The Language of Weblogs:

A study of genre and individual differences

Scott Nowson

Doctor of Philosophy

Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh

2006

Abstract

This thesis describes a linguistic investigation of individual differences in online personal diaries, or 'blogs.' There is substantial evidence of gender differences in language (Lakoff, 1975), and to a lesser extent linguistic projection of personality (Pennebaker & King, 1999). Recent work has investigated these latter differences in the area of computer-mediated communication (CMC), specifically e-mail (Gill, 2004).

This thesis employs a number of analytic techniques, both top-down (dictionary-based) and bottom-up (data-driven), in order to explore personality and gender differences in the language of blogs. A corpus was constructed by asking authors to submit a month of text and complete a sociobiographic questionnaire. The corpus consists of over 400,000 words and five-factor personality data (Buchanan, 2001) for 71 subjects.

The thesis begins by framing blogs in the context of other genres, both CMC and traditional, in order to show both the distinctiveness and representativeness of the genre. Top-down content analysis techniques are then employed to investigate the relationship between personality and linguistic features. A number of features correlate with each trait, but upon regression, very little variance is explained.

Bottom-up techniques are more successful. The corpus was stratified into high, low and neutral personality groups to identify distinctive collocations for each. Returning to the raw personality scores, it becomes clear that even a small amount of n-gram context helps account for much more variance in personality. A measure of contextuality (Heylighen & Dewaele, 2002) shows that authors considered high in Agreeableness pay more attention to differences between their extra-linguistic context and that of their audience.

Attention turns to gender, where similar methods are applied to investigate gender differences in language. Many previous findings are confirmed in the blog corpus. In addition, women are found to write more in their blogs than men. More generally, using the British National Corpus, it is shown that women are more contextual, except in the least contextual of genres (academic writing) where there is no difference.

The study concludes by confirming that both gender and personality are projected by language in blogs; furthermore, approaches which take the context of language features into account can be used to detect more variation than those which do not.

Acknowledgements

Acknowledgements tend to begin with supervisors, and who am I to buck the trend. I extend very many thanks to and heap loads of appreciation upon Jon Oberlander and Judy Robertson. Jon has a boundless amount of enthusiasm, whilst Judy has what it takes to keep us on the straight and narrow. Jon may often be as elusive as Baroness Orczy's Pimpernel, but like Batman he always comes through in the nick of time, whilst Judy can always be relied on for sage wisdom and greatly appreciated positive words of encouragement. A quality shout out must also go to Alastair Gill, who having tread a similar path, left some very helpful and entertaining route markers.

I also wish to thank Keith Stenning and Jean-Marc Dewaele (Birbeck) for an enjoyable viva and for suggestions which have made this a more complete piece of work.

In a professional manner I would like the extend thanks to the Economic and Social Research Council, and the AMI project (particularly Jean Carletta) for their financial support; Tom Buchanan at Westminster, Paul Rayson of UCREL, Lancaster and Elizabeth Austin here for helping an unknown student in need of assistance;

Many thanks to all the bloggers in the world for taking an idea and running in every direction with it; special thanks to those that took part in my experiment.

On a personal level I'd like to thank my friends and family; old and new, lost and found. Thanks to Tim Smith, for being the best office mate, always ready to be distracted for talk of comics, television, and life in general; his friendship has been greatly enjoyed and valued over the last four years and I have high hopes it will continue for many years to come. Likewise Gavin Anderson for always being an e-mail away, brimming with positivity and superb distractions of his own. Thanks to Neil Meikle for being around as long as I have, and to Simon Gray for being around even longer.

I always wondered why we save the most important people to the end. I guess because they are the ones we want to remember the most. Saying that, how could I ever forget everything my beautiful wife Amanda has given me and shown me over the years: the love, support, friendship, encouragement, support, great times, laughter, experiences, air miles, support, patience, kind heart, warm smile, unwavering support, unbounded love...without all of this, without her constant pride in me and my work, not only would this thesis be so much less than it is, but I would be less of a man.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Scott Nowson*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis begins with the introductory chapter. Firstly, it introduces the main focus of the study, before making the objectives clearer. Following the statement of the aims of the thesis is a discussion of its boundaries. The structure is then outlined before a summary and statement of hypotheses are presented.

1.1 Introduction to thesis focus

Consider these quotes from the corpus of personal weblogs built for this thesis:

- I don't know how many of you ever experience a similar thing, but well, I just see possibilities around me everyday to be evil, and I have to make an active decision NOT to do it. Like mothers who leave their children in prams outside shops, and people who leave their cars open when they are 'popping into the shop' or whatever.
  I see all these things and think to myself "I could steal that" or "I could run off with that" and other such things. I don't. I'd like to make that perfectly clear. I don't. But I do think about it.
- I am writing this hesitantly, because I am conscious that you reading this may be thinking "What category do I fit into?" I am also conscious that I am coming over as egotistical in assuming that you care; all I know is that when someone cites me, I feel warm. When someone de-links me I feel disappointed.

It's clear that both quotes concern the author's attitude toward other people: the latter is clearly more concerned about what other people feel and think of them than

the former. It is probably obvious to the reader that the quotes are from different individuals. But how do they know this, and what does each sample tell us about its author?

Individual differences play an important role in every aspect of human life. Whether the reasons for our differences are biological or psychological, instilled by nature or learnt from nurture, they affect us every day. Perhaps two of the most important individual differences are those of gender and personality.

Gender: the most observable of differences; traditionally straightforwardly defined; affects people on many levels – physically and mentally, externally and internally. The differences between men and women are manifest.

Personality: often less obvious, but no less important; has a contentious definition; based in our heads, manifests in many ways.

Both a person's gender and their personality are important in identifying them as an individual. They not only affect how people are, but how they are perceived. Today the majority of gender and personality trait recognition is done in face-to-face communication. It is easiest to get an impression of a person's character for example when their every mannerism can be observed, every intonation heard. However, it is not just *how* a person says something that reveals their gender, that conveys a sense of their personality. *What* they actually say, the words and phrases they choose for production reveals a lot about them; language can carry rich suggestions of both gender and personality. (cf. Lakoff, 1975; Pennebaker & King, 1999). Men swear more than women, while women use more pronouns; Extraverts talk more, while high Neurotics are more immediate in their writing style.

Recent work by Gill (2004) has investigated further the claims that personality is projected in language; specifically the language of e-mail. The internet is in fact increasingly being considered as a resource for linguistic study (Keller, Lapata and Ourioupina, 2002). A number of studies have focused on the nature of various types of computer-mediated communication (CMC), such as asynchronous e-mail and synchronous chat. Both gender differences and to a lesser extent personality have been investigated in CMC (cf. Herring, 2000; Gill, 2004).

The internet is still relatively young and new genres continue to emerge (Crowston & Williams, 2000). The fluid nature of these genres means that there is little in the way of expectation as to their nature; they are not restricted by a set of standards or explicitly taught rules. This allows a great deal of room for individuality. One genre that has been little studied at this level is that of the online personal diaries, or 'blogs' as they are more commonly known, that opened this section. There are very many millions of blogs in the world, many of which are updated on a daily basis (Rainie, 2005). They therefore provide a wealth of individually authored text with which to study individual differences.

This thesis will explore the projection of both personality and gender in the language of blogs. In addition to this, it will explore the linguistic properties of blogs as a genre, looking at how representative of language in general they can be. If the ways in which individual differences affect language can be understood, it becomes easier to recognise character without the traditional physical cues. Understanding what type of language is being employed can further inform this.

While it is not a direct concern of this thesis, a natural application of understanding language use is to better inform generation methods. So, one further motivation for this work is the long term goal of personality rich natural language generation. From a human-computer interaction perspective it has been shown that users relate better to systems that imply a personality more similar to their own (Nass & Lee, 2000). There has been also been work attempting to create autonomous agents that can convey a sense of character (cf. Mateas, 1997). A future application of this work would be to provide a list of features which can be used to generate language that projects a sense of personality.

1.2 Objectives

This thesis reports work of a mostly exploratory nature: its main focus is an exploration of individual differences and language. The individual differences that this thesis will focus on are personality and gender. This therefore presents the general hypotheses of this thesis: personality and gender are projected by language; there are linguistically

identifiable differences between genders and within personality traits. The aim is to identify linguistic features which can be used in the future to detect or indeed project a specific personality type or gender.

