3. As these two examples suggest, the determination of which register the analyst is dealing with is a priority and precursory to the task of authorship profiling. Chapter 5 has also pointed out that what matters is the general linguistic pattern rather than a single variable, since a general tendency compatible with the literature is not always found using exactly the same linguistic variables. Generally speaking, the fact that the same functional patterns found in other registers were also found in malicious forensic texts adds to the validity of these patterns to be used in forensic scenarios and it also validates them for future uses, even though care should be taken as register variation can confound the results. In conclusions, the findings of Chapter 5 move the area of research of authorship profiling towards a better understanding of the relationship between the four social factors here selected and language use.

After confirming the presence of the major patterns of linguistic variations found in previous studies for the FMT corpus, Chapter 6 presented the results of a series of regression analyses aimed at predicting the social factors. The regression analyses performed in this Chapter confirm that in general the social factors can be predicted with about 70% accuracy even in texts as short as the ones considered for the present study, as long as the register of the text is accounted for. The results of this study address gap number (4), as this study provides an initial list of models that constitute a first step towards a systematic method for profiling malicious texts. Thanks to the contextualisation of these models in previous research coming from several disciplines, it is possible to use these models to answer various

research questions that can be useful to improve the accuracy of the profile. Future research should use these findings as a start to generate new hypothesis which will in turn lead to better profiling protocols.

In conclusion, the present study suggests that social factors can be profiled using patterns of linguistic variables. This knowledge can be used in the future to create a methodology of forensic authorship profiling. However, contrarily to the current ideology that is evident in some of the works carried out for automatic authorship profiling, the present study also suggests that the automatic profiling of the author of a text is still far away. Given the complex interrelationships between the linguistic variables, the social factors and the register of the text, it is unlikely that a computer program can profile efficiently the author of a disputed text with the present state of the art, even though this might be possible at some stage in the future. At the moment, it is more likely that the task of the computer is to aid the analyst to spot and quantify features that would otherwise be unnoticed or unquantified. With the aid of a trained human analyst, however, the present study suggests that authorship profiling can be successfully applied, since many general patterns of social differentiation in the use of certain linguistic variables are evident when the register is controlled. The first step towards a standard protocol is to confirm that the findings of this study are replicated in similar experiments on new data. The regression models presented in Chapter 6 represent a start towards this endeavour. Future research should be dedicated to the confirmation or falsification of these models in the light of linguistic and non-linguistic theories. Following this step, more conclusions can be drawn on the nature of the linguistic variation exhibited by the social factors and on what these patterns actually suggest to the analyst. However, the present work points out that two elements should be at the basis of this future protocol: (a) a focus on register analysis as a preliminary step to be performed before the profiling; and (b) a focus on general patterns of linguistic variation rather than on single variables. A future method of authorship profiling would therefore be compatible with the algorithm recently proposed by Nini and Grant (2013), in which the analysis of register is foregrounded and considered essential for the analysis of social variation and paired with multivariate analysis of the co-variation of many linguistic variables. Most importantly, the findings of the present work strongly suggest that future research should focus on expanding our present knowledge of the linguistic variables, which should remain at the centre of the linguistic analysis of forensic texts.

8 References

Argamon, S., Dhawle, S., Koppel, M. and Pennebaker, J. W. (2005) Lexical predictors of personality type, In *Proceedings of Classification Society of North America*, St. Louis.

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), pp. 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), p. 119.

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

Bamman, D., Eisenstein, J. and Schnoebelen, T. (2012) Gender in Twitter: Styles, stances, and social networks, *arXiv preprint arXiv:1210.4567*.

Barbieri, F. (2008) Patterns of age-based linguistic variation in American English, *Journal of Sociolinguistics*, **12**(1), pp. 58–88.

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

Berman, R., Nayditz, R. and Ravid, D. (2011) Linguistic diagnostics of written texts in two school-age populations, *Written Language and Literacy*, John Benjamins Publishing Company, **14**(2), pp. 161–187.

Bernstein, B. (1960) Language and social class, *British Journal of Sociology*, **11**(3), pp. 271–276.

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

Biber, D. (1989) A typology of English texts, *Linguistics*, **27**(1), pp. 3–43.

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), pp. 9–37.

Biber, D. (1993) Representativeness in Corpus Design, *Literary and Linguistic Computing*, **8**(4), pp. 243–257.

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

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.

Biber, D. and Finegan, E. (1989) Styles of stance in English: Lexical and grammatical marking of evidentiality and affect, *Text*, **9**(1), pp. 93–124.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *The Longman Grammar of Spoken and Written English*, Harlow, Longman.

Biber, D. and Jones, J. (2005) Merging corpus linguistic and discourse analytic research goals: Discourse units in biology research articles, *Corpus Linguistics and Linguistic Theory*, **1**(2), pp. 151–182.

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

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

Brown, C., Snodgrass, T., Kemper, S., Herman, R. and Covington, M. (2008) Automatic measurement of propositional idea density from part-of-speech tagging, *Behavior Research Methods*, **40**(2), pp. 540–545.

Byrd, M. (1993) Adult age differences in the ability to write prose passages, *Educational Gerontology: An International Quarterly*, **19**, pp. 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), pp. 385–407.

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

Cheshire, J. (2005) Syntactic variation and beyond: Gender and social class variation in the use of discourse-new markers, *Journal of Sociolinguistics*, **9**(4), pp. 479–508.

Cheung, H. and Kemper, S. (1992) Competing complexity metrics and adults' production of complex sentences, *Applied Psycholinguistics*, **13**, pp. 53–76.

Cloran, C. (1989) Learning through language: the social construction of gender, In *Language Development: Learning Language, Learning Culture*, Hasan Martin, J, R. (ed.), Norwood, NJ, Ablex Publishing Corporation, pp. 111–151.

Cosgrove, K. P., Mazure, C. M. and Staley, J. K. (2007) Evolving knowledge of sex differences in brain structure, function, and chemistry, *Biological Psychiatry*, **62**(8), pp. 847–55.

Covington, M. (2012) CPIDR, [online] Available from: http://www.ai.uga.edu/caspr/.

Crosby, F. and Nyquist, L. (1977) The female register: An empirical study of Lakoff's hypotheses, *Language in Society*, Cambridge Univ Press, **6**, pp. 313–322.

Dodsworth, R. (2011) Social class, In *The SAGE Handbook of Sociolinguistics*, Wodak, R., Johnstone, B., and Kerswill, P. (eds.), London, SAGE, pp. 192–208.

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

Eckert, P. (1998) Age as a sociolinguistic variable, In *The Handbook of Sociolinguistics*, Coulmas, F. (ed.), Oxford, UK; Cambridge, Mass, Blackwell Publishers, pp. 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**, pp. 461–490.

Engelman, M., Agree, E. M., Meoni, L. A. and Klag, M. J. (2010) Propositional density and cognitive function in later life: Findings from the precursors study, *Journal of Gerontology: Psychological Sciences*, **65B**(6), pp. 706–711.

Fast, G. (n.d.) Lingua::EN::Syllable Perl module, [online] Available from: http://search.cpan.org/~gregfast/Lingua-EN-Syllable-0.251/Syllable.pm.

Fast, L. a and Funder, D. C. (2010) Gender differences in the correlates of self-referent word use: authority, entitlement, and depressive symptoms, *Journal of Personality*, **78**(1), pp. 313–38.

FBI (n.d.) FBI Vault, [online] Available from: http://vault.fbi.gov/.

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

Flesch, R. (1949) *The Art of Readable Writing*, New York, Harper.

Foster, D. (2001) *Author Unknown: On the Trail of Anonymous*, London, Macmillan.

Francis, H. (1974) Social-class, reference and context, *Language and Speech*, **17**, pp. 193–198.

Fraser, B. (1998) Threatening revisited, *Forensic Linguistics*, **5**(2), pp. 159–173.

Gleitman, H., Fridlund, A. J. and Reisberg, D. (2004) *Psychology*, New York; London, W. W. Norton.

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

Halari, R., Hines, M., Kumari, V., Mehrotra, R., Wheeler, M., Ng, V. and Sharma, T. (2005) Sex differences and individual differences in cognitive performance and their relationship to endogenous gonadal hormones and gonadotropins, *Behavioral Neuroscience*, **119**(1), pp. 104–17.

Halliday, M. (2004) The spoken language corpus: a foundation for grammatical theory, In *Advances in corpus linguistics: papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg 22-26 May 2002*, Rodopi, pp. 11–38.

Halliday, M. A. K. (1999) The grammatical construction of scientific knowledge: The framing of the English clause, In *Incommensurability and Translation: Kuhnian Perspectives on Scientific Communication and Theory Change*, Rossini Favretti, R., Sandri, G., and Scazzieri, R. (eds.), Chelthenham, Edward Elgar.

Halliday, M. A. K. (2004) *The Language of Science*, Webster, J. (ed.), New York; London, Continuum.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004) *An Introduction to Functional Grammar*, London, Arnold.

Harley, T. a., Jessiman, L. J. and MacAndrew, S. B. G. (2011) Decline and fall: A biological, developmental, and psycholinguistic account of deliberative language processes and ageing, *Aphasiology*, **25**(2), pp. 123–153.

Hasan, R. (2009) A sociolinguistic interpretation of everyday talk between mothers and children, In *The Collected Works of Ruqaiya Hasan Vol. 2: Semantic Variation: Meaning in Society and in Sociolinguistics*, Webster, J. (ed.), London, Equinox, pp. 73–118.

Hawkins, P. R. (1969) Social class, the nominal group and reference, *Language and Speech*, **12**(2), pp. 125–135.

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), pp. 439–459.

Heylighen, F. and Dewaele, J. (1999) Variation in the contextuality of language: an empirical measure, *Interner Bericht, Center "Leo Apostel", Vrije …*.

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

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

Jenkinson, T. and Weymouth, A. (1976) Pronominal usage, cohesion and explicitness in working-class speech: Towards an evaluative technique, *Language and Speech*, SAGE Publications, **19**(2), pp. 101–116.

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), p. 317.

Kaiser, A., Haller, S., Schmitz, S. and Nitsch, C. (2009) On sex/gender related similarities and differences in fMRI language research., *Brain Research Reviews*, Elsevier B.V., **61**(2), pp. 49–59.

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

Kemper, S., Bontempo, D., Mckedy, W., Schmalzried, R., Tagliaferri, B. and Kieweg, D. (2011) Tracking sentence planning and production, *Journal of Gerontology: Psychological Sciences*, **66B**(2), pp. 160–168.

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), pp. 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), pp. 49–66.

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

Kemper, S., Thompson, M. and Marquis, J. (2001) Longitudinal change in language production: Effects of aging and dementia on grammatical complexity and propositional content, *Psychology and Aging*, **16**(4), pp. 600–614.

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), pp. 401–412.

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

Lakoff, R. (1973) Language and woman's place, *Language in Society*, **2**(01), pp. 45–80.

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*, Berkeley.

Macaulay, R. (2002) Extremely interesting, very interesting, or only quite interesting? Adverbs and social class, *Journal of Sociolinguistics*, **6**(3), pp. 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**, pp. 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), pp. 429–473.

Mulac, A., Bradac, J. J. and Gibbons, P. (2001) Empirical support for the gender-as-culture hypothesis. An intercultural analysis of male/female language differences, *Human Communication Research*, **27**(1), pp. 121–152.

Mulac, A. and Lundell, T. L. (1994) Effects of gender-linked language differences in adults' written discourse: Multivariate tests of language effects, *Language & Communication*, **14**(3), pp. 299–309.

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), pp. 211–236.

Nini, A. (2014) Multidimensional Analysis Tagger 1.0 - Manual, [online] Available from: https://sites.google.com/site/multidimensionaltagger/.

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), pp. 173–202.