The area of language in which this hypothesis will be explored is computer-mediated communication. More specifically, the language to be studied is taken from online personal diaries, or blogs. Yet why are blogs different from any other choice of text, and, conversely, what makes them representative enough of language in general to be worth studying? Answering these questions is the secondary objective of this thesis; to explore the properties of the language in blogs that both sets them apart and makes them similar to other genres.

There is one further methodological objective, which answers the question why study gender when so many have before? The intention is that finding gender differences that have been observed in other genres will help confirm the hypothesis above: that blogs are representative of, if not language as a whole, then at least language in CMC in general. Additionally, this will show that the methodologies employed in the thesis are clearly capable of identifying differences in language due to gender, and add validity to findings relating to personality.

Note that these objectives are merely the general aims of this thesis. More specific hypotheses will be put forward at each stage of work. These will be informed by previous findings, an understanding of the nature of the differences being explored, and the properties of the techniques being employed. By addressing more specific questions at each stage, the general hypotheses will be tested.

1.3 Boundaries of the thesis

In addition to identifying the aims of this thesis—explaining what the thesis will attempt to do—it is also important to identify its boundaries—what it will not do. This allows focus to be kept on those areas which are of most interest, without losing track by trying to cover everything.

This thesis may be about personality, but it is not concerned with *personality theory*. The crux of this is that although extensive reference is made to particular personal-

ity traits, these are used merely to inform the study while not being the object of study themselves. It is not the aim of this study to explore the validity of personality traits, and certainly not trait theory itself. The thesis will however provide a background to these themes in order to situate the study within them.

Similarly, gender is also not the object of study. This thesis will not speculate as to *why* differences in language between genders may occur. It may be that there are cognitive difference between males and females, or sociological effects that alter an individual's projection of self. But that is not of importance here. It is merely assumed that differences are related to underlying phenomena.

And, while language *is* the focus of this thesis, language *production* is not. Why personality might inter-relate with cognitive language production capabilities is not a concern here. Nor is the question of how findings could specifically be used to inform natural language generation systems.

Finally, while this thesis aims to make claims about language in general, only language of blogs is actually under scrutiny. However, it is considered that blogs are representative of, and share many similarities with other computer-mediated forms of communication. Results may not be generalisable to all forms of language, but relevance within CMC is assured.

1.4 Structure of the thesis

This thesis is structured as follows. Chapter two presents a survey of the literature relevant to the work to be reported. Personality trait theory is first introduced, highlighting the reasons for selecting the model of personality to be used. This is then introduced with a description of the traits which are the focus of this aspect of the thesis, and how these traits would appear to relate to language. Alternative personality theories to those of traits are then briefly discussed. This is followed by a report on previous linguistic studies of personality, with a focus on recent work on personality projection in e-mails. Attention then turns to gender: first with an introduction to language-centred gender studies; and then with a summary of findings from the literature. The area of the secondary objective is then introduced: genre. In attempting to define genre, a number

of approaches are highlighted. This leads to discussion of work looking specifically at genres within computer-mediated communication. The introduction of CMC leads to more background on the specific genre of interest here: personal diary blogs. The reasons for choosing to study diary blogs are made clear, and an introduction to the area as a whole is provided. This then prompts a review of work exploring language use in CMC, and specifically those studies which have focused on weblogs. The rest of the chapter introduces the tools that will be used for linguistic analysis of text, including a discussion of methodological issues.

Chapter three describes the creation of the blog corpus upon which the work of this thesis is based. It explains how the data was collected and prepared. It will also look at some basic statistics and explain some approaches to the data adopted during the course of the thesis.

Chapter four deals with the corpus as a whole, reporting work which attempts to delineate the distinctiveness of blogs as genre. Two analyses are reported which use unitary linguistic measures to place blogs in the context of genres drawn from the BNC.

Chapter five reports the results of work of top-down (dictionary-based) content analysis. Here, a number of dictionaries reporting psychological categories and psycholinguistic information of words are used. Relationships between these categories and properties with personality traits are examined.

After concluding chapter five with a discussion of some of the drawbacks of the dictionary approach, chapter six adopts a number of bottom-up (data-driven) analyses in the study of individual differences. Techniques are employed from a number of sources in the corpus linguistics community.

In chapter seven attention turns from personality to gender. The toolset of analyses built thus far are used to explore differences between the language of male- and female-authored blogs.

Chapter eight, the final chapter, provides both a summary of the thesis, alongside conclusions drawn from the work reported within. It also discusses implications of the work and suggests future directions for study.

1.5 Summary and statement of Hypotheses

This chapter has introduced the key areas of study for this thesis — the relationship between language and individual differences, namely gender and personality — and described why this is an important field for study. This chapter has also indicated the objectives and boundaries of the thesis. The structure of the rest of the thesis was then outlined. The major goals of this thesis are to test whether and the what extent:

Hypothesis 1: Blogs are distinct yet representative of more general language.
Hypothesis 2: Personality is projected linguistically in blogs.
Hypothesis 3: Gender is projected linguistically in blogs.

Chapter 2

Literature Review

This chapter presents a survey of the fields from which work on this thesis is drawn, the most significant being personality, gender, genre and computer-mediated communication (CMC). The chapter begins with an introduction to trait theories of personality, and highlights the selection of the five-factor model to be used in this study. After presenting some alternatives to trait theory, previous findings for language differences due to personality are discussed. The focus is next on gender, first with an introduction to gender difference work, and then a summary of results from studies of language. After next introducing definitions of genre, there is a review of work looking specifically at genres in CMC. This leads to an introduction to the CMC genre to be used in this study, weblogs. The reasons for choosing them, along with further details of what they are and how they have been used and studied, are reported. There then follows a discussion of studies of language in CMC generally, before specifically focusing on weblogs. The final section introduces the tools and techniques to be employed, along with methodological issues arising from their use.

2.1 Introduction to Personality

2.1.1 Trait theories of personality

This thesis is mainly concerned with personality as modelled by the five-factor model (Digman, 1990; Costa and McCrae, 1992; Wiggins and Pincus, 1992; Goldberg, 1993),

though it also makes reference to Eysenck's three-factor model (Eysenck and Eysenck, 1991; Eysenck et al., 1985). Both these models are 'trait' approaches, whereby personality is reduced to a number of measurable factors, or traits.

Factors are seen as a scale, with possible scores ranging from 'low' to 'high'. Factors are considered to be orthogonal and independent of one another, the score on one trait predicting nothing of another trait. In practise however there may be some relationship between traits (cf. Matthews et al., 2001; section 3.4.4 for data from this study).

The trait approach assumes that individuals have stable personality characteristics (Cloninger, 1996). This distinguishes personality from more transitory states such as mood or emotions. Traits are intrinsically linked with behaviour, both in a causal and informing relationship, though this is often highlighted as a criticism.

> Mike is aggressive. How do we know? We have seen him beating up on people. Why does he? His trait of aggressiveness causes him to beat up on people. The trait explains the behaviour, and the behaviour is the reason that we infer the trait. That is circular reasoning. It does not offer a satisfactory explanation of behaviour. (Cloninger, p76)

When it comes to studying personality, there are alternatives to trait theory. Though thoroughly placing traits within the field of psychology is not the intention of this thesis, some of these alternatives will be briefly discussed in section 2.1.3. Despite these alternatives however, the main proponents of trait theories see their increasing use in experimental situations as acceptance of their validity.

Not only is there debate as to the nature of personality theory, but there have been many trait models proposed with the previously mentioned three- and five- factor models amongst the most commonly studied today. Interestingly, the first two traits of both the models to be discussed here are the same: Extraversion (also known as Extraversion-Introversion) and Neuroticism (Emotionality-Stability). The nature of these traits is undisputed, and form the heart of many theories of personality (Matthews et al., 2001).

Where the models diverge however is not just in the number and definition of the remaining factors, but more importantly their theoretical basis. The five-factor model employed in this thesis is derived from lexical studies, and adds the traits of Openness,

Agreeableness and Conscientiousness. The 'lexical approach' is concerned with the dimensions people use when describing themselves and others. This research suggested five major factors that ordinary people use to describe personality. Goldberg (1981) referred to these as the 'Big Five'. These factors have shown validity after replication (McCrae & Costa, 1987, 1997; Funder, 2001).

Eysenck however, claims a 'biological basis' for his model (Eysenck, 1970; Eysenck & Eysenck, 1991) and adds a factor termed Psychoticism. Eysenck emphasises his traits' validity since they are based on invariable aspects of human existence (Eysenck, 1993).