O'Barr, W. M. and Atkins, B. K. (1980) "Women's language" or "powerless language"?, In *Women and Languages in Literature and Society*, McConnell-Ginet, S., Borker, R., and Furman, N. (eds.), New York, Praeger, pp. 93–110.

Olsson, J. (2003) *Forensic Linguistics: An Introduction to Language, Crime and the Law*, London, Continuum.

Pennebaker, J. W. (2011) *The Secret Life of Pronouns: What our Words Say about us*, New York; London, Bloomsbury Press.

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*, Citeseer, **113**(1), pp. 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**, pp. 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), pp. 291–301.

Plum, G. A. and Cowling, A. (1987) Social constraints on grammatical variables: Tense choice in English, In *Language Topics: Essays in Honour of Michael Halliday Vol. 2*, Steele, R. and Threadgold, T. (eds.), Amsterdam, J. Benjamins, pp. 281–305.

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

Poole, M. E. (1973) Comparison of factorial structure of written coding patterns for a middle-class and a working-class group, *Language and Speech*, **16**, pp. 93–109.

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

Poole, M. E. (1983) Socioeconomic status and written language, In *The Psychology of Written Language: Developmental and Educational Perspectives*, Martlew, M. (ed.), New York, John Wiley, pp. 335–376.

Pustet, R. (2004) Zipf and his heirs, *Language Sciences*, **26**(1), pp. 1–25.

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), pp. 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), pp. 133–152.

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

Rushton, J. and Young, G. (1975) Context and complexity in working-class language, *Language and Speech*, **18**, pp. 366–387.

Saily, 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).

Sankoff, D. and Labov, W. (1979) On the uses of variable rules, *Language in Society*, **8**(2), pp. 189–222.

Savage, M., Devine, F., Cunningham, N., Taylor, M., Li, Y., Hjellbrekke, J., Le Roux, B., Friedman, S. and Miles, A. (2013) A new model of social class? Findings from the BBC's Great British Class Survey experiment, *Sociology*, **47**(2), pp. 219–250.

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*.

Schmid, H.-J. (2003) Do women and men really live in different cultures? Evidence from the BNC, In *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.), Frankfurt, Peter Lang, pp. 185–221.

Shuy, R. (1996) *Language Crimes: Use and Abuse of Language Evidence in the Court Room*, Oxford, Blackwell.

Solan, L. and Tiersma, P. (2005) *Speaking of Crime: The Language of Criminal Justice*, Chicago, University of Chicago Press.

Soto, C. J., John, O. P., Gosling, S. D. and Potter, J. (2011) Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample, *Journal of Personality and Social Psychology*, **100**(2), pp. 330–48.

Spencer, E., Craig, H., Ferguson, A. and Colyvas, K. (2012) Language and ageing - exploring propositional density in written language - stability over time, *Clinical Linguistics & Phonetics*, **26**(9), pp. 743–754.

Springer, S. P. and Deutsch, G. (1997) *Left Brain, Right Brain: Perspectives from Cognitve Neuroscience*, *Neuroscience*, 5th ed, Basingstoke, W. H. Freeman.

Tausczik, Y. R. and Pennebaker, J. W. (2009) The psychological meaning of words: LIWC and computerized text analysis methods, *Journal of Language and Social Psychology*, **29**(1), pp. 24–54.

Toutanova, K., Klein, D., Manning, C. and Singer, Y. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network, In *Proceedings of HLT-NAACL 2003*, pp. 252–259.

Witte, S. P. and Davis, A. S. (1980) The stability of T-unit length: A preliminary investigation, *Research in the Teaching of English*, JSTOR, **14**(1), pp. 5–17.

Wodak, R. and Benke, G. (1998) Gender as a sociolinguistic variable: New perspectives on variation studies, In *The Handbook of Sociolinguistics*, Coulmas, F. (ed.), Oxford, UK; Cambridge, Mass, Blackwell Publishers, pp. 127–150.

9 APPENDIX

9.1 List of AMT texts

The table below displays all the texts of the AMT corpus. The texts that did not reach the text length requirement of 100 tokens are displayed with their text length highlighted. The asterisk next to an element within a cell indicates that the information is uncertain.

Code	Author ID	Tokens	Name	Gender	Age	Nationality	Year	Source
UNK53_01	63	654	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Al Gore, Sr., File 1)
UNK67_01	77	227	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Carl Sagan, File 1)
UNK26_01	33	**57**	Charles Augustus Livezly	Male	60	American	N/A	FBI Vault (Claudia Johnson, File 1)
UNK26_02	33	**57**	Charles Augustus Livezly	Male	60	American	N/A	FBI Vault (Claudia Johnson, File 1)
UNK27_01	34	**67**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Claudia Johnson, File 1)
ERHA_01	97	582	Eric Harris	Male	18	American	1999	FBI Vault (Columbine High School, File 1)
ERHA_02	97	184	Eric Harris	Male	18	American	1999	FBI Vault (Columbine High School, File 1)
UNK72_01	82	111	Unknown	Unknown	Unknown	American*	1939	FBI Vault (Eddie Cantor, File 3)
UNK73_01	83	240	Unknown	Unknown	Unknown	American*	1942	FBI Vault (Eddie Cantor, File 4)

UNK24_01	31	174	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Eisenhower, File 2)
UNK24_02	31	171	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Eisenhower, File 2)
UNK25_01	32	165	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Eisenhower, File 5)
UNK50_01	59	**98**	Unknown	Unknown	Unknown	Unknown	N/A	FBI Vault (Eleanor Roosevelt, File 27)
UNK51_01	60	**79**	Unknown	Unknown	Unknown	American*	1941	FBI Vault (Eleanor Roosevelt, File 27)
UNK52_01	61	397	Unknown	Unknown	Unknown	American*	1943	FBI Vault (Eleanor Roosevelt, File 31)
UNK56_01	66	**79**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Elizabeth Taylor, File 1)
UNK57_01	67	307	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Elizabeth Taylor, File 1)
LLCL_01	62	329	Lloyd K. Clemens	Male	Unknown	American	N/A	FBI Vault (Eugene McCarthy, File 2)
UNK43_01	52	541	Unknown	Male	65	American	N/A	FBI Vault (Frances Perkins, File 14)
UNK43_02	52	424	Unknown	Male	65	American	N/A	FBI Vault (Frances Perkins, File 15)

UNK82_01	92	571	Unknown	Unknown	Unknown	American*	1951	FBI Vault (Frank Sinatra, File 4)
UNK82_02	92	1026	Unknown	Unknown	Unknown	American*	1951	FBI Vault (Frank Sinatra, File 4)
UNK83_01	93	183	Unknown	Unknown	Unknown	American*	1976	FBI Vault (Frank Sinatra, File 6)
UNK84_01	94	165	Unknown	Male	Unknown	American	1981	FBI Vault (Frank Sinatra, File 7)
UNK85_01	95	1602	Unknown	Female	Unknown	American*	1985	FBI Vault (Frank Sinatra, File 7)
UNK86_01	96	**51**	Unknown	Unknown	Unknown	American*	1980	FBI Vault (Frank Sinatra, File 7)
GJBR_01	25	560	Unknown	Unknown	Unknown	American	N/A	FBI Vault (George Jackson Brigade, File 1)
UNK63_01	73	153	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (George Steinbrenner, File 12)
UNK64_01	74	1080	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (George Steinbrenner, File 12)
UNK64_02	74	1136	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (George Steinbrenner, File 12)
UNK23_01	30	1101	Unknown	Unknown	Unknown	N/A	N/A	FBI Vault (Humphrey, File 15)
UNK21_01	28	303	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Humphrey, File 4)
UNK22_01	29	118	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Humphrey, File 8)

UNK28_01	35	**99**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 1)
UNK29_01	36	**55**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 1)
UNK30_01	37	181	Unknown	Male	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 1)
UNK31_01	38	108	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 1)
UNK32_01	39	187	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 2)
UNK33_01	40	**90**	Unknown	Male	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 2)
UNK33_02	40	**72**	Unknown	Male	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 2)
PLPC_01	41	**80**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 2)
PLPC_02	41	304	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 2)
UNK35_01	43	**54**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 3 or 4)

UNK36_01	44	**59**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 3 or 4)
UNK34_01	42	116	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 3)
UNK37_01	45	**76**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 5)
UNK38_01	46	**45**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 6)
UNK39_01	47	**45**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 7)
WASM_01	48	129	Walter Smalley	Male	74	American	N/A	FBI Vault (Jesse Helms, File 7)
UNK40_01	49	**48**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Jesse Helms, File 7)
UNK54_01	64	164	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (John Murtha, File 11)
UNK55_01	65	**71**	Unknown	Male	Unknown	American	N/A	FBI Vault (John Murtha, File 29)
UNK41_01	50	104	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph McCarthy, File 1)
UNK42_01	51	**61**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph McCarthy, File 5)
UNK46_01	55	153	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph P. Kennedy, File 1)
UNK46_02	55	**45**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph P. Kennedy, File 1)

UNK47_01	56	284	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph P. Kennedy, File 2)
UNK48_01	57	542	Unknown	Female*	Unknown	American*	N/A	FBI Vault (Joseph P. Kennedy, File 4)
UNK49_01	58	**92**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Joseph P. Kennedy, File 4)
UNK65_01	75	212	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Lucille Ball, File 1)
UNK66_01	76	233	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Lucille Ball, File 1)
UNK69_01	79	293	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Marlene Dietrich, File 1)
UNK81_01	91	122	Unknown	Male	34	American	1992	FBI Vault (Micheal Jackson, File 9 part 1)
UNK81_02	91	292	Unknown	Male	34	American	1992	FBI Vault (Micheal Jackson, File 9 part 1)
UNK81_03	91	295	Unknown	Male	34	American	1992	FBI Vault (Micheal Jackson, File 9 part 1)
UNK45_01	54	188	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Robert F. Kennedy, File 9)
UNK71_01	81	**58**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Roberto Clemente, File 1)

UNK68_01	78	107	Unknown	Male*	23*	American*	N/A	FBI Vault (Rocky Marciano, File 1)
UNK68_02	78	216	Unknown	Male*	23*	American*	N/A	FBI Vault (Rocky Marciano, File 1)
UNK68_03	78	120	Unknown	Male*	23*	American*	N/A	FBI Vault (Rocky Marciano, File 1)
UNK70_01	80	**62**	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Sammy Davis Jr, File 3)
UNK44_01	53	278	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Sen. George Norris, File 1)
UNK80_01	90	739	Unknown	Unknown	Unknown	American*	1960	FBI Vault (Steve Allen, File 1)
UNK80_02	90	715	Unknown	Unknown	Unknown	American*	1960	FBI Vault (Steve Allen, File 1)
UNK80_03	90	763	Unknown	Unknown	Unknown	American*	1960	FBI Vault (Steve Allen, File 1)
UNK20_01	27	296	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Ted Kennedy, File 11)
UNK19_01	26	309	Unknown	Unknown	Unknown	American*	N/A	FBI Vault (Ted Kennedy, File 2)

UNK74_01	84	**28**	Unknown	Unknown	Unknown	American*	1964	FBI Vault (The Beatles, File 9)
UNK77_01	87	166	Unknown	Unknown	Unknown	American*	1944*	FBI Vault (Walter Winchell, File 19)
UNK78_01	88	173	Unknown	Unknown	Unknown	American*	1944	FBI Vault (Walter Winchell, File 19)
UNK79_01	89	113	Unknown	Unknown	Unknown	American*	1944	FBI Vault (Walter Winchell, File 22)
UNK75_01	85	**89**	Unknown	Unknown	Unknown	American*	1937	FBI Vault (Walter Winchell, File 4)
UNK76_01	86	951	Unknown	Unknown	Unknown	American*	1938	FBI Vault (Walter Winchell, File 5)
ERRU_01	6	279	Eric Rudolph	Male	32	American	2000	Gales, T. A. 2010. Ideologies of Violence: a Corpus and Discourse Analytic Approach to Stance in Threatening Communications
UNK07_01	12	328	Unknown	Male*	40+*	N/A	N/A	Gales, T. A. 2010. Ideologies of Violence: a Corpus and Discourse Analytic Approach to Stance in Threatening Communications
UNK08_01	13	100	Unknown	Unknown	Unknown	N/A	N/A	Gales, T. A. 2010. Ideologies of Violence: a Corpus and Discourse Analytic Approach to Stance in Threatening Communications
LUHE_01	1	414	Luke Helder	Male	21	American	1980	Olsson, J. 2003. Forensic Linguistics: An Introduction to Language, Crime and the Law
MISA_01	2	272	Micheal Sams	Male	51	English	1990	Olsson, J. 2003. Forensic Linguistics: An Introduction to Language, Crime and the Law
UNK09_01	14	381	Unknown	Unknown	Unknown	American*	1996	Olsson, J. 2003. Forensic Linguistics: An Introduction to Language, Crime and the Law
UNK14_01	20	162	Unknown	Unknown	Unknown	N/A	N/A	Olsson, J. 2003. Forensic Linguistics: An Introduction to Language, Crime and the Law

UNK87_01	99	183	Unknown	Unknown	Unknown	American*	Unknown	Pennebaker, J. 2011. The Secret Life of Pronouns
UNK01_01	3	353	Unknown	Female	70	English	2010	Private collection
UNK02_01	4	506	Unknown	Male	48	English	N/A	Private collection
UNK02_02	4	267	Unknown	Male	48	English	N/A	Private collection
JOMC_01	5	524	Anonymised	Male	46	English	N/A	Private collection
UNK03_01	7	148	Unknown	Male	30*	N/A	N/A	Private collection
ROSP_01	8	325	Anonymised	Male	30+*	American	N/A	Private collection
ROSP_02	8	542	Anonymised	Male	30+*	American		Private collection
UNK05_01	10	**93**	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK10_01	15	353	Unknown	Female	35+	N/A	N/A	Private collection
UNK11_01	16	388	Unknown	Male	55+	N/A	N/A	Private collection
UNK12_01	17	603	Unknown	Unknown	Unknown	N/A	N/A	Private collection
JOSA_01	18	125	Anonymous	Male	19	N/A	N/A	Private collection
UNK13_01	19	437	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK15_01	21	286	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK15_02	21	270	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK16_01	22	257	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK17_01	23	495	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK18_01	24	1078	Unknown	Unknown	Unknown	N/A	N/A	Private collection
UNK58_01	68	402	Unknown	Unknown	Unknown	Unknown	N/A	Private collection

UNK59_01	69	870	Unknown	Unknown	Unknown	Unknown	N/A	Private collection
UNK60_01	70	210	Unknown	Male	30	American	N/A	Private collection
UNK61_01	71	624	Unknown	Male	30	American	N/A	Private collection
UNK61_02	71	407	Unknown	Male	30	American	N/A	Private collection
UNK62_01	72	371	Unknown	Male	35	English	N/A	Private collection
UNK89_01	101	286	Unknown	Unknown	Unknown	Canadian	2009	Web search (http://www.vancouverite.com/2009/07/16/home-grown-canadian-terrorists-threaten-b-c-gas-operations/)
UNK88_01	100	104	Unknown	Unknown	Unknown	English	2010	Web search (http://news.bbc.co.uk/1/shared/spl/hi/pop_ups/08/uk_animal_rights_trial/html/3.stm)
UNK90_01	102	123	Unknown	Unknown	Unknown	Unknown	2011	Web search (http://shoqvalue.com/a-threatening-letter-i-received-after-post-about-david-house)

UNK90_02	102	177	Unknown	Unknown	Unknown	Unknown	2011	Web search (http://shoqvalue.com/a-threatening-letter-i-received-after-post-about-david-house)
UNK90_03	102	116	Unknown	Unknown	Unknown	Unknown	2011	Web search (http://shoqvalue.com/a-threatening-letter-i-received-after-post-about-david-house)
UNK92_01	104	154	Unknown	Unknown	Unknown	American*	2009	Web search (http://slog.thestranger.com/slog/archives/2009/01/06/gay_bars_receive_threatening)
UNK91_01	103	351	Unknown	Unknown	Unknown	American*	2007	Web search (http://www.anthraxinvestigation.com/GS-thoughts.html)
UNK94_01	106	123	Unknown	Unknown	Unknown	Unknown	2013	Web search (http://www.bobsblitz.com/2013/04/heres-threatening-letter-rc-sent-to.html)
UNK88_02	100	139	Unknown	Unknown	Unknown	English	2010	Web search (http://www.dailymail.co.uk/news/article-1323573/Animal-rights-activists-targeted-Huntingdon-Life-Sciences-jailed.html)
UNK95_01	107	104	Price	Male	Middle-age	American	1996	Web search (http://www.examiner.com/article/terrorism-and-domestic-violence-america)

UNK04_01	9	102	Unknown	Unknown	Unknown	N/A	N/A	Web search (http://www.krqe.com/dpp/news/crime/who-wrote-threatening-letter)
UNK93_01	105	250	Unknown	Unknown	Unknown	Unknown	2005	Web search (http://www.rawa.org/debellis.htm)
UNK96_01	108	165	Aslam Butt	Male*	Unknown	Unknown	2013	Web search (http://www.scoop.co.nz/stories/WO1304/S00103/pakistan-open-letter-on-a-threatening-email-from-journalist.htm)
ZOKI_01	98	406	Zodiac killer	Unknown	Unknown	Unknown	1969	Web search (http://www.zodiackiller.com/letters_index.html)
ZOKI_02	98	601	Zodiac killer	Unknown	Unknown	Unknown	1969	Web search (http://www.zodiackiller.com/letters_index.html)
ZOKI_03	98	205	Zodiac killer	Unknown	Unknown	Unknown	1969	Web search (http://www.zodiackiller.com/letters_index.html)

UNK06_01	11	461	Unknown	Female*	26*	American	2011	Web search (https://sites.google.com/a/truthaboutbills.com/www/News-Feed/threateninglettertowirepsenators)

9.2 The Experiment tasks

CODE:

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

9.3 The questionnaire

Code:

Please fill in this field by using the first two letters of your name followed by the first two letters of your surname. Example: Mr John Smith -> Code: JOSM

Gender

- Male
- Female

Age

What is the highest level of education that you have achieved so far?

- Primary school
- Secondary school
- College
- Bachelor degree
- Postgraduate degree
- PhD
- Other:

What is your occupation?

What is or was your mother's occupation?

What is or was your father's occupation?

How would you define your ethnic background?

Where are you from?

Tick the box below if you are happy NOT to receive any compensation for your participation. This research project is supported by limited funding and your voluntary help would give us the possibility to recruit more participants and therefore increase the potential and usefulness of our findings

Tick

Submit

9.4 The consent form

Consent form

Title of project: Authorship profiling for forensic purposes

Name of researcher: Andrea Nini

- I agree to participate in this study
- I understand that the information that I will provide for the questionnaire will be treated confidentially and that will not be disclosed to anyone else who is not involved in the research project
- I understand and it has been explained to me that this study aims at studying the written language of several strata of society and that therefore, for this purpose, the information I provide in the questionnaire will be used to create a sample stratified according to gender, age, education level, ethnicity and social positions
- I understand that my participation is voluntary and that I can withdraw at any time

---------- ----------

Name and Surname Signature

---------- ----------

Signature of researcher Date

9.5 Variables surveyed from the literature

Table 9-1 - Table summarising all the variables used in the studies surveyed from the literature

Variable	Process of analysis	Reference	High score predicts:	Considered?
Abstractness of nouns	Semi-automatic	Berman (2008)	Higher education	No
Adverbial booster SW	Semi-automatic	McEnery (2006)	Male gender	No
Classifiers	Manual	Hawkins (1977)	Lower social class [level of education, occupation of parents]	No
Compactness	Manual	Byrd (1993)	Younger age	No
Degree of cohesiveness	Manual	Byrd (1993)	Younger age	No
D-Level	Automatic	Kemper & Sumner (2001), Kemper *et al.* (2001)	Younger age	No
Embedded clauses	Manual	Rabaglia & Salthouse (2011)	Younger age	No
Emphatic adverb/adj SW	Semi-automatic	McEnery (2006)	Male gender	No
Emphatic *so*	Semi-automatic	Crosby & Nysquist (1977)	Female gender	No
Empty adjectives	Semi-automatic	Crosby & Nysquist (1977)	Female gender	No

Evaluative adjectives	Semi-automatic	Barbieri (2008), Macaulay (2002)	Younger age, Higher social class [occupation, education, residence]	No
Exophoric references	Manual	Hawkins (1977)	Lower social class [level of education, occupation of parents]	No
Fragments	Manual	Mitzner & Kemper (2003)	Lower education	No
Future tense verbs	Semi-automatic	Pennebaker *et al.* (2003)	Older age	No
General expletives SW	Semi-automatic	McEnery (2006)	Female gender	No
Germanic words	Semi-automatic	Berman (2008)	Higher education	No
Hedges	Semi-automatic	Crosby & Nysquist (1977)	Fermale gender	No
Hypocoristic adjectives	Semi-automatic	Hawkins (1977)	Female gender, Lower social class [level of education, occupation of parents]	No
Idiomatic SW	Semi-automatic	McEnery (2006)	Female gender, Older age	No
Left-branching clauses	Manual	Kemper *et al.* (1989), Rabaglia & Salthouse (2011), Labov & Auger (1993)	Younger age, Higher social class [occupation]	No
Lexical density of clauses	Manual	Berman (2008)	Higher education	No
Lexical item *I think*	Automatic	Poole (1979)	Female gender	No

Literal SW	Semi-automatic	McEnery (2006)	Higher social class [occupation]	No
Loban weighted index of subordination	Manual	Poole (1976)	Higher social class [father's occupation, education]	No
Main clauses	Manual	Kemper *et al.* (1989)	Lower education	No
Mean noun phrase length	Manual	Berman (2008)	Higher education	No
Mean number of nodes in noun phrase	Manual	Berman (2008)	Higher education	No
Mean pre-verb length	Manual	Poole (1979)	Higher social class [school area]	No
Mild SW	Automatic	McEnery (2006)	Female gender, Older age, Higher social class [occupation]	No
Moderate SW	Semi-automatic	McEnery (2006)	Female gender, Older age, Higher social class [occupation]	No
New referents type	Manual	Cheshire (2005)	-	No
Non-finite subordination	Manual	Berman (2008)	Higher education	No
Noun phrases	Manual	-	-	No
Personal SW	Semi-automatic	McEnery (2006)	Younger age	No
Politeness forms	Manual/Semi-automatic	Crosby & Nysquist (1977)	Female gender	No
Powerless register	Semi-automatic	Crosby & Nysquist (1977)	Fermale gender	No

Predicative negative adjectives SW	Semi-automatic	McEnery (2006)	Higher social class [occupation]	No
Premodifying intensifying negative adjectives SW	Semi-automatic	McEnery (2006)	Female gender	No
Pronominal SW	Semi-automatic	McEnery (2006)	Higher social class [occupation]	No
Proper nouns (people)	Semi-automatic	Rayson *et al.* (1997)	Female gender	No
Proper nouns (places)	Semi-automatic	Rayson *et al.* (1997)	Male gender	No
Proportion of (*of*)/(*in + into*)	Automatic	Poole (1979)	Male gender	No
Qualifiers	Manual	Hawkins (1977)	Lower social class [level of education, occupation of parents]	No
Ratio (*I*)/ (total personal pronouns)	Automatic	Poole (1979), Poole (1976)	Female gender, Higher social class [school area]	No
Ratio (uncommon adjectives)/(adjectives)	Automatic	Poole (1979), Poole (1976)	Female gender, Higher social class [school area]	No
Ratio (unusual adverbs)/(adverbs)	Automatic	Poole (1979)	Male gender	No
Relative clauses	Manual	Berman (2008)	Higher education	No
Right-branching clauses	Manual	Kemper *et al.* (1989)	Higher education	No
Romance words	Semi-automatic	Berman (2008)	Higher education	No
Slang	Semi-automatic	Barbieri (2008)	Younger age	No
Strong SW	Semi-automatic	McEnery (2006)	Male gender, Younger age, Lower social class [occupation]	No