No matter the basis, debate continues: Eysenck maintains that Agreeableness and Conscientiousness are merely (negatively related) facets of Psychoticism (Eysenck, 1993); conversely, Costa and McCrae, amongst other, have argued that five factors are required in order to describe personality fully (Costa and McCrae, 1992; McCrae and Costa, 1997; Digman, 1990; Goldberg, 1993).

It is not a goal of this thesis to prove or disprove any particularly theory or model of personality, nor will it continue discussion of the debate any further. For the rest of this thesis, work is concerned with the five-factor model of personality (Costa & McCrae, 1992). The specific measurement instrument to be used will be discussed in section 2.7.3.1. This is not necessarily a preference for one model over the other. However, it is necessary to explain the choice behind this model, particularly in light of the work of Gill (see section 2.2.1). As will become apparent, this work most closely resembles that of Gill, yet he chose Eysenck's three-factor EPQ-R (Eysenck & Eysenck, 1991). By discussing the reasons for his choice, the reason the five-factor model is chosen here should be clear.

> *Theoretically*, a biological or neural description of personality is desirable, since this research is conducted from a cognitive science perspective, and we may want to integrate theories of language production with theories of personality. (Gill, p12)

Indeed, integrating theories *is* desirable (cf. Dewaele & Furnham, 2000; Dewaele, 2002). However, just because Costa and McCrae's dimensions were not founded on biology does not mean there is no fundamental basis for the model. McCrae et al.

(2000) maintain that their factors are indicators underlying dispositions of human nature which are genetically influenced. Indeed, since personality traits can be at least partially inherited, Matthews et al. (2003) posit that there must be a biological influence on traits. As mentioned previously, there are more trait models than discussed here. One that combines a more than passing resemblance to at least four of Costa and McCrae's five-factors with the biological inspiration of Eysenck is the work of Zuckerman (1995).

In addition to the basis of the EPQ model, Gill also provides support for his choice with a quote from Kline (1993(*a*)) which highlights its strengths. However, the quote also highlights a flaw in the broadness of the EPQ categories. Gill maintains the breadth is an advantage since the three factors provide a reduced model of personality to work with compared to the five. Indeed the broader the model the easier it is to work with, but this can also mean that effects can be lost within a broad factor. Working with a more refined model allows for a more fine grained comparison. In this case it is particularly relevant: Neuroticism and Extraversion are generally considered to be equivalent between the two models; and even Eysenck himself considers Agreeableness and Conscientiousness to be facets of Psychoticism. By extending the model in this way, by allowing more factors of variation, it is hoped that more subtle differences can be identified. Following Gill's three-factor study it seems appropriate to extend the model and conduct a five-factor study.

To quote De Raad and Perugini (2002) in final support of the choice of the five-factor model:

> The Big Five model has acquired the status of a reference model ... its five main constructs capture so much of the subject matter of personality psychology.

2.1.2 Personality traits

In describing the personality traits, this section refers to the facet scales of the NEO personality inventory (adapted from Costa, McCrae & Dye, 1991). No attempt is made here to discuss associated behaviours of each trait. However, inferences will be made as to the expected effects that these facets will have on features and properties of the

language that extreme types will use.

2.1.2.1 Neuroticism

The six facets of Neuroticism are: Anxiety, Hostility, Depression, Self-consciousness, Impulsiveness, Vulnerability. This presents a number of expectations as to the language used by high Neurotics: they will use more words associated with negative emotions, particularly anxiety, anger and sadness; they will talk more about themselves than other people, reflected in greater use of first-person pronouns; their impulsive nature suggests their writing will take less consideration of their audience as they they are more reactive.

2.1.2.2 Extraversion

Extraverts are considered to display: Warmth, Gregariousness, Assertiveness, Activity, Excitement seeking, Positive emotions. Extravert language (as opposed to Introvert language) would therefore be characterised by: more references to positive emotions; fewer tentative and hedge terms; more references to other people, third-person pronouns, and social activities; greater use of verbs.

2.1.2.3 Openness

Openness is categorised by the following traits: Fantasy, Aesthetics, Feelings, Actions, Ideas, Values. These suggest that the language of individuals high in Openness will contain: more references to feelings, good or bad; more abstract terms, less concrete language; greater use of words relating to higher level cognitive processes and beliefs; greater and broader use of verbs.

2.1.2.4 Agreeableness

The facets considered aspects of Agreeableness are: Trust, Straightforwardness, Altruism, Compliance, Modesty, Tender-mindedness. This suggests that highly Agreeable individual's language will: reflect more consideration of their audience; less aggression, swearing and negative emotions; fewer references to self;

2.1.2.5 Conscientiousness

The facets of Conscientiousness are: Competence, Order, Dutifulness, Achievement Striving, Self-discipline, Deliberation. Highly Conscientious language should be reflected by: clearer, more concise language; greater discussion of achievements; fewer topic changes.

2.1.3 Alternative theories of personality

Trait theories are not the only models of personality. This section is not an in-depth study of personality analysis, but merely serves to briefly introduce alternative approaches. Traits are useful in this study because they are quantifiable. Traits fall under the branch of psychology in which aspects of the mind are scientifically measured, psychometry. One alternative methodology is psychoanalysis, which serves to explore the relationship between aspects of the conscious and unconscious mind, which are, for the most part, far less measurable.

Perhaps the most well known psychoanalytic theory is that of Sigmund Freud's psychodynamics. Freud maintained that personality was defined by the use of a fixed amount of instinctual energy (the 'libido') invested across mental structures concerning basic biological drives ('id'), and the reality-focused sense of individuality ('ego') and conscience ('super-ego'). Perhaps most famously Freud also theorised stages of psychosexual development and associated complexes, such as the Oedipus complex. Psychoanalysis maintains that many aspects of personality are evoked by conflicts between the mental structures. For example, the gratification required by the id often goes against the ego's need to maintain social rules.

Many researchers have felt that Freudian theories did not adequately explain the phenomena they were discovering, however, and so developed theories of their own. In many recent studies, higher prominence has been given to those parts of the mind considered unconscious. The essential idea is that a large proportion of mental processing is inaccessible to the conscious mind. Whereas it has been argued that the key ideas of psychoanalysis are not scientifically testable (Popper, 1957), unconscious cognitive processes can be studied with the use of subliminal stimuli.

One assumption that trait theory makes is that traits are not only stable across time, but that they are stable across individuals. Trait theory can be seen to formalise the awareness people have that persons of a similar disposition can be grouped together. Advocates of humanistic and phenomenological approaches maintain that there is more individuality behind personality. They argue that unique personal experience is important since, for example no two 'Extraverts' share the same idiographic background or have had the exact same life experiences; it is one's life that makes one different.

Closely related to these approaches is the situationist criticism of traits. This argument, most significantly levelled by Mischel (1968) maintains that traits cannot singularly explain behaviour. From this criticism rose interactionism: the balance that both person and situation contribute to behaviour. While no researcher would hold with either of those explaining personality in isolation, widespread situational studies pose more significant issues. In the scope of this thesis, situation is disregarded in favour of obtaining general patterns of language.

This section has highlighted some of the main areas in which theories of personality are based. There are many studies which have investigated the relationship between aspects of the above methodologies and different personality traits, with varying degrees of success. Many of these studies serve to support trait theory, while others help to situate traits in a larger domain of psychology. For a more in depth discussion of alternatives to trait theory and their relationship with traits, see Matthews et al. (2001).

2.2 Personality and Language

As mentioned previously, with regards to personality this thesis is concerned with trait theory. With this in mind, focus here is on observation of language use as it relates to exhibition of these traits. Reviewing this literature leads to a number of observations: firstly, of the little work that has been conducted in the field, it tends to employ 'inconsistent' methods (Furnham, 1990) and be spread across many disciplines; secondly, much of the work has been concerned with speech rather than writing; thirdly, research generally concentrates on Extraversion, and to a lesser extent Neuroticism, rather than the other traits of the three- (Psychoticism) and five-factor models (Openness, Agree-

ableness and Conscientiousness). Indeed, on this last observation, Pennebaker et al. (2003) noted that they were only aware of one single study which was concerned with language and the 'big five' personality dimensions.