Subjects filled by nouns	Manual	Johnston (1977)	Higher social class [father's occupation]	No
Subjects filled by pronouns	Manual	Johnston (1977)	Lower social class [father's occupation]	No
Subordinate clauses	Manual	Rabaglia & Salthouse (2011), Johnston (1977), Poole (1979)	Younger age, Higher social class [father's occupation], Higher social class [school area]	No
Tag questions	Manual	Crosby & Nysquist (1977)	Female gender	No
Uncommon adjectives	Semi-automatic	Poole (1979), Hawkins (1977)	Male gender, Higher social class [parent's occupation, education]	No
Unusual adverbs	Semi-automatic	Poole (1979), Poole (1976)	Male gender, Higher social class [school area]	No
Very Mild SW	Semi-automatic	McEnery (2006)	Female gender, Older age, Higher social class [occupation]	No
Very Strong SW	Semi-automatic	McEnery (2006)	Male gender, Younger age, Lower social class [occupation]	No
Word *I*	Automatic	Poole (1979)	Female gender	No
Word *in*	Automatic	Poole (1979)	Male gender	No
Word *into*	Automatic	Poole (1979)	Male gender	No
Word *its*	Automatic	Argamon *et al.* (2003)	Male gender	No
Word *of*	Automatic	Argamon *et al.* (2003), Koppel *et al.* (2002)	Male gender	No
Words longer than five letters	Automatic	Rabaglia & Salthouse (2011)	Older age	No

Adjectives	Automatic	Heylighen & Dewaele (1999), Poole (1979), Macaulay (2002)	Male gender, Higher education, Higher social class [occupation, education, residence]	Yes
Advanced Guiraud 1000	Automatic	Mollet *et al.* (2010), Byrd (1993)	Higher education	Yes
Adverbs	Automatic	Heylighen & Dewaele (1999), Poole (1979), Macaulay (2002), Poole (1976), Rayson *et al.* (1997)	Female gender, Lower education, Higher social class [occupation, education, residence], Higher social class [father's occupation, level of education], Higher social class [occupation]	Yes
Articles	Automatic	Heylighen & Dewaele (1999), Newman *et al.* (2008)	Male gender, Higher education	Yes
Attributive adjectives	Automatic	Argamon *et al.* (2003)	Male gender	Yes
Average clauses per sentence	Automatic	Kemper *et al.* (1989)	Younger age	Yes
Average word length	Automatic	Argamon *et al.* (2003), Bromley (1991)	Male gender, Higher education	Yes
Baayen's P	Automatic	Mollet *et al.* (2010)	Higher education	Yes
Clauses	Automatic	-	-	Yes
Contractions	Automatic	Argamon *et al.* (2003)	Female gender	Yes
Deep formality	Automatic	Heylighen & Dewaele (1999)	Male gender, Higher education	Yes
Dependent clauses	Automatic	-	-	Yes

Determiners	Automatic	Argamon *et al.* (2003), Koppel *et al.* (2002)	Male gender	Yes
Determiners/Nouns	Automatic	Argamon *et al.* (2003)	Male gender	Yes
Dimension 1	Automatic	Biber & Reppen (1998), Schler *et al.* (2006)	Female gender, Younger age	Yes
Emphatics	Automatic	Biber & Reppen (1998)	Female gender	Yes
Fichtner's C	Automatic	Kemper *et al*. (2001)	Younger age	Yes
First person pronouns	Automatic	Argamon *et al.* (2003), Pennebaker & Stone (2003)	Female gender, Younger age	Yes
Flesch readability score	Automatic	Bromley (1991), Flesch (1949)	Higher education	Yes
Genitives	Automatic	Hawkins (1977)	Female gender, Lower social class [level of education, occupation of parents]	Yes
Innovative stance adverbs	Automatic	Barbieri (2008)	Younger age	Yes
Intensifiers	Automatic	Barbieri (2008)	Younger age	Yes
Interjections	Automatic	Heylighen & Dewaele (1999)	Female gender	Yes
Lemma *say*	Automatic	Rayson *et al.* (1997)	Lower social class [occupation]	Yes
Lexical density of text	Automatic	Berman (2008)	Higher education	Yes
Long T-units	Manual	Hunt (1983)	Higher education	Yes
Mean clause length	Automatic	Berman (2008), Hunt (1971), Hunt (1983)	Higher education	Yes
Mean sentence length	Automatic	Kemper *et al.* (2001), Hunt (1983), Mitzner & Kemper (2003), Poole (1979)	Younger age, Higher education, Higher social class [school area]	Yes

Mean T-unit length	Automatic	Hunt (1971), Hunt (1983), Loban (1967)	Higher education, Higher social class [parent's occupation]	Yes
Modal verbs	Semi-automatic	Barbieri (2008), Plum & Cowling (1987)	Older age, Higher social class [not specified]	Yes
Negative emotion words	Automatic	Pennebaker & Stone (2003)	Younger age	Yes
Negative marker	Automatic	Argamon *et al.* (2003)	Female gender	Yes
Nouns	Automatic	Heylighen & Dewaele (1999), Hawkins (1977)	Male gender, Higher education, Higher social class [level of education, occupation of parents]	Yes
Numbers	Automatic	Rayson *et al.* (1997)	Male gender	Yes
Passive voice	Automatic	Berman (2008)	Higher education	Yes
Past tense verbs	Automatic	Pennebaker & Stone (2003), Plum & Cowling (1987)	Younger age, Higher social class	Yes
P-Density	Automatic	Kemper *et al*. (2001), Mollet *et al.* (2010)	Younger age, Higher education	Yes
Personal pronouns	Automatic	Koppel *et al.* (2002), Newman *et al.* (2008), Poole (1979), Rayson *et al.* (1997)	Female gender, Lower social class [occupation]	Yes
Positive emotion words	Automatic	Pennebaker & Stone (2003)	Older age	Yes
Possessive determiners	Automatic	Hawkins (1977)	Female gender, Lower social class [level of education, occupation of parents]	Yes
Prepositions	Automatic	Argamon *et al.* (2003), Heylighen & Dewaele (1999),	Male gender, Higher education	Yes

		Koppel *et al.* (2002), Poole (1979)		
Present tense verbs	Automatic	Argamon *et al.* (2003), Plum & Cowling (1987)	Female gender, Lower social class [not specified]	Yes
Pronouns	Automatic	Heylighen & Dewaele (1999), Johnston (1977), Hawkins77	Female gender, Lower education, Lower social class [parent's occupation]	Yes
Proper nouns	Automatic	Rayson97	Female gender	Yes
Ratio (Clause)/(T-unit)	Automatic	Hunt83, Mitzner & Kemper (2003), Loban (1967)	Higher education, Higher social class [parent's occupation]	Yes
Ratio (Dependent clauses)/(T-unit)	Automatic	Labov & Auger (1993), Loban (1967)	Higher social class [occupation], Higher education	Yes
Ratio (nouns)/(personal pronouns)	Automatic	Saily *et al.* (2011)	Male gender	Yes
Ratio (syllables)/(words)	Automatic	Rabaglia & Salthouse (2011), Berman (2008)	Older age, Higher education	Yes
Ratio (T-unit)/(Sentence)	Automatic	Hunt (1983)	Lower education	Yes
Second person pronouns	Automatic	Argamon *et al.* (2003), Rayson *et al.* (1997)	Female gender, Lower social class [occupation]	Yes
Sentences	Automatic	-	-	Yes
Short T-units	Manual	Hunt (1983)	Lower education	Yes
Social words	Automatic	Newman *et al.* (2008)	Female gender	Yes
Subordinating connectives	Automatic	Loban (1967)	Higher education, Higher social class [parent's occupation]	Yes

Text length	Automatic	Bromley (1991), Johnston (1977)	Higher education, Higher social class [parent's occupation]	Yes
Third person pronouns	Automatic	Argamon *et al.* (2003), Newman *et al.* (2008), Barbieri (2008), Rayson *et al.* (1997)	Female gender, Low social class [occupation], Older age	Yes
Time words	Automatic	Pennebaker & Stone (2003)	Younger age	Yes
Total SW	Automatic	McEnery (2006), Newman *et al.* (2008), Barbieri (2008), Rayson *et al.* (1997)	Younger age, Lower social class [occupation]	Yes
Traditional stance adverbs	Automatic	Barbieri (2008)	Older age	Yes
T-units	Manual	-	-	Yes
Type/token ratio	Automatic	Kemper *et al.* (2001), Rabaglia & Salthouse (2011), Byrd (1993)	Older age, Higher education	Yes
Verbs	Automatic	Heylighen & Dewaele (1999), Johnston (1977)	Female gender, Lower education, Lower social class [parent's occupation]	Yes
Word *and*	Automatic	Koppel *et al.* (2002)	Female gender	Yes
Word *for*	Automatic	Koppel *et al.* (2002)	Female gender	Yes
Word *with*	Automatic	Koppel *et al.* (2002)	Female gender	Yes
Words longer than six letters	Automatic	Newman *et al.* (2008), Pennebaker & Stone (2003)	Male gender, Older age	Yes
Words longer than ten letters	Automatic	Bromley (1991)	Higher education	Yes

9.6 The list of variables used in the analyses

This Appendix lists the variables used in the present study for the analysis presented in Chapter 4, Chapter 5 and Chapter 6 in alphabetical order. If an abbreviation was used in the graphs for the variable this is listed next to the name of the variable in parentheses. For each variable a description of how it was calculated is given. In the descriptions, a word that is entirely capitalised indicates a lemma.

Advanced Guiraud 1000

This variable was calculated using the following formula as described in Mollet *et al*. (2010):

$$AG = \frac{(types_{total} - types_{common})}{\sqrt{tokens}}$$

The number of common types consisted in the number of types in the text that could be found in the first 1000 types of the British National Corpus word list. The script was checked on 20% of the AMT corpus.

Agentless passives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a passive is assigned when one of the two following patterns is found: (a) any form of BE followed by a participle plus one or two optional intervening adverbs or negations; (b) any form of BE followed by a nominal form and a participle.

Amplifiers

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where amplifiers were defined as any of the items in this list: *absolutely, altogether, completely, enormously, entirely, extremely, fully, greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very*.

Analytic negations

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an analytic negation is assigned when *not* or its contraction are found.

Attributive adjective

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an attributive adjective is assigned for any adjective that was not tagged as predicative adjective.

Average clause length

This variable was calculated by dividing the total number of tokens by the number of clauses.

Average sentence length

This variable was calculated by dividing the total number of tokens by the number of sentences.

Average t-unit length

This variable was calculated by dividing the total number of tokens by the number of t-units.

Average word length

Average word length was calculated automatically by MAT dividing the total number of characters of a text by the total number of tokens.

Average word length in syllables

This variable was calculated by dividing the total number of syllables by the number of tokens.

Baayen's P

This variable was calculated as described in Mollet *et al*. (2010) by dividing the number of *hapax legomena* by the number of tokens. This operation was done with a Perl script.

***Be* as main verb**

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *be* as main verb is assigned when BE is not preceded by *there* and it is followed by a determiner, a cardinal number, a personal pronoun or a possessive pronoun or a preposition or an adjective, taking into account intervening adverbs or negations.

By-passives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a by-passive is assigned when the conditions for passive is met and the preposition *by* follows the pattern.