The interdisciplinary nature of this research question appears to be the main explanation for the lack of work in the field. It touches on aspects of many fields including personality, social psychology, socio- and psycholinguistics (Furnham, 1990). This interdisciplinarity also explains the inconsistency to the approaches adopted. Certainly one reason for the concentration on Extraversion and Neuroticism may stem from the debate around the remaining traits, as briefly outlined in the previous section. Indeed the instability of some traits discourages those not directly subscribing to one model or another from exploring those avenues of research. That studies have concentrated on speech is clear due to the extra paralinguistic information it offers: pronunciation, intonation or volume, all of which are easily observable.

One reason for lack of work in the area of personality and Second Language Acquisition was a series of studies by Naiman et al. (1975, 1978). After failing to replicate their earlier findings with Extraversion, rather than re-assess their methods, they completely dismissed the instrument (the Eysenck Personality Inventory) as failing to adequately measure what they felt extraversion–introversion was. This had large repercussions in the SLA field, with very few studies that followed concerning themselves with personality (cf. Dewaele, 2005).

In terms of speech versus writing, it is not necessarily the case that more work has been conducted in one than the other. Dewaele and Furnham (1999) reviewed 33 SLA studies, and they found many significant correlations reported in studies of oral language. However, they found no relationships were ever found between Extraversion and data derived from written material.

2.2.1 Previous findings

In the following description of previous work on personality and language, focus is on four areas most relevant to written language : fluency, morphology and syntax, conversational behaviour (cf. Scherer, 1979; Furnham, 1990) and content analysis. As already mentioned, much of the work in this field has concentrated on Extraversion,

and so the majority of findings are for this trait. Results concerning other traits will also be reported, but only those relating to the three- and five-factors models.

Personality in e-mail Prior to this work, Gill (2004) performed a similar study investigating language and personality, framing his study in e-mail text. Along with collecting sociobiographic data (using the EPQ-R three-factor inventory) Gill asked subjects to write two 'e-mails' to a good friend whom they hadn't seen for quite some time. The subjects were instructed to first write about '*what has happened to you, or what have you done in the past week*,' while the second was about '*what your plans are for the next week*.' The e-mail corpus consisted of 105 subjects and approximately 60,000 words.

This work has employed (often with a degree of adaptation) a number of the Gill's methods. Therefore, in discussing previous results, close attention is paid to the work reported by Gill and colleagues.

2.2.1.1 Fluency

The majority of fluency studies report an Extravert advantage, and are mostly reported in terms of speech rate. Extraverts have higher speech rates (Siegman, 1987) in both formal and informal situations (Dewaele, 1998; Dewaele and Furnham, 2000). Extravert speech is also less likely to contain silence (Siegman, 1978; cf Dewaele, 1998, for issues of silent pauses and measurement of speech rate). Similarly, in complex verbal tasks, Introverts' pauses before speaking were significantly longer (Ramsay, 1968). In formal situations Extraverts have been found to show less hesitation ('er') but make a higher proportion of semantic errors (Dewaele and Furnham, 2000).

2.2.1.2 Syntax

This section reports findings of mainly grammatical features. 'Zestful' individuals, most closely related to Extraverts, have been shown to use more pronouns, adverbs, verbs and words in total (cf. Furnham 1990; Dewaele and Furnham, 1999). Similar characteristics were also found for Extravert non-native speakers: a factor analysis of syntactic tokens led Dewaele and Furnham (2000) to describe this as implicit language

(a preference for pronouns, adverbs and verbs), which contrasts the more explicit language of Introverts (nouns, modifiers and prepositions). This finding related to both formal and informal situations, and mirrored previous analysis of the individual categories (Dewaele, 1996).

However, similar measures do not identify a similar relationship. Beyond the idea of implicitness, Heylighen and Dewaele (2002) developed a measure of contextuality/formality based on parts of speech (for greater discussion of this measure, see section 2.7.2.2). Their F-measure balanced relative frequency counts of deictic parts-of-speech (pronouns, verbs, adverbs and interjections) against non-deictic (nouns, adjectives, preposition and articles). Previous findings suggest that Extraverts would use more contextual speech, while Introverts more formal. They found no relationship with the F-measure for level of Extraversion, except in high-stress speech-based situations (such as an oral examination), when extreme Introverts were more formal. They did however hypothesise that Openness, in its capacity as the factor of intellect, would show a positive relationship with F-measure, though they were unable to verify this claim.

Similarly, Oberlander and Gill (2004) investigated this 'implicit-extravert hypothesis' on their e-mail corpus. Following a number of previous findings (cf. Pennebaker & King, 1999; Gill & Oberlander, 2003) they also justified their exploration of a 'implicit-neurotic hypothesis', whereby high neurotics also used implicit language. They only found some parts-of-speech to be used with significantly different relative frequencies by the sub-groups of their subjects. However, n-gram analysis for part-of-speech sequences revealed some support for both hypotheses. Subsequent re-analysis has confirmed this effect (Oberlander & Gill, 2005). Unigram analysis found high Extraverts to make greater use of conjunctions and adjectives, high neurotics to use more conjunctions, and low neurotics more nouns and adverbs (Oberlander & Gill, 2005).

2.2.1.3 Conversational Behaviour

This section looks briefly at how personality can effect interpersonal communication. As would be expected from the social aspect of Extraversion, Extraverts show greater desire to communicate and initiate interaction (McCroskey & Richmond, 1990; cf.

Yellen et al., 1995, for similar finding in CMC). Extraverts also initiate more laughter within a conversation, and talk more (Gifford & Hine, 1994). Other studies have also found that Extraverts use a greater total number of words (Carment et al., 1965; Campbell & Rushton, 1978). However, studies of second language speakers have shown that while the overall text or speech produced is longer, the longest utterances are actually shorter, especially in informal situations (Dewaele, 1995; Dewaele and Furnham, 2000): Extraverts say more, but in shorter bursts.

2.2.1.4 Content Analysis

Here, work on personality relationships with the content of language is reported. Coding of conversational speech acts has found that Introverts use more hedges and talk of problems, but Extraverts express more pleasure talk, agreement, compliments and tend to focus on discussing extracurricular activities (Thorne, 1987, it is worth noting that this study found no significant differences between groups for talk time or number of speech acts). Extraverts have also been shown to use more self-referencing expressions (Gifford & Hine, 1994).

The Linguistic Inquiry and Word Count (LIWC; Pennebaker & Francis, 1999; Pennebaker et al., 2001[1]) has been used in a number of studies to investigate personality. Results are discussed in more detail here since this is an analysis method to be used in the context of this thesis. For more specific details on the LIWC text analysis method see section 2.7.1.1.

Pennebaker and King (1999) applied LIWC analysis to texts written by authors for whom five-factor personality information was available. Using factors derived from LIWC variables, they found: Neuroticism correlated strongly with 'Immediacy' (greater use of First-person singular, Present tense words, Discrepancies and fewer Articles and Words of greater than 6 letters); Extraversion correlated negatively with 'Making distinctions' (greater use of Negations, and Discrepancy, Exclusive, Inclusive and Tentative words), but also positively with some aspects of the 'Social past' (most significantly Social words and Positive emotion words); Openness correlated nega-

[1]Following Gill, this thesis uses the earlier version of the LIWC, and so only this version shall be cited in the future.

tively with great strength with ‘Immediacy’; Agreeableness less so but positively like Neuroticism; Conscientiousness follows Extraversion with a strong negative correlation with ‘Making distinctions.’[2]

Gill (2004) replicated both the factor analysis and the correlation study, albeit with the traits of the EPQ-R. While he did reproduce similar factors, there were fewer after applying Pennebaker and King’s selection approach to his own variables. The only significant relationships Gill’s data provided for Neuroticism. High Neurotics use fewer terms associated with the ‘Social past’ and perhaps contradictorily more Inclusive words.

Gill did however extend his study to include all LIWC variable. He found a number of them correlated at least marginally significantly with the three personality dimensions of his study. However, upon linear regression, he found few remained related and those that did explained little variance in the dimensions. High Neurotics used more Inclusive words and Total first-person pronouns. High Psychotics used fewer First-person singular pronouns, but more words reflecting Cognitive mechanism. At the highest level, nothing was retained for Extraversion.

Cloninger (1996) has suggested that language could be used to investigate the personality of those who are unwilling or unable to complete personality tests, such as the famous or deceased. Pennebaker and Lay (2002) have done just that with former New York Mayor, Rudolph Giuliani using the LIWC. It was perceived that during his time as Mayor, Giuliani underwent a number of apparent personality changes, due to personal crises, and later the terrorist attacks on September 11 2001. The language used in 35 of Giulianis press conferences given between 1993 and late 2001 was analysed, and it was found that his linguistic style had indeed undergone significant change.