Cardinal numbers

This variable was calculated automatically by MAT and it is one of the tags assigned by the Stanford Tagger. This variable includes any cardinal number

Causative adverbial subordinators

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where causative adverbial subordinators were defined as any occurrence of the word *because*.

Clauses

Clauses were identified automatically using a Perl script that identified all the lexical verbs and all the forms of BE, HAVE and DO that were not auxiliaries.

Clauses per t-units

This variable was calculated by dividing the total number of clauses by the number of t-units.

Common nouns

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Concessive adverbial subordinators

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where causative adverbial subordinators were defined as any occurrence of the words *although* and *though*.

Conditional adverbial subordinators

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where causative adverbial subordinators were defined as any occurrence of the words *if* and *unless*.

Conjuncts

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a conjunct is assigned when any of the items in this list is found: punctuation+*else*, punctuation+*altogether*, punctuation+*rather*, *alternatively, consequently, conversely, e.g., furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, similarly, therefore, thus, viz., in comparison, in contrast, in particular, in addition, in conclusion, in consequence, in sum, in summary, for example, for instance, instead of, by contrast, by comparison, in any event, in any case, in other words, as a result, as a consequence, on the contrary, on the other hand.*

Contractions

This variable was calculated automatically by MAT. The program identified a contraction every time an apostrophe was found or when the item *n't* was found.

Coordinating conjunctions

This variable was calculated automatically by MAT and it is one of the tags assigned by the Stanford Tagger. This variable includes any occurrence of the words *but* and *or* as well as any occurrence of the word *and* that was not tagged as independent clause coordination or phrasal coordination.

Deep formality

This variable was calculated following the procedure described in Heylighen & Dewaele (1999) using MAT variables. The formula used for the present study therefore was:

DF = (nominalisations [NOMZ] + gerunds [GER] + total nouns [NN] + attributive adjectives [JJ] + predicative adjectives [PRED] + prepositions [PIN] + articles [DT] – first person pronouns [FPP1] – second person pronouns [SPP2] – third person pronouns [TPP3] – pronoun *it* [PIT] – verb bases [VB] – past tenses [VBD] – present participles [VBG] – past participles [VBN] – present tenses [VPRT] – place adverbials [PLACE] – time adverbials [TIME] – total adverbs [RB] – conjuncts [CONJ] – downtoners [DWNT] – amplifiers [AMP] – interjections [UH] + 100)/2

The notation "DF" in the present study was preferred to the simple "F" used by Heylighen & Dewaele (1999) as it is less prone to ambiguity. In general, the clearer term "deep formality" rather than just "formality" is used in the present work in order to avoid confusion with more traditional concepts of linguistic formality.

Deep formality 2

Since Heylighen & Dewaele (1999) specify that deep formality can be calculated just using deep formal items against non-deep formal items, a new deep formality calculation was generated using MAT variables that fall within those two categories. This new variable, called DF2, was calculated using the following formula (the new variables are underlined):

DF2 = (nominalisations [NOMZ] + gerunds [GER] + total nouns [NN] + attributive adjectives [JJ] + predicative adjectives [PRED] + prepositions [PIN] + articles [DT] + *that* as adjective complement [THAC] + past participial WHIZ deletion relatives [WZPAST] + present participial WHIZ deletion relatives [WZPRES] + *that* relative clauses on subject position [TSUB] + *that* relative clauses on object position [TOBJ] + WH relative clauses on subject position [WHSUB] + WH relative clauses on object position [WHOBJ] – first person pronouns [FPP1] – second person pronouns [SPP2] – third person pronouns [TPP3] – pronoun *it* [PIT] – demonstrative pronouns [DEMP] – indefinite pronouns [INPR] – pro-verb *do* [PROD] – WH-clauses [WHCL] – verb bases [VB] – past tenses [VBD] – present participles [VBG] – past participles [VBN] – present tenses [VPRT] – place adverbials [PLACE] – time

adverbials [TIME] – total adverbs [RB] – conjuncts [CONJ] – downtoners [DWNT] – amplifiers [AMP] – interjections [UH] + 100)/2

These variables were chosen as they met the requirements of being respectively formal or non-formal. All the variables added on the formal side are different ways to expand the scope and specificity of nouns whereas all the elements added in the non-formal side are other forms of pronouns.

Demonstrative pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a demonstrative pronoun is assigned when the words *those, this, these* are followed by a verb or auxiliary verb or a punctuation mark or a WH pronoun or the word *and*. The word *that* is also tagged as a demonstrative pronoun when it follows the pattern above or when it is followed by *'s* or *is* and, at the same time, it has not been already tagged as a *that* relative clauses in object position, *that* relative clauses in subject position, *that* adjective complements, or *that* as a verb complement.

Demonstratives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a demonstrative is assigned when the words *that, this, these, those* are found and when these have not been tagged as either demonstrative pronouns, *that* relative clauses in object position, *that* relative clauses in subject position, *that* adjective complements, or *that* as a verb complement.

Dependent clauses

The number of dependent clauses was obtained by subtracting the number of t-units from the number of clauses, since any t-unit contains by definition only one main clause.

Dependent clauses per sentence

This variable was calculated by dividing the total number of dependent clauses by the number of sentences.

Determiners

This variable was calculated automatically by MAT and it is one of the tags assigned by the Stanford Tagger. This variable includes any occurrence of the words *a, an, the*.

Determiners per nouns

This variable was calculated by dividing the number of determiners (DT) by the numbers of common nouns (NN).

Dimension 1

Dimension 1 is the opposition between Involved and Informational discourse. Low scores on this variable indicate that the text is informationally dense, as for example academic prose, whereas high scores indicate that the text is affective and interactional, as for example a casual conversation. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Dimension 2

Dimension 2 is the opposition between Narrative and Non-Narrative Concerns. Low scores on this variable indicate that the text is non-narrative whereas high scores indicate that the text is narrative, as for example a novel. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Dimension 3

Dimension 3 is the opposition between Context-Independent Discourse and Context-Dependent Discourse. Low scores on this variable indicate that the text is dependent on the context, as in the case of a sport broadcast, whereas a high score indicate that the text is not dependent on the context, as for example academic prose. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Dimension 4

Dimension 4 measures Overt Expression of Persuasion. High scores on this variable indicate that the text explicitly marks the author's point of view as well as their assessment of likelihood and/or certainty, as for example in professional letters. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Dimension 5

Dimension 5 is the opposition between Abstract and Non-Abstract Information. High scores on this variable indicate that the text provides information in a technical, abstract and formal way, as for example in scientific discourse. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Dimension 6

Dimension 6 measures On-line Informational Elaboration. High scores on this variable indicate that the text is informational in nature but produced under certain time constraints, as for example in speeches. The variable is automatically calculated by MAT using the instructions provided by Biber (1988).

Discourse particles

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a discourse particle is assigned when any of the words *well, now, anyhow, anyways* is preceded by a punctuation mark.

Downtoners

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a downtoner is assigned when any of these items is found: *almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat*.

Emphatics

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an emphatic is assigned when any of these words or patterns is found: *just, really, most, more, real*+adjective, *so*+adjective, any form of DO followed by a verb, *for sure, a lot, such a*.

Evaluative adjectives

This variable was calculated with a Perl script that searched for the following words: *serious, crazy, stupid, weird, awesome, funny, biggest, pissed, horrible, shitty, strange, hard, tight, nervous, popular, personal, entire, hot, cold, scary, honest, natural, mad, short, good, massive, sick, disgusting, brilliant.* The count was then normalised.

Existential *there*

This variable was calculated automatically by MAT and it is one of the tags assigned by the Stanford Tagger. This variable includes any occurrence of the word *there* that the Stanford Tagger analysed as being a case of existential *there*.

Fichtner's C

This variable measures the level of syntactic complexity by expressing a measure of how many levels of embedded sentences there are a in a text. This variable was found by Mollet *et al*. (2010) to be a good proxy for D-Level and for syntactic complexity in general. The formula used in Mollet *et al*. (2010) to calculate Fichtner's C was the following:

$$FC = \frac{verbs}{sentences} * \frac{tokens}{sentences}$$

In the present study, however, the number of clauses was used instead of the number of lexical verbs, since the number of clauses found in the present study is already a measure of lexical verbs.

First person pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a first person pronoun is assigned when any of these words is found: *I, me, us, my, we, our, myself, ourselves*.

Flesch-Kincaid grade level

Flesch-Kincaid grade level (Flesch, 1949) was determined using the formula:

$$Flesch - Kincaid\ grade\ level = 0.39(ASL) + 11.8(ASW) - 15.59$$

Flesch readability score

Flesch readability score (Flesch, 1949) was determined using the formula:

$$Flesch\ readability\ score = 206.835 - 1.015(ASL) - 84.6(ASW)$$

General adverbs

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Genitives

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Gerunds

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a gerund is assigned when any nominal form ending in *–ing* or *–ings* is found.

Hedges

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a hedge is assigned when any word or pattern in this list is found: *maybe, at about, something like, more or less, sort of, kind of* (these two items must be preceded by a determiner, a quantifier, a cardinal number, an adjective, a possessive pronouns or WH word).

Indefinite pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an indefinite pronoun is assigned when any word in this list is found: *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something*.

Independent clause coordination

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an independent clause coordination is assigned when the word *and* is found in one of the following patterns: (1) preceded by a comma and followed by *it, so, then, you, there* + BE, or a demonstrative pronoun or the subject forms of a personal pronouns; (2) preceded by any punctuation; (3) followed by a WH pronoun or any WH word, an adverbial subordinator or a discourse particle or a conjunct.

Infinitives

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger. MAT assigns the category 'infinitives' only to those *to* that are not tagged as prepositions.

Innovative stance adverbs

This variable was calculated with a Perl script that searched for the following words: *kind of, sort of, really, actually, definitely, totally.* The count was then normalised.

Intensifiers

This variable was calculated by summing the relative frequencies of the two MAT variables amplifiers and emphatics. Although slightly differently from what proposed by the hypothesis Barbieri08, this solution was chosen as more comprehensive.

Lemma SAY

This variable was calculated with a Perl script that searched for occurrences of: *say, says, saying, said.* The count was then normalised per the total tokens.

Lexical density

Lexical density was calculated using a script that individuated content words. Following Biber *et al.* (1999), a content word was defined as a noun, a lexical verb, an adjective or an adverb. However, only adverbs ending in *–ly* were considered as being content words, given that most of the adverbs non ending in *–ly* are more likely to be considered function words. The script therefore counted as content words: all singular and plural common and proper nouns, all nominalisations, all gerunds, all attributive adjectives, all predicative adjectives, all verbs except for the lemmas HAVE, DO and BE, and all the adverbs ending in *–ly*. The count for content words was then divided by the number of tokens to obtain the classic measure of lexical density.

Lexical density H

Following Halliday (2004), another measure of lexical density was calculated. Halliday (2004: 344) defined lexical density as 'the number of lexical items (content words) per ranking (non-embedded) clause'. Although the number of ranking clauses is impossible to calculate automatically and would require a careful and time-consuming manual analysis, a proxy to it is the number of t-units. Lexical density H is therefore the number of content words divided by the number of t-units.

Long t-units

As detailed in Hunt (1983) a long t-unit is defined as a t-unit that contains more than 20 tokens. A script was created that counted how many t-units longer than 20 tokens were found in a text and this count was then normalised by the total number of t-units.

Mean rarity score

The mean rarity score of a text was calculated using the word list of the British National Corpus (BNC) retrieved from:

<http://www.lexically.net/downloads/version4/downloading%20BNC.htm>

A script was created to identify for each word the rank that it had on the BNC word list. For each text, the sum of the ranks was obtained and the mean was calculated. This score was then multiplied by 0.01 in order to have a more manageable 2-digit number.

Necessity modals

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a necessity modal is assigned when any word in this list is found: *ought, should, must*.

Negative emotion words

This variable belongs to the set of variables produced by LIWC (Tausczik and Pennebaker, 2009). The list of stems that LIWC includes in this category is: *abandon*, abuse*, abusi*, ache*, aching, advers*, afraid, aggravat*, aggress*, agitat*, agoniz*, agony, alarm*, alone, anger*, angr*, anguish*, annoy*, antagoni*, anxi*, apath*, appall*, apprehens*, argh*, argu*, arrogan*, asham*, assault*, asshole*, attack*, aversi*, avoid*, awful, awkward*, bad, bashful*, bastard*, battl*, beaten, bitch*, bitter*, blam*, bore*, boring, bother*, broke, brutal*, burden*, careless*, cheat*, complain*, confront*, confus*, contempt*, contradic*, crap, crappy, craz*, cried, cries, critical, critici*, crude*, cruel*, crushed, cry, crying, cunt*, cut, cynic, damag*, damn*, danger*, daze*, decay*, defeat*, defect*, defenc*, defens*, degrad*, depress*, depriv*, despair*, desperat*, despis*, destroy*, destruct*, devastat*, devil*, difficult*, disadvantage*, disagree*, disappoint*, disaster*, discomfort*,*

discourag, disgust*, dishearten*, disillusion*, dislike, disliked, dislikes, disliking, dismay*, dissatisf*, distract*, distraught, distress*, distrust*, disturb*, domina*, doom*, dork*, doubt*, dread*, dull*, dumb*, dump*, dwell*, egotis*, embarrass*, emotional, empt*, enemie*, enemy*, enrag*, envie*, envious, envy*, evil*, excruciat*, exhaust*, fail*, fake, fatal*, fatigu*, fault*, fear, feared, fearful*, fearing, fears, feroc*, feud*, fiery, fight*, fired, flunk*, foe*, fool*, forbid*, fought, frantic*, freak*, fright*, frustrat*, fuck, fucked*, fucker*, fuckin*, fucks, fume*, fuming, furious*, fury, geek*, gloom*, goddam*, gossip*, grave*, greed*, grief, griev*, grim*, gross*, grouch*, grr*, guilt*, harass*, harm, harmed, harmful*, harming, harms, hate, hated, hateful*, hater*, hates, hating, hatred, heartbreak*, heartbroke*, heartless*, hell, hellish, helpless*, hesita*, hit, homesick*, hopeless*, horr*, hostil*, humiliat*, hurt*, idiot*, ignor*, immoral*, impatien*, impersonal, impolite*, inadequa*, indecis*, ineffect*, inferior*, inhib*, insecur*, insincer*, insult*, interrup*, intimidat*, irrational*, irrita*, isolat*, jaded, jealous*, jerk, jerked, jerks, kill*, lame*, lazie*, lazy, liabilit*, liar*, lied, lies, lone*, longing*, lose, loser*, loses, losing, loss*, lost, lous*, low*, luckless*, ludicrous*, lying, mad, maddening, madder, maddest, maniac*, masochis*, melanchol*, mess, messy, miser*, miss, missed, misses, missing, mistak*, mock, mocked, mocker*, mocking, mocks, molest*, mooch*, moodi*, moody, moron*, mourn*, murder*, nag*, nast*, needy, neglect*, nerd*, nervous*, neurotic*, numb*, obnoxious*, obsess*, offence*, offend*, offens*, outrag*, overwhelm*, pain, pained, painf*, paining, pains, panic*, paranoi*, pathetic*, peculiar*, perver*, pessimis*, petrif*, pettie*, petty*, phobi*, piss*, piti*, pity*, poison*, prejudic*, pressur*, prick*, problem*, protest, protested, protesting, puk*, punish*, rage*, raging, rancid*, rape*, raping, rapist*, rebel*, reek*, regret*, reject*, reluctan*, remorse*, repress*, resent*, resign*, restless*, revenge*, ridicul*, rigid*, risk*, rotten, rude*, ruin*, sad, sadde*, sadly, sadness, sarcas*, savage*, scare*, scaring, scary, sceptic*, scream*, screw*, selfish*, serious, seriously, seriousness, severe*, shake*, shaki*, shaky, shame*, shit*, shock*, shook, shy*, sicken*, sin, sinister, sins, skeptic*, slut*, smother*, smug*, snob*, sob, sobbed, sobbing, sobs, solemn*, sorrow*, sorry, spite*, stammer*, stank, startl*, steal*, stench*, stink*, strain*, strange, stress*, struggl*, stubborn*, stunk, stunned, stuns, stupid*, stutter*, submissive*, suck, sucked, sucker*, sucks, sucky, suffer, suffered, sufferer*, suffering, suffers, suspicio*, tantrum*, tears, teas*, temper, tempers, tense*, tensing, tension*, terribl*, terrified, terrifies, terrify, terrifying, terror*, thief, thieve*, threat*, ticked, timid*, tortur*, tough*, traged*, tragic*, trauma*, trembl*, trick*, trite, trivi*, troubl*, turmoil, ugh, ugl*, unattractive, uncertain*, uncomfortabl*, uncontrol*, uneas*, unfortunate*, unfriendly, ungrateful*, unhapp*, unimportant, unimpress*, unkind, unlov*, unpleasant, unprotected, unsavo*, unsuccessful*, unsure*, unwelcom*, upset*, uptight*, useless*, vain, vanity, vicious*, victim*, vile, villain*, violat*, violent*, vulnerab*, vulture*, war, warfare*, warred, warring, wars, weak*, weapon*, weep*, weird*, wept, whine*, whining, whore*, wicked*, wimp*, witch, woe*, worr*, worse*, worst, worthless*, wrong*, yearn*.*

Nominalisations

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a nominalisation is assigned when a noun ends in *-tion, -ment, -ness* or *-ity*.

***Of* preceded by a noun**

This variable was calculated using a Perl script that identified any occurrence of the word *of* preceded by a nominal form.

Other adverbial subordinators

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *other adverbial subordinator* is assigned when any of these words is found: *since, while, whilst, whereupon, whereas, whereby, such that, so that* (plus a word that is neither a noun nor an adjective), *such that* (plus a word that is neither a noun nor an adjective), *inasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as.*

Past participial clauses

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a past participial clause is assigned when a punctuation mark is followed by a past participial form of a verb followed by a preposition or an adverb.

Past participial WHIZ deletion relatives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a past participial WHIZ deletion relative is assigned when the following pattern is found: a noun or quantifier pronoun followed by a past participial form of a verb followed by a preposition or an adverb or a form of BE.

Past participles

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Past tenses

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

P-Density

P-Density was described in Section 2.2.1. Following Kemper's and Sumner's (2001) and other studies, this variable was calculated using the software CPIDR (Covington, 2012).

Perfect aspects

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a perfect aspect is assigned when HAVE is followed by a past or participle form of any verb taking into account any intervening adverb or negation. The interrogative version of this pattern is found by counting how many times a form of HAVE is followed by a nominal form and then followed by a past or participle form of any verb.

Phrasal coordinations

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a phrasal coordination is assigned when *and* is preceded and followed by the same tag.

Pied-piping relatives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a pied-piping relative is assigned when any preposition is followed by *who, who, whose* or *which*.

Place adverbials

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a place adverbial is assigned when one of the following words is found: *aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west*.

Plural proper nouns

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Positive emotion words

This variable belongs to the set of variables produced by LIWC (Tausczik and Pennebaker, 2009). The list of stems that LIWC includes in this category is: *accept, accepta*, accepted, accepting, accepts, active*, admir*, ador*, advantag*, adventur*, affection*, agree, agreeab*, agreed, agreeing, agreement*, agrees, alright*, amaz*, amor*, amus*, aok, appreciat*, assur*, attachment*, attract*, award*, awesome, beaut*, beloved, benefic*, benefit, benefits, benefitt*, benevolen*, benign*, best, better, bless*, bold*, bonus*, brave*, bright*, brillian*, calm*, care, cared, carefree, careful*, cares,*

caring, casual, casually, certain, challeng*, champ*, charit*, charm*, cheer*, cherish*, chuckl*, clever*, comed*, comfort*, commitment*, compassion*, compliment*, confidence, confident, confidently, considerate, contented*, contentment, convinc*, cool, courag*, create*, creati*, credit*, cute*, cutie*, daring, darlin*, dear*, definite, definitely, delectabl*, delicate*, delicious*, deligh*, determina*, determined, devot*, digni*, divin*, dynam*, eager*, ease*, easie*, easily, easiness, easing, easy*, ecsta*, efficien*, elegan*, encourag*, energ*, engag*, enjoy*, entertain*, enthus*, excel*, excit*, fab, fabulous*, faith*, fantastic*, favor*, favour*, fearless*, festiv*, fiesta*, fine, flatter*, flawless*, flexib*, flirt*, fond, fondly, fondness, forgave, forgiv*, free, freeb*, freed*, freeing, freely, freeness, freer, frees*, friend*, fun, funn*, genero*, gentle, gentler, gentlest, gently, giggl*, giver*, giving, glad, gladly, glamor*, glamour*, glori*, glory, good, goodness, gorgeous*, grace, graced, graceful*, graces, graci*, grand, grande*, gratef*, grati*, great, grin, grinn*, grins, ha, haha*, handsom*, happi*, happy, harmless*, harmon*, heartfelt, heartwarm*, heaven*, heh*, helper*, helpful*, helping, helps, hero*, hilarious, hoho*, honest*, honor*, honour*, hope, hoped, hopeful, hopefully, hopefulness, hopes, hoping, hug, hugg*, hugs, humor*, humour*, hurra*, ideal*, importan*, impress*, improve*, improving, incentive*, innocen*, inspir*, intell*, interest*, invigor*, joke*, joking, joll*, joy*, keen*, kidding, kind, kindly, kindn*, kiss*, laidback, laugh*, libert*, like, likeab*, liked, likes, liking, livel*, lmao, lol, love, loved, lovely, lover*, loves, loving*, loyal*, luck, lucked, lucki*, lucks, lucky, madly, magnific*, merit*, merr*, neat*, nice*, nurtur*, ok, okay, okays, oks, openminded*, openness, opport*, optimal*, optimi*, original, outgoing, painl*, palatabl*, paradise, partie*, party*, passion*, peace*, perfect*, play, played, playful*, playing, plays, pleasant*, please*, pleasing, pleasur*, popular*, positiv*, prais*, precious*, prettie*, pretty, pride, privileg*, prize*, profit*, promis*, proud*, radian*, readiness, ready, reassur*, relax*, relief, reliev*, resolv*, respect, revigor*, reward*, rich*, rofl, romanc*, romantic*, safe*, satisf*, save, scrumptious*, secur*, sentimental*, share, shared, shares, sharing, silli*, silly, sincer*, smart*, smil*, sociab*, soulmate*, special, splend*, strength*, strong*, succeed*, success*, sunnier, sunniest, sunny, sunshin*, super, superior*, support, supported, supporter*, supporting, supportive*, supports, suprem*, sure*, surpris*, sweet, sweetheart*, sweetie*, sweetly, sweetness*, sweets, talent*, tehe, tender*, terrific*, thank, thanked, thankf*, thanks, thoughtful*, thrill*, toleran*, tranquil*, treasur*, treat, triumph*, true, trueness, truer, truest, truly, trust*, truth*, useful*, valuabl*, value, valued, values, valuing, vigor*, vigour*, virtue*, virtuo*, vital*, warm*, wealth*, welcom*, well, win, winn*, wins, wisdom, wise*, won, wonderf*, worship*, worthwhile, wow*, yay, yays.*

Possessive WH-pronouns

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger to the word *whose*.

Possibility modals

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a possibility modal is assigned when one of the following words is found: *can, may, might, could.*

Pre-determiners

This variable was calculated automatically by MAT and it is one of the tags assigned by the Stanford Tagger. The Stanford tagger assigns this tag every time it finds one of the following words that precede an article or possessive pronoun: *all, both, half, many, nary, quite, rather, such.*

Predicative adjectives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a predicative adjective is assigned when an adjective is preceded by any form of BE and followed by a word that is not another adjective, an adverb or a noun. If any adverb or negation is intervening between the adjective and the word after it, the tag is still assigned.