2.2.1.5 Further Analysis

This section reports on other types of analysis implemented by Gill in his study of personality in the language of e-mails. As well as the LIWC (Pennebaker & Francis, 1999), Gill also used the MRC Psycholinguistic Database (Coltheart, 1981; Wil-

[2]With respect to the individual LIWC categories which were found to correlate, full discussion can be found in section 5.2, while the full results are replicated in table B.4.

son, 1987; for greater explanation see section 2.7.1.2). After linear regression, Gill again found that little variance was accounted for by the properties in the database (Gill, 2004). Low Extraverts, like high Neurotics, use more concrete language. High Neurotics also showed a preference to language common in speech. High scores on Psychoticism tended to use more unusual non-dictionary words, though of the more standard words they used, used language of more varied frequency.

The e-mail corpus has also been used for a word n-gram analysis, with the subjects stratified into high, neutral and low personality groups (Oberlander & Gill, 2005; again, greater discussion of this methodology can be found in section 2.7.2.1). A number of bi- and trigrams were found to be significantly overused by each of the extreme personality groups of Neuroticism and Extraversion. They identified patterns of n-grams containing nominals and inclusive words in the Extraverts group, suitably reflecting their more sociable nature. They also found Extraverts using more phrases reflecting certainty, while Introverts were more tentative. There was also a social effect in the low Neuroticism group, with a higher use of third-person references. High neurotics were shown to used multiple punctuation collocations.

2.3 Gender and Language

2.3.1 Gender differences

Unlike many other individual differences, such as personality, the link between language and gender has been extensively studied. However, before discussing the results of such studies (see section 2.3.2) it is worth observing the generalisations being made. Much work on gender differences treats each gender as a distinct conforming group: that is to say it is generally assumed that all men behave in a similar manner while women are equivalently consistent.

However, in recent times this binary difference has come under increasing scrutiny. The main criticism is that the diversity within each gender is completely ignored. It is fair to say that differences among men and women may be as great or greater than the difference between the two groups. The most basic argument is that individuals are not solely defined by their gender; there are many socio- and ethnographic differences

between individuals. This has led to an increasing focus on gender diversity.

Early work by variationists (summarised in Labov, 1990) showed that when considering language change, men and women lead in different ways. However, Nichols (1998) found that while younger creole women in South Carolina shifted to a more modern style of language more than men of a similar generation, older women were more traditional than their male peers. Eckert (1997) identified linguistic traits of two social groups within a high school setting, and found that within both these groups, it was the female members that were more advanced in the use of group variants.

Eckert's explanation for this concerned women's status consciousness. Women often have to fight harder to show that they are worthy of group membership, and are likely to do so stylistically, via language use for example. She maintained that this shows as much inequality of gender as difference shaping language.

The work of variationists focuses mostly on determining different linguistic patterns. Alternative work in discourse styles seeks to explore how language differences reflect social processes: for example the masculinity of male language. Among the earliest work in this area is that of Lakoff (1975). She maintained that women's lack of power in society, which leads to a lack of confidence, was reflected in less assertive speech. This argument was based on her observations, rather than empirically derived findings, that women have a higher degree of politeness, less frequently use swearing, and more frequently use tag questions, intensifiers and hedging expressions.

Lakoff's theories traditionally fell under the 'dominance' approach to gender discourse styles, whereby linguistic behaviour can be explained by men's power and women's submissiveness. However, it has also been argued that her theories fit a 'deficit' approach in which male language is seen as the norm from which female language diverges. A more balanced approach is that of unbiased 'difference.' Tannen (1990) sees men and women as different in the same way that people of different nationalities are: there is no superior-inferior relationship. She has suggested that the difference is due to individuals simply internalising different learnt norms of communication.

Of course, despite the differing perspectives on language, discourse style returns gender to a binary variable, ignoring the diversity within. One point worth noting is

that gender here is considered distinct from sex. The language of interest in this study is that which is relevant to socially differentiating between men and women. That is to say it is not necessarily the case that male and female language differs due to different biological mechanisms, but because social conditioning and relationships affect the language of men and women.

This study does generalise the differences between genders, assuming a normal level of variance within each. The same however can also be said across each personality class studied. The inter-relationship of aspects studied here is not considered; not least due to the small number of subjects. It is worth noting, however, that there is no significant difference between the genders on any personality traits.

2.3.2 Previous findings

As mentioned above, this study treats the language of men and women as being of even difference, not looking within gender. This section looks at some of the more relevant results in gender difference studies. There have been a number of surveys of the field which have highlighted many of the most commonly found differences (Mulac et al., 2001; Pennebaker, Mehl and Niederhoffer, 2003; Groom & Pennebaker, 2005): typical male language consists of references to quantity, adjectives reflecting judgement, higher incidences of articles and prepositions suggesting concreteness, notably greater use of taboo/swearing words, and men are more likely to discuss impersonal topics such as occupation, money and sports; female language is much more personal, with greater references to emotions, higher use of pronouns and references to other people, uncertainty verbs and hedges.

Within the Conversational sub-corpus of the BNC, studies at word level (Rayson, Leech and Hodges, 1997) also found that men swear more and women use more female pronouns and first person. In general it was found that men used more common nouns, while women had greater relative frequency of proper nouns, pronouns, verbs. The difference between proper nouns and nouns was attributed to women's preference to talk more about other people.

Also within the BNC, gender differences have been examined within parts of the fiction and non-fiction sub-corpora (Argamon et al., 2003; Koppel et al., 2002). The

authors distinguished their work from that of others by highlighting that most gender language studies are conducted on speech, with a few in informal written genres. Speech allows for intonational cues for example, which do not appear in writing. Even using transcribed speech (Rayson et al., 1997) some features of spoken language such as fillers (eg. 'umm' or 'err') can still be used. Features studied from the written BNC included over 400 function words, and the most frequently occurring parts-of-speech n-grams. This allowed for identification of both general trends and more specific differences: men tended to use more prepositions generally, though women used 'for' and 'with' significantly more. Similarly, while women overall used more pronouns, men used 'he' just as much.

Interestingly Koppel et al. (2002) found slightly different function words to be the most significant at distinguishing gender between the fiction and non-fiction texts. Similarly, the review of Mulac et al. (2001) did not find a reliable difference between genders for use of first and second person pronouns, although many of the studies reported were conducted on quite a small scale. There is, however, clearly variability within the field. Further more, Koppel et al. found that many of the differences between male/female texts were the same as those between non-fiction/fiction.

This corresponds well with their findings that tied male language to Biber's 'informational' dimension (Argamon et al., 2003). In Biber's original study (Biber, 1988) non-fiction was indeed more informational than fiction (see section 4.1.3 for a replication of this scale).

Many studies have found effects for various parts of speech. Heylighen and Dewaele (2002) applied their F-measure (computed from relative frequencies of several parts of speech, see section 2.7.2.2 for more details) to texts of known gender and found a distinct difference: female language scores lower, preferring a more contextual style (greater use of pronouns, verbs, adverbs and interjections); men prefer a more formal style (more nouns, adjectives, prepositions and articles). They found this result was to be consistent with previous findings from socio-linguistic and psychological studies, and appears consistent with those results presented here.

In their factor analysis of the LIWC, Pennebaker and King (1999) also looked at gender, finding a strong correlation with their 'Immediacy' factor. Women were

determined to write in a more more immediate style. This not only corresponds well to the findings of Heylighen and Dewaele above, but also to those of Argamon et al. (2003): male writing is more formal, less immediate and more informational; female writing is more contextual, more immediate and more involved.

Gender language differences have also been studied in computer-mediated communication (see section 2.6 for broader CMC and language discussion). From the early popularisation of the internet, it has been held up as an anonymous medium for communication. It was originally the popular view that 'in cyberspace others only know what you choose to present about yourself' (Herring, 2000). However, studies have shown that gender is often visible on the basis of features in language that the individual may not be aware they are producing. The differences found in CMC are much the same as those found in traditional language. Herring (2000) provides a summary: men are more verbose and post longer message to discussion boards, use crude language, are more critical, and assert their opinions as fact; women were found to post shorter messages, are more likely to qualify their assertions, and, as Lakoff (1975) suggested, are more polite.

Specific work on e-mails has found again that women prefer more sociable and domestic topics, while men prefer to discuss impersonal and external matters (Colley and Todd, 2000). Judges have also been able to distinguish between genders in e-mail based on a number of features (Thomson and Murachver, 2001): women use more modal auxiliaries, intensifying adverbs, are more likely to discuss emotions and share personal information, and e-mails are more likely to contain questions, compliments, apologies and self-deprecation; men's e-mail on the other hand are more likely to contain opinions and insults.

2.4 Genre

So far this chapter has introduced the research areas of personality and gender, as well as looked at previous studies on language within these fields. This section is concerned with the remaining aim of the thesis, and concerns genre. Genre is derived from the French word for 'kind' or 'class' and genre studies date back to Aristotle. This thesis

aims to examine blogs within a larger genre space. To this end, this section first looks at how genre is defined. Following this, work within computer-mediated communication which looks at genre is reported.

2.4.1 Definitions of Genre

Everyone is familiar with the idea of genre: a name given to a group of things which are similar in some way. For example in film theory Westerns and Horror are established genres, while game shows and sitcoms exist on television, and newspaper stories, academic papers and statistical lists exist within text. However, attempts to define just what makes a genre, and what situates an entity within it are not as straightforward.

The difficulty lies in the fact that there is no scientifically measurable notion of genre; it is merely an abstract concept (Feuer, 1992). Additionally, there is no fixed set of genres, with the general public prone to create their own de facto genre labels.[3] Stam (2000) identifies four problems with genre labels (as they relate to film): *extension*, the breadth of a label or lack thereof; *normativism*, preconceived ideas of membership to a genre; *monolithic* definitions, the belief that entities only have one genre; and *biologism*, in which genres are seen to evolve with a life cycle.

Genres can be defined in a number of ways (see Swales, 1990 for further discussion). The traditional definitions are based on conventions of content and form (eg. Westerns are movies about cowboys). These definitions are easily recognisable, however entities can often exhibit conventions from multiple genres. Contemporary theories tend to to describe genres more loosely in terms of family resemblances. The more similar two entities are, the more likely they belong to the same genre. In addition to the *definitional* and *family resemblance* approaches, there is the idea of *prototypicality*, borrowed from psychology. According to this, there will be some entities which are more typical of a genre than others. Genres are therefore fuzzy, with degrees of membership.

Beyond content and form however, recent times have seen the addition of intended purpose into genre analysis (Miller, 1984; Swales 1990). A genre can be seen as a

[3]The film ‘Shaun of the Dead’ was described by its creators as a ‘zomromcom’ — a play on the romantic comedy abbreviation *romcom* since the movie heavily features zombies.

shared code between producers and interpreters of material. A creator knows what their intention is, what their content shall be, and what form it shall take. The viewer/reader understands these, and will share the conceived genre of the creator. However, purpose is also not always clear cut, and can do as much harm as good (Askehave and Swales, 2001).

2.4.2 Genres in CMC

There is a large body of work concerned with genre in literature and film theory; there are significantly fewer studies focused on computer-mediated communication (CMC). However, there is increasing interest in CMC, not least because of the ease of access to data which is already in electronic format. Weblogs are an increasingly prominent form of CMC. Before introducing weblogs properly, this section reports on studies in the identification of genres within CMC. Note that related work examining language in CMC will be reported in section 2.6.

The first argument researchers posit is that CMC cannot be treated as a single genre, much as film or literature is not. Yates and Graddol (1996) examined a number of different types of CMC and showed that they were all distinct forms of communication. In studying academic e-mail discussion lists, Gruber (2000) determined that they could also be classed as a stable genre. He would not commit to a single genre however, for he did find distinct differences between lists. Likewise in Cho's study of e-mail (1996), while messages did share linguistic features, there was also variation between messages of the same CMC type. Cho attributed this inter-individual difference to the fact that at that time there were no stable genre expectations of e-mail. Of course, in light of work already discussed here (cf. Gill, 2004) it is eminently possible these were due to individual differences of the authors.

Many novel genres have emerged from the internet (Crowston and Williams, 2000) such as personal homepages, hotlists and FAQs. A number of generalised classifications of web genres have been identified (Crowston and Williams, 2000; Shepherd and Watters, 1999): reproduced/replicated, adapted/variant and novel/spontaneous. *Reproduced* genres are traditional paper based genres that have simply been directly reproduced on the web, academic papers for example. *Adapted* genres are similar to

reproduced, only with adjustments made to take advantage of the functionality of the web. Online manuals are an example of an adapted genre, since they are in many ways similar to traditional papers manuals, but with the added benefit of direct hypertext cross referencing. *Novel* and *spontaneous* genres (also referred to as emergent) are those such as personal homepages or FAQs which have emerged from the internet with no traditional antecedent.

However, while genres are still emerging as the internet evolves, there are many webpages which cannot be classified. Recent work by Santini (2005) is aimed at automatically clustering webpages by identifying textual patterns. She also reports that Biber has started sketching a typology of web registers (Biber, 2004) following his multi-dimensional analysis approach (Biber, 1988). In a less traditional approach, Shepherd et al. (2004) used web-specific features such as number of links and the presence of javascript in order to determine home page classes. A recent survey of user-perception of web-based genres (Santini, 2006) shows that they are still evolving, with users disagreeing on the labelling of many genres.

Looking more specifically at weblogs, Herring et al. (2004a) investigated just what made them a legitimate genre. Drawing from previous research on genre analysis and particularly the work of Yates and Orlikowski (1992), they found a number of criteria by which to define a genre. They determined that comparing weblogs to other electronic genres along with more traditional ones, helped to explain their nature.

With reference to the work of Crowston and Williams (2000) Herring et al. felt that since journal blogs were related in some way to off-line, paper-based diaries, then weblogs are at least a partially reproduced genre.

Their study led them to propose that weblogs are a hybrid of existing genres, and that they are rendered unique by the combination of features from the source genres that they adapt, along with their distinctive technical affordances. They suggested a continuum of online genres, placing weblogs between standard HTML documents such as webpages and asynchronous CMC such as newsgroups.

On automatically distinguishing weblogs from non-weblogs, the task appears surprisingly easy. Elgersma and de Rijke (2006) report good classification accuracy using intuitively derived features such as number of posts and the webhost of the page, many

of which are clearly specific to blogs. There is as yet little work looking at different classes of weblogs (see section 2.5.2), though a survey of studies concerning weblogs can be found in section 2.5.4.

2.5 Weblogs

With the introduction of weblogs in the previous section, it is appropriate to discuss these in greater depth. The specific genre of interest in this thesis is personal diary weblogs. 27% of internet users in the US read weblogs, but 62% still don't know what they are. This is despite the fact that it is estimated over 8 million weblogs have been created in the US alone (Rainie, 2005).

Weblogs are an increasingly popular mode of communication in the ever changing online world, and they provide the data that supports this project. This section will introduce weblogs and give an idea of their general popularity. However, before any discussion of what weblogs are, the reasons behind their choice for this study are outlined.

2.5.1 Selection of blogs as object of this study

The work of this thesis owes a certain amount to the work of Gill who utilised a number of different approaches to study personality differences in language of e-mail. In choosing to study individual difference in language, there were a number of factors to be considered when deciding on the text genre to use:

1. Ease of attaining data

   Ideally, the data should be as easy to collect as possible. The text should be in electronic format, and the personality test results should be easily and reliably collected.

2. Volume of data required

   Gill had 105 subjects with an average of approximately 600 words per subject. At least as much again is required so as to be able to replicate his results, but

ideally there will be more subjects and more data per subject. While it is not a direct concern of this thesis, it has been shown that machine learning techniques perform best with texts of at least 1000 words (Stamatatos et al., 2000a). It was found that the majority of errors were caused by shorter texts.

3. Register of the data.

   Gill analysed personal e-mails, which tend to be in an informal register. More formal registers tend to use more constrained language to suit the situation or purpose of the text. At this point, the field of personality and language studies, particularly with regard to traits others than Neuroticism and Extraversion, still appears to be in its infancy. With this in mind, the greater room for linguistic variability within the data the better. To this end, data should ideally be as similarly informal as e-mail.