Predictive modals

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a predictive modal is assigned when one of the following words is found: *will, would, shall* and their contractions.

Present participial clauses

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a present participial clause is assigned when a punctuation mark is followed by a present participial form of a verb followed by a preposition, a determiner, a WH pronoun, a WH possessive pronoun, any WH word, any pronoun or any adverb.

Present participial WHIZ deletion relatives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a present participial WHIZ deletion relative is assigned when a present participial form is preceded by a noun.

Present participles

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Present tenses

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Private verbs

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a private verb is assigned when any of these words is found: *accept, accepts, accepting, accepted, anticipate, anticipates, anticipating, anticipated, ascertain, ascertains, ascertaining, ascertained, assume, assumes, assuming, assumed, believe, believes, believing, believed, calculate, calculates, calculating, calculated, check, checks, checking, checked, conclude, concludes, concluding, concluded, conjecture, conjectures, conjecturing, conjectured, consider, considers, considering, considered, decide, decides, deciding, decided, deduce, deduces, deducing, deduced, deem, deems, deeming, deemed, demonstrate, demonstrates, demonstrating, demonstrated, determine, determines, determining, determined, discern, discerns, discerning, discerned, discover, discovers, discovering, discovered, doubt, doubts, doubting, doubted, dream, dreams, dreaming, dreamt, dreamed, ensure, ensures, ensuring, ensured, establish, establishes, establishing, established, estimate, estimates, estimating, estimated, expect, expects, expecting, expected, fancy, fancies, fancying, fancied, fear, fears, fearing, feared, feel, feels, feeling, felt, find, finds, finding, found, foresee, foresees, foreseeing, foresaw, forget, forgets, forgetting, forgot, forgotten, gather, gathers, gathering, gathered, guess, guesses, guessing, guessed, hear, hears, hearing, heard, hold, holds, holding, held, hope, hopes, hoping, hoped, imagine, imagines, imagining, imagined, imply, implies, implying, implied, indicate, indicates, indicating, indicated, infer, infers, inferring, inferred, insure, insures, insuring, insured, judge, judges, judging, judged, know, knows, knowing, knew, known, learn, learns, learning, learnt, learned, mean, means, meaning, meant, note, notes, noting, noted, notice, notices, noticing, noticed, observe, observes, observing, observed, perceive, perceives, perceiving, perceived, presume, presumes, presuming, presumed, presuppose, presupposes, presupposing, presupposed, pretend, pretend, pretending, pretended, prove, proves, proving, proved, realize, realise, realising, realizing, realises, realizes, realised, realized, reason, reasons, reasoning, reasoned, recall, recalls, recalling, recalled, reckon, reckons, reckoning, reckoned, recognize, recognise, recognizes, recognises, recognizing, recognising, recognized, recognised, reflect, reflects, reflecting, reflected, remember, remembers, remembering, remembered, reveal, reveals, revealing, revealed, see, sees, seeing, saw, seen, sense, senses, sensing, sensed, show, shows, showing, showed, shown, signify, signifies, signifying, signified, suppose, supposes, supposing, supposed, suspect, suspects, suspecting, suspected, think, thinks, thinking, thought, understand, understands, understanding, understood.*

Pronoun *it*

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a pronoun *it* is assigned when the word *it* is found.

Pro-verb *do*

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a pro-verb *do* is assigned when DO is not in neither of the following patterns: (a) followed by a verb or followed by adverbs, negations and then a verb; (b) preceded by a punctuation mark or a WH pronoun.

Public verb

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a public verb is assigned when any of these words is found: *acknowledge, acknowledged, acknowledges, acknowledging, add, adds, adding, added, admit, admits, admitting, admitted, affirm, affirms, affirming, affirmed, agree, agrees, agreeing, agreed, allege, alleges, alleging, alleged, announce, announces, announcing, announced, argue, argues, arguing, argued, assert, asserts, asserting, asserted, bet, bets, betting, boast, boasts, boasting, boasted, certify, certifies, certifying, certified, claim, claims, claiming, claimed, comment, comments, commenting, commented, complain, complains, complaining, complained, concede, concedes, conceding, conceded, confess, confesses, confessing, confessed, confide, confides, confiding, confided, confirm, confirms, confirming, confirmed, contend, contends, contending, contended, convey, conveys, conveying, conveyed, declare, declares, declaring, declared, deny, denies, denying, denied, disclose, discloses, disclosing, disclosed, exclaim, exclaims, exclaiming, exclaimed, explain, explains, explaining, explained, forecast, forecasts, forecasting, forecasted, foretell, foretells, foretelling, foretold, guarantee, guarantees, guaranteeing, guaranteed, hint, hints, hinting, hinted, insist, insists, insisting, insisted, maintain, maintains, maintaining, maintained, mention, mentions, mentioning, mentioned, object, objects, objecting, objected, predict, predicts, predicting, predicted, proclaim, proclaims, proclaiming, proclaimed, promise, promises, promising, promised, pronounce, pronounces, pronouncing, pronounced, prophesy, prophesies, prophesying, prophesied, protest, protests, protesting, protested, remark, remarks, remarking, remarked, repeat, repeats, repeating, repeated, reply, replies, replying, replied, report, reports, reporting, reported, say, says, saying, said, state, states, stating, stated, submit, submits, submitting, submitted, suggest, suggests, suggesting, suggested, swear, swears, swearing, swore, sworn, testify, testifies, testifying, testified, vow, vows, vowing, vowed, warn, warns, warning, warned, write, writes, writing, wrote, written.*

Quantifier pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a quantifier pronoun is assigned when one of the following words is found: *everybody, somebody, anybody, everyone, someone, anyone, everything, something, anything*.

Quantifiers

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a quantifier is assigned when one of the following words is found: *each, all, every, many, much, few, several, some, any*.

Relative frequency of *and*

This variable was calculated using a Perl script that counted how many times the word *and* appeared in a text and then dividing this number by the total number of tokens of the text. This ratio was then multiplied by 100 to obtain the relative frequency.

Relative frequency of *for*

This variable was calculated using a Perl script that counted how many times the word *for* appeared in a text and then dividing this number by the total number of tokens of the text. This ratio was then multiplied by 100 to obtain the relative frequency.

Relative frequency of *with*

This variable was calculated using a Perl script that counted how many times the word *with* appeared in a text and then dividing this number by the total number of tokens of the text. This ratio was then multiplied by 100 to obtain the relative frequency.

Second person pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a second person pronoun is assigned when any of these words is found: *you, your, yourself, yourselves, thy, thee, thyself, thou.*

Seem-appear

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where an instance of *seem-appear* is assigned when SEEM or APPEAR are found.

Sentences

A sentence was identified every time a string of words started with capital letter and ended with an end of sentence punctuation. However, some participants and some authors of AMT texts did not use any

sentence boundary and used a new line to mark a new sentence. A script was therefore created to transform new lines before a capital letter into an end of sentence punctuation. These instances were then accounted in the number of total sentences.

Sentence relatives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a sentence relative is assigned when a punctuation mark is followed by *who, who, whose* or *which*.

Short t-units

As detailed in Hunt (1983) a short t-unit is defined as a t-unit that contains less than 10 tokens. A script was created that counted how many t-units shorter than 10 tokens were found in a text and this count was then normalised by the total number of t-units.

Singular proper nouns

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger.

Social words

This variable belongs to the set of variables produced by LIWC (Tausczik and Pennebaker, 2009). The list of stems that LIWC includes in this category is: *acquainta*, admit, admits, admitted, admitting, adult, adults, advice, advis*, affair*, amigo*, anybod*, anyone*, apolog*, argu*, armies, army, ask, asked, asking, asks, assembl*, aunt*, babe*, babies, baby*, bambino*, band, bands, bf*, blam*, boy, boy's, boyf*, boys*, bro, bros, brother*, bud, buddies*, buddy*, bye, call, called, caller*, calling, calls, captain, celebrat*, cell, cellphon*, cells, cellular*, chat*, chick, chick'*, child, child's, children*, citizen, citizen'*, citizens, colleague*, comment*, commun*, companion, companions, companionship*, compassion*, complain*, comrad*, confess*, confide, confided, confides, confiding, congregat*, consult*, contact*, contradic*, convers*, counc*, couns*, cousin*, coworker*, crowd*, cultur*, dad*, dating, daughter*, deal, describe, described, describes, describing, disclo*, discuss*, divorc*, email, email'*, emailed, emailer*, emailing, emails, encourag*, enemie*, enemy*, everybod*, everyone*, everything*, ex, exbf*, exboyfriend*, excus*, exes, exgf*, exgirl*, exhubby*, exhusband*, explain, explained, explaining, explains, express*, exwife*, exwive*, families*, family, father*, fellow*, female*, feud*, fiance*, fight*, flatter*, folks, forgave, forgiv*, fought, friend*, game*, gather*, gave, gentlem*, gf*, girl, girl's, girlfriend*, girls*, give, giver*, gives, giving, gossip*, grandchil*, granddad*, granddau*, grandf*, grandkid*, grandm*, grandpa*, grandson*, granny, group*, grownup*, grudge*, guest*, guy*, he, he'd, he'll, he's, hear, heard, hearing, hears, hed, hello*, help, helper*, helpful*, helping, helps, her, hers, herself, hes, hey, hi, him, himself, his, honey, hubby, human*, husband*,*

individual, infant, infant's, infants*, inform, informs, insult*, interact*, interrup*, interview*, involv*, kid, kid'*, kidding, kids*, kin, ladies, lady, lady's, language*, let's, lets, letter, listen, listened, listener*, listening, listens, love, loved, lover*, loves, loving*, ma, ma'am, ma's, mail, mailed, mailer*, mailing, mails, male, male's, males, mam, man, man's, marriag*, marrie*, mate, mate's, mates, mating, meet, meeting*, meets, members, men, men'*, mention*, messag*, met, mob, mobb*, mobs, mom, mom's, momma*, mommy*, moms, mother, motherly, mothers, mr, mrs, mum, mum's, mummy*, mums, name, negotiat*, neighbor*, neighbour*, nephew*, newborn*, niece*, offer*, organiz*, our, ours, ourselves, outsider*, overhear*, owner*, pa, pa's, pal, pals, pappy, parent*, participant*, participat*, partie*, partner*, party*, people*, person, person's, personal, persons, persua*, phone*, phoning, prais*, private, provide, public, question*, reassur*, receiv*, refus*, relationship*, relatives, replie*, reply*, request*, respond*, role*, roomate*, roomed, roomie*, rooming, roommate*, rumor*, rumour*, said, say*, secret, secretive*, secrets, self, send*, sent, share, shared, shares, sharing, she, she'd, she'll, she's, shes, sir, sis, sister*, social*, societ*, somebod*, someone*, son, son's, sons, soulmate*, speak, speaking, speaks, spoke*, spous*, stepchild*, stepfat*, stepkid*, stepmot*, stories, story, suggest*, sweetheart*, sweetie*, talk, talkative*, talked, talker*, talking, talks, team*, teas*, telephon*, tell, telling, tells, thee, their*, them, themselves, they, they'd, they'll, they're, they've, theyd, theyll, theyre, theyve, thine, thou, thoust, thy, told, transact*, uncle, uncle's, uncles, ur, us, visit*, we, we'd, we'll, we're, we've, wed, wedding*, weds, welcom*, weve, who, who'd, who'll, who's, whod, wholl, whom, whos, whose, wife*, willing, wive*, woman, woman's, womanhood, womanly, women*, word*, write, writing, wrote, y'all, ya, yall, ye, you, you'd, you'll, you're, you've, youd, youll, your, youre, yours, you've*.

Split auxiliaries

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a split auxiliary is assigned when an auxiliary is followed by one or two adverbs and a verb base form.