It was considered that after e-mails, diaries would be a good source of text, because these are both personal and plentiful. The difficulty came in both gaining access to personal diaries and processing the data. Blogs are electronically kept diaries. Like most computer-mediated communication (CMC) this offers the virtue of large amounts of natural language data at relatively low cost; for example transcription costs are minimal compared with spontaneous speech. This addresses consideration one above. The second is addressed by the sheer number of blogs in existence, as mentioned above and to come in the next section. Finally, the register of the data is informal since this study concerns journal weblogs rather than more impersonal information based ones (cf. sections 2.5.2.1, 2.5.2.2 and 2.5.2.3). Note that the concerns of experimenting in an online environment are discussed in the Methodological Issues section (section 2.7.3.1).

2.5.2 What is a weblog?

In 1990 Tim Berners-Lee of CERN developed HTML, and the World Wide Web was born. One of the first websites was his log of websites, keeping track of them as they first came online; it was the first weblog. Traditionally, weblogs were rather straightforward logs of the web, containing nothing more than a log of other websites,

but the ease with which they can be created has led to increasing popularity and rapid evolution.

The Routledge Encyclopedia of Narrative Theory (Herman et al., 2005) has this definition:

> A weblog, or *blog*, is a frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first.[4]

With the increase in the number of people with access to the internet, and the availability of tools for easily creating a weblog such as Blogger and LiveJournal,[5] there has been a great increase in the number of weblogs. In the year 2000, LiveJournal was seeing an average of less than 10 new diary weblogs a day at the start of the year. This had increased to 200-300 by the end of the year, and by the end of 2004 there were approximately 9000 new journals created each day. In a 2003 survey (Henning, 2003), it was predicted that over 4 million weblogs had been created up to that time, and as mentioned previously, by 2005 there were 8 million in the US alone.

There are many different kinds of weblogs ranging from personal online journals to sites that track news on specific topics. With increasing use of audio and video technology there are now sub-genres such as photoblogs.

The term *blog*, originally the shortened form of weblog, is by far the more common term,[6] and in 2004 the word was selected by US dictionary publisher Mirriam-Webster as their word of the year: the word whose definition was most requested (BBC, 2004). They define a blog as:

> A web site that contains an online personal journal with reflections, comments and often hyperlinks.

This highlights a common perception about blogs: the assumption that they feature the far more personal content of online journals rather than the more news-based weblogs. To further understand the distinction between different types of weblogs, it is perhaps best to look at the three main types of weblogs that this thesis recognises: *news*, *commentary*, and *journal*.

[4] The full definition can be read at the author's website `http://jilltxt.net/?p=227`

[5] `http://www.blogger.com` and `http://www.livejournal.com` respectively.

[6] On 7th March 2005, comparing hits on google.com, *blog* achieves 181 million compared to just 43.8 million for *weblog*.

2.5.2.1 News weblogs

As already mentioned, the very first weblog was a website that listed all the other websites as they came on-line in the early days of the internet. Many weblogs exist today to serve a similar function: they catalogue news from various sources on particular topics.

The kind of news collated can vary: it can be general political, technological or national news for example; or it can be very specific news on very specific topics, such as Wi-Fi technology, or local institution news. Figure 2.1 shows a sample taken from a popular political news weblog. Instapundit[7] is written by law professor Glenn Reynolds and is ranked as one of the most read weblogs in the world.[8]

One of the defining characteristics of news weblogs is that they are updated frequently, often several times daily. Figure 2.1 shows three posts made on the same day, all made by 10:30 am.[9] Each news story contains at least one link to the original source, and as is increasingly common, there are adverts within the page. Note also the content: as with many weblogs in September 2005, focus is very heavily on the hurricane disaster in the Southern United States.

Many sites merely post links to other websites or report on stories, but others include comments from the author. This is often all that distinguishes the many weblogs that report the same topic: the personality of the writer. They can add their thoughts to the news and act as a guide in the field.

Of course the more author comment that is included the more opinion is given and the less objective the reporting. This would lead to a weblog being categorised as *commentary*, the next category described here, rather than strictly news reporting.

2.5.2.2 Commentary weblogs

As with news weblogs, commentary weblogs tend to refer to outside material and are often just as focused, but they are not necessarily so time-pressured. Authors often

[7] `http://www.instapundit.com/`

[8] According to The Truth Laid Bear (`http://www.truthlaidbear.com/TrafficRanking.php`), which lists weblog statistics, Instapundit receives over 100,000 viewings a day.

[9] This is clear due to the reverse chronological order that posts appear in.

Figure 2.1: An example of a news blog

describe their work in varying ways: analysis, rants, musings. Figure 2.2 shows a weblog for news in the weblog community.

This weblog looks similar to a news weblog, and in many respects it is. The posts pictured both relate to the field, and the first is a matter of news. However, there are not necessarily links to sources, and the items both have a heavy personal spin: the first is the author's response to the news item, and the second is purely the author's thoughts on items which they have previously posted about. It is much easier to get a sense of author the more they write from personal opinion. A political *news* weblog is

Figure 2.2: An example of a commentary blog

more likely to be objective, than one in which the author responds and comments on the stories.

There are many technological and political weblogs, but it is also possible to find ones that discuss religious matters, review books or television, and even cover doll collecting. Of increasing popularity however, following closely behind the readership of political weblogs, are those concerned with gossip. Among the most popular of these is Gawker,[10] a Manhattan based weblog that focuses on celebrity gossip.

[10] http://www.gawker.com

2.5.2.3 Journal weblogs

Journal weblogs are simply on-line diaries, and they are the main focus of this thesis. The level of personal detail in the diaries varies, and often the author chooses to remain anonymous, but they still concern the day to day thoughts and actions of an individual. Figure 2.3 shows a fairly typical journal weblog with many of the properties expected: the text is personal in nature; previous posts are archived but still accessible; posts occur at various intervals; the third post contains a number of links which have interested the author. Note this blog uses an anonymity approach of using generic terms to refer to family members, e.g. ‘son # 2’.

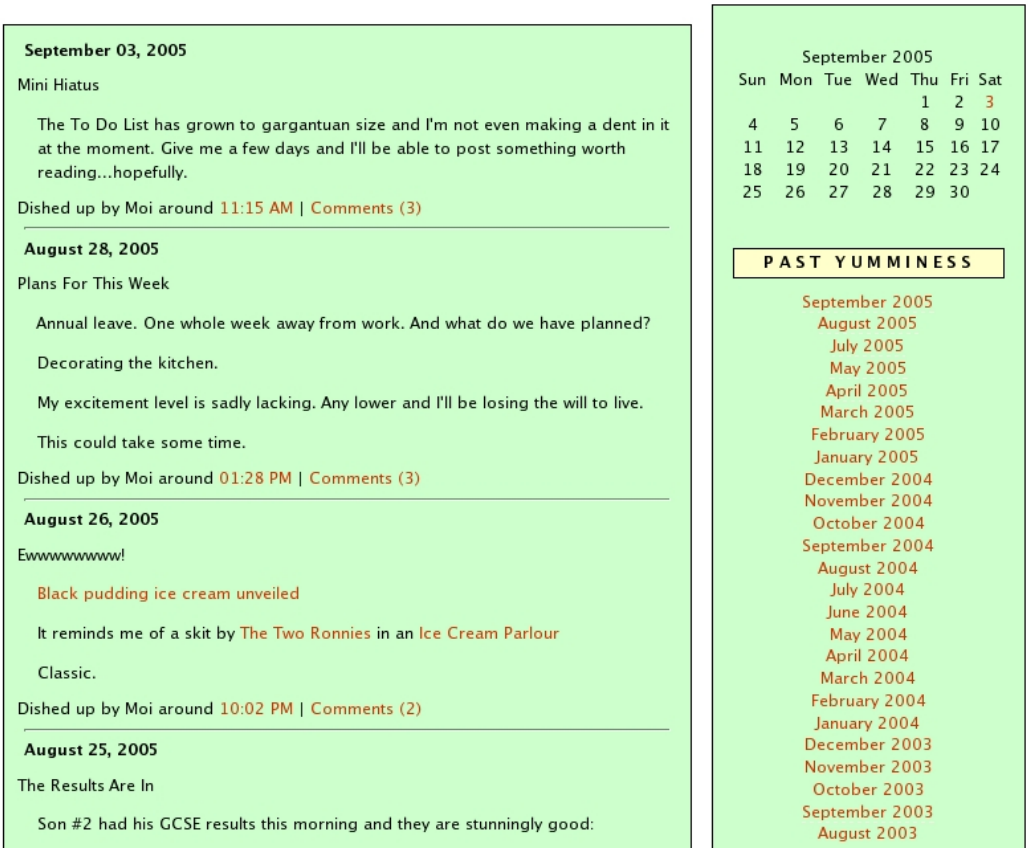

Figure 2.3: An example of a journal blog

There are weblogs that do not necessarily fit into these three categories or are a mixture of them. There are many weblogs which are nothing but random collections of links deemed *interesting* or *amusing* by the authors. The categories discussed so far were defined early in the progress of this work, but there has been work subsequently published that attempts to categorise blogs. Perhaps the most popular categories are those of Herring et al. (2004a) who say that blogs can be *filters* (because they filter the internet based on the author's interests and opinions), *knowledge logs* ('information and observations focused around an external topic') and individually authored *personal journals*. These categories map reasonably onto those above. In his study, Krishnamurthy (2002) proposed to classify blogs along two dimensions: personal versus topical, and individual versus community. However, differing approaches to classification are not an issue to be concerned with since categorisation per se is not a goal of this work. The categories are merely used to illustrate what a weblog is and highlight where journal weblogs fit in the medium.