Split infinitives

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a split infinitive is assigned when an infinitive marker *to* is followed by one or two adverbs and a verb base form.

Stranded prepositions

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a stranded preposition is assigned when a preposition is followed by a punctuation mark.

Suasive verbs

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a suasive verb is assigned when one of the following words is found: *agree, agrees, agreeing, agreed, allow, allows, allowing, allowed, arrange, arranges, arranging, arranged, ask, asks, asking, asked, beg, begs, begging, begged, command, commands, commanding, commanded, concede, concedes, conceding, conceded, decide, decides, deciding, decided, decree, decrees, decreeing, decreed, demand, demands, demanding, demanded, desire, desires, desiring, desired, determine, determines, determining, determined, enjoin, enjoins, enjoining, enjoined, ensure, ensures, ensuring, ensured, entreat, entreats, entreating, entreated, grant, grants, granting, granted, insist, insists, insisting, insisted, instruct, instructs, instructing, instructed, intend, intends, intending, intended, move, moves, moving, moved, ordain, ordains, ordaining, ordained, order, orders, ordering, ordered, pledge, pledges, pledging, pledged, pray, prays, praying, prayed, prefer, prefers, preferring, preferred, pronounce, pronounces, pronouncing, pronounced, propose, proposes, proposing, proposed, recommend, recommends, recommending, recommended, request, requests, requesting, requested, require, requires, requiring, required, resolve, resolves, resolving, resolved, rule, rules, ruling, ruled, stipulate, stipulates, stipulating, stipulated, suggest, suggests, suggesting, suggested, urge, urges, urging, urged, vote, votes, voting, voted.*

Subordinating connectives

This variable was introduced vaguely in Loban (1967) as listing words such as *however, moreover, therefore, because* or *although*. To calculate it, the best approximation was to sum the relative frequencies for those MAT variables that were considered to be more relevant. These were: causative adverbial subordinators, concessive adverbial subordinators, conditional adverbial subordinators, other adverbial subordinators, and conjuncts. These variables not only probably include all the words analysed by Loban (1967) but they are also more comprehensive in including any other frequent subordinator or conjunctive adjunct.

Subordinator *that* deletion

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a subordinator *that* deletion is assigned when one of the following patterns is found: (1) a public, private or suasive verb followed by a demonstrative pronoun or a subject form of a personal pronoun; (2) a public, private or suasive verb is followed by a pronoun or a noun and then by a verb or auxiliary verb; (3) a public, private or suasive verb is followed by an adjective, an adverb, a determiner or a possessive pronoun and then a noun and then a verb or auxiliary verb, with the possibility of an intervening adjective between the noun and its preceding word.

Swear words

This variable belongs to the set of variables produced by LIWC (Tausczik and Pennebaker, 2009). The list of stems that LIWC includes in this category is: *arse, arsehole*, arses, ass, asses, asshole*, bastard*, bitch*, bloody, boob*, butt, butt's, butts, cock, cocks*, crap, crappy, cunt*, damn*, dang, darn, dick, dicks, dumb*, dyke*, fuck, fucked*, fucker*, fuckin*, fucks, goddam*, heck, hell, homo, jeez, mofo, motherf*, nigger*, piss*, prick*, pussy*, queer*, screw*, shit*, sob, sonofa*, suck, sucked, sucks, tit, tits, titties, titty, wanker*.*

Syllables

The number of syllables in a text was determined using a script created with Perl based on a module of Perl called 'Syllable.pm' (Fast, n.d.). The script to count syllables was tested for reliability on 20% of the AMT corpus by dividing the number of errors by the number of syllables. The results were excellent with an average accuracy rate of 99.5%.

Synthetic negations

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a synthetic negation is assigned when the word *no* is followed by any adjective or any noun or for any *neither* and *nor*.

T-units

The t-units were identified manually using the definition given by Hunt (1983). A t-unit boundary was marked every time a new independent clause started, including in cases in which a coordinator was present. A sample of text marked for t-unit is given below (the symbol # marks the beginning of a new t-unit):

I am writing to you because I am not satisfied with the package holiday I purchased and paid for # my first complaint is the plain was 4 hours late # then I had to pay extra for a changed flight # then when I landed the coach wasn't even there # so I had to pay for a taxi which cost me €40 Euros # then when I got to the hotel it was like a squat # it was very dirty # the hotel staff were drinking while they was working # I had to wait in the hotel lobby for an hour and forty five minutes because they didn't even prepare my room for me # then I was expecting a lunch considering I paid all exclusive # the hotel was supposed to be a four star establishment # but the hotel porter told me that they only supply breakfast and a evening meal # at that point I tried to phone your company # but some receptionist told me that you was out for lunch # so I went for a walk considering the beach was supposed to be 10 minutes walk from the hotel # it ended up being a 45 minuntes (sic) walk instead # I was fuming having to make my children walk for 45 minutes in the blazing heat # so I am writing to express my anger towards the way I was treated by the hotel staff # and my family were disgusted # I have tried to contact you by phone but with no success # so I have taken legal action against you unless you give me a refund of £500 # I look forward to your reply

T-units per sentence

This variable was calculated by dividing the total number of t-units by the number of sentences.

***That* adjective complements**

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *that* adjective complement is any *that* preceded by an adjective.

***That* relative clauses on object position**

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *that* relative clauses on object position is any *that* that is preceded by a noun and followed by a determiner, a subject form of a personal pronoun, a possessive pronoun, the pronoun *it*, an adjective, a plural noun, a proper noun or a possessive noun.

***That* relative clauses on subject position**

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *that* relative clauses on subject position is any *that* preceded by a noun and followed by an auxiliary verb or a verb, with the possibility of an intervening adverb or negation.

***That* verb complements**

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a *that* verb complement is any *that* that matches one of the following patterns: (1) preceded by *and, nor, but, or, also* or any punctuation mark and followed by a determiner, a pronoun, *there*, a plural noun or a proper noun; (2) preceded by a public, private or suasive verb or a form of *seem* or *appear* and followed by any word that is NOT a verb, auxiliary, a punctuation or the word *and*; (3) preceded by a public, private or suasive verb or a form of *seem* or *appear* and a preposition and up to four words that are not nouns.

Third person pronouns

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a third person pronoun is assigned when any of these words is found: *she, he, they, her, him, them, his, their, himself, herself, themselves.*

Time adverbials

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a time adverbial is assigned when one of the following words is found: *afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later,*

momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, subsequently, today, to-day, tomorrow, to-morrow, tonight, to-night, yesterday.

Time words

This variable belongs to the set of variables produced by LIWC (Tausczik and Pennebaker, 2009). The list of stems that LIWC includes in this category is: *abrupt*, after, afterlife*, aftermath*, afternoon*, afterthought*, afterward*, again, age, aged, ages, aging, ago, ahead, already, always, ancient*, annual*, anymore, anytime, april, august, autumn, awhile, back, before, began, begin, beginn*, begins, begun, biannu*, bimonth*, birth*, biweek*, born, busy, bye, cease*, ceasing, centur*, childhood, christmas*, clock*, common, constant, constantly, continu*, current*, cycle*, dail*, date*, day*, decade*, decay*, december, delay*, due, during, earli*, early, end, ended, ending, ends, era, etern*, eve, evening*, event, eventually, ever, everyday, fade*, fading*, fast, faster, fastest, february, final, finally, finish*, first, firstly, firsts, followup*, forever, former*, forward*, frequent, frequented, frequenting, frequently, frequents, friday*, futur*, generation*, happening, histor*, hour*, hurrie*, hurry*, immediate, immediately, immediateness, immortal*, inciden*, infinit*, initial*, initiat*, instan*, interval*, january, july, june, last*, late, lately, later, latest, like, long, longe*, march*, meantime, meanwhile, min, minute*, modern*, moment*, monday*, month*, morning*, never, new, newer, newest, newly, next, night, nightly, nights, noon*, november, now, o'clock*, occasional*, oclock*, october, old, olden, older, oldest, once, origin, past, period*, perpetual*, preced*, present, presently, prior, proceed*, quick*, recency, recent*, recur*, repeat*, repetit*, respectively, return*, rhythm*, saturday*, schedul*, season*, seconds, senior*, september*, sequen*, simultaneous*, slow*, sometime, sometimes, soon, soone*, sped, speed*, spring, start, started, starter*, starting, starts, startup*, still, stop, stopped, stopper*, stopping, stops, subsequen*, sudden*, summer*, sunday*, synch*, tempora*, term, terminat*, then, thursday*, til, till, time*, timing, today*, tomorrow*, tonight*, tuesday*, until, updat*, usual, usually, wednesday*, week, week'*, weekend*, weekl*, weeks, when, whenever, while, whilst, winter*, year, yearly, years, yesterday*, yet, young*, youth*.*

Tokens

Text length given by the total number of tokens of each text. The total number of tokens was calculated automatically by MAT.

Total adverbs

This variable was calculated by summing the relative frequencies given by MAT for place adverbials, time adverbials, general adverbs, conjuncts, downtoners, hedges, amplifiers, and emphatics.

Total adjectives

This variable was calculated by summing the relative frequencies given by MAT for attributive adjectives and predicative adjectives.

Total emotion words

This variable was calculated by summing the relative frequencies of positive emotion words and negative emotion words.

Total modal verbs

This variable was calculated by summing the relative frequencies given by MAT for possibility modals, necessity modals and predictive modals.

Total nouns

This variable consists in the relative frequency of proper and common nouns, singular and plural. Nominalisations and gerunds are excluded from this count.

Total personal pronouns

This variable was calculated by summing the relative frequencies of first person pronouns, second person pronouns, third person pronouns and pronoun *it*.

Total proper nouns

This variable was calculated by summing the relative frequencies of singular proper nouns and plural proper nouns.

Total prepositional phrases

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a preposition is assigned when any of the prepositions listed by Biber (1988) is found.

Total verbs

The frequency of total verbs was calculated by summing the frequencies of verb bases, past tenses, present participles, past participles, and present tenses.

Traditional stance adverbs

This variable was calculated with a Perl script that searched for the following words: *maybe, probably, certainly, absolutely, of course, indeed, generally, in fact, usually, roughly, apparently, typically, finally, frequently*. The count was then normalised per the total tokens.

Type-token ratio

The type-token ratio of a text was calculated automatically by MAT by counting the number of types in the first 100 tokens of each text.

Verb bases

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger to the base forms of verbs.

WH questions

This variable was calculated automatically by MAT. The program finds a WH question when the following pattern is matched: any punctuation followed by a WH word and followed by any auxiliary verb, allowing an intervening word between the punctuation mark and the WH word.

WH relative clauses on object position

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a WH relative clause on object position is assigned when the following pattern is found: any word that is not a form of the words ASK or TELL followed by any word, followed by a noun, followed by any word that is not an adverb, a negation, a verb or an auxiliary verb.

WH relative clauses on subject position

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a WH relative clause on subject position is assigned when the following pattern is found: any word that is not a form of the words ASK or TELL followed by a noun, then a WH pronoun, then by any verb or auxiliary verb, with the possibility of an intervening adverb or negation between the WH pronoun and the verb.

WH-clauses

This variable was calculated automatically by MAT. The program used the algorithm implemented by Biber (1988) where a WH-clause is assigned when the following pattern is found: any public, private or suasive verb followed by any WH word, followed by a word that is not an auxiliary.

WH-determiners

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger to the words *which* and *that* when used as relative pronouns.

WH-pronouns

This variable was calculated automatically by MAT and it is one of the tags assigned by the algorithms of the Stanford Tagger to the words *what, who* and *whom*.

Words longer than six letters

A script was created to count any word longer than six letters and this number was then normalised to obtain a frequency per 100 tokens.

Words longer than ten letters

A script was created to count any word longer than ten letters and this number was then normalised to obtain a frequency per 100 tokens.