2.5.2.4 Terminology

As discussed earlier, there are many definitions of *weblogs* and *blogs*, and there is much debate as to the differences between *journals* and *diaries*.[11] This section defines the terms as they will be used in this thesis.

- **Diary**: *A daily record of events or transactions, a journal; specifically, a daily record of matters affecting the writer personally, or which come under his personal observation.* (From the Oxford English Dictionary.[12])

- **Journal**: the same as a diary.

- **Weblog**: *a frequently updated website consisting of dated entries arranged in reverse chronological order so the most recent post appears first.* (From the Routledge Encyclopedia of Narrative Theory.)

[11] See http://www.wild-mind.net/index.php?m=200307#2 for discussion.

[12] `http://www.oed.com`

- **Blog**: this is the shortened from of weblog that is here used to describe a personal *journal* weblog as described above.

  Blog can also be used as a verb to describe the act of writing a blog, or single blog entry.

- **Blogger**: one who keeps a blog.

- **Blogging**: the act of keeping a blog, or writing a blog entry.

Henceforth, in general discussion the term *weblog* will be used, but when referring to the personal weblogs with which this work is concerned, the term *blog* will be used.

2.5.3 Trends in Weblogs

Weblogs are becoming harder to avoid, both on the internet and in the more traditional paper-based media. Technology websites report on how weblogs caused noise for search engines (Orlowski, 2003a). Newspapers introduce bloggers who pronounce how great blogging is and tell some amusing internet based anecdotes. An increasing number of celebrities are now blogging, either with commentary style blogs on issues that concern them[13] or more personal musings on their life.[14] There are companies, such as Gawker Media, which exist purely to publish weblogs covering a variety of different topics.

Weblogs are often touted as the new journalism. They are in many ways "*more spontaneous than traditional commentary*" (Weintraub, 2003), and bloggers are frequently becoming unofficial eye-witness sources to news stories (Ward, 2003). Response times are faster than traditional media, and bloggers are not constrained by formality. In the 90 minutes following the terror attacks on London in July 2005, blog-monitoring site Technorati reported that 1300 blogs had already mentioned the catastrophic events (BBC, 2005).

Two of the most significant uses of weblogs in recent times concern the American Presidential elections in 2004, and the war in Iraq. Weblogs have been used to discuss

[13] See the blog of Anita Roddick, founder of the body shop at `http://www.anitaroddick.com`.
[14] Long time blogger Wil Wheaton at `http://www.wilwheaton.net`.

the recent war from various perspectives[15] including those of the journalists who have been covering the war and even Iraqis living in the midst of the conflict. The most well known of these is Salam Pax[16] who maintained an anonymous blog from the heart of Baghdad at a very dangerous time. Though he no longer maintains his blog, it has been published as a book, and the author has moved to more traditional journalism with a column in the Guardian newspaper.

The Presidential race was also blogged from many perspectives but first hit the news when it was adopted by early candidate for the Democratic nomination Howard Dean. Dean was well known for his use of the internet as a campaigning tool, and his team maintained a weblog of the campaign.[17]

In January 2005, the biggest story in the news was the devastating tsunami which hit south-east Asia. Blogs once again hit the media (Boyd, 2004) being used to tell personal accounts of the tragedy[18] and also to coordinate relief efforts and information.[19]

However, not all weblogs attract the attention, nor readership, lavished on a select few by the media. Henning (2003) likened the situation to an iceberg:

> When you say "blog" most people think of the most popular weblogs, which are often updated multiple times a day and which by definition have tens of thousands of daily readers. [...]
>
> What is below the water line are the literally millions of blogs that are rarely pointed to by others, since they are only of interest to the family, friends, fellow students and co-workers of their teenage and 20-something bloggers. Think of them as blogs for nanoaudiences.

It is estimated that roughly a quarter of weblogs created are abandoned after just one day, with many more not lasting beyond the first year. This is perhaps a reflection of the ease with which they can now be created. There are many who feel that too much is made of the technology and that what makes them popular is actually the people who put it to good use (Orlowski, 2003b).

---

[15] `http://www.warblogs.cc` highlights the best stories from various weblogs.

[16] Where is Raed can be found at `http://dear_raed.blogspot.com`

[17] Dean's blog can be found at `http://blogforamerica.com`, though it now covers more general democratic concerns.

[18] `http://morquendi.blogspot.com/2004_12_28_morquendi_archive.html`

[19] The South-East Asia Earthquake and Tsunami blog at `http://tsunamihelp.blogspot.com/`.

Weblogs appear with increasing frequency in the mainstream media. Stories are now written about blogs and discussions about them are held in newspapers, on radio and even television.

In his study of social structure in weblogs, Marlow (2004) used Lexis Nexis to search for blog related terms in newspapers and magazines. He found that while the number of articles concerning weblogs was growing, the term was being used less frequently per article. This led him to posit that "more recent articles are likely to be influenced more by weblogs, and less about the medium itself."

One topic much discussed is the increasing dangers of blogging. There are increasing reports of bloggers losing their jobs because of their blogs (Crawford, 2005, Twist, 2005). This has created much concern in the community and with employers. There are concerns over freedom of speech and infringement of copyright, and this is leading to companies working out specific blogging policies with their employees (Campbell, 2005, Jesdanun, 2005).

Another debate concerns the legitimacy of bloggers' journalistic claims and the power that they have. When bloggers find something amiss, they have the power, resources and network to investigate thoroughly, uncover the truth and report it to the world (Anderson, 2005). However, many are beginning to fear this power is open to abuse. It can be used by political extremists to unjustifiably slur opponents and damage reputations (Rall, 2005).

2.5.4 Previous work on weblogs

It is not just the media for whom weblogs are becoming common place: academia is beginning to embrace them, both as an instrument and object of study (Mortensen and Walker, 2002, Rosenbloom, 2004). Just as weblogs are being used to discuss developments in commercial fields, academics are increasingly using blogs to discuss their work.[20] It is felt that both senior academics and students can use blogs as an informal outlet for their ideas, which will widen the potential audience from which they could receive feedback. In fact Warwick University in the UK is actively encouraging blogging, becoming arguably the largest academic blogging project in the world

[20]This project is blogged at `http://blogademia.blogspot.com`.

(Adenekan, 2005).

One of the larger bodies of work arises from interest in the community-based nature of many weblogs - knowledge sharing (Rittenbruch et al., 2003, Cubranic et al., 2003, Efimova et al., 2004). Perhaps the most obvious application of this is in the field of education (Efimova and Fiedler, 2004, Anderson, 2004).

Knowledge sharing requires a social network around which to build a community, but recent studies have also shown blogging to be a social activity (Marlow, 2004; Efimova and de Moor, 2005). Nardi et al. (2004) carried out an ethnographic study of blogging and found that unlike traditional diaries, blogs are

> ...a studied minuet between blogger and audience. Bloggers consider audience attention, feedback and feelings as they write [...] consciousness of audience is central to the blogging experience.

Work on weblogs has exploded in popularity in recent times. This is evidenced by the level of attendance at the recent AAAI spring symposium on Computational Approaches to Analysing Weblogs: almost one third of attendees to the 8 symposia were 'blogademics.'[21] It is becoming increasingly clear that such work can fall into two camps: those who use weblogs as the base genre for work in a specific area, such as this study; and those who study weblogs for their own sake.

Work in the former group includes studies attempting to automatically identify bloggers by age and gender (Burger and Henderson, 2006; Schler et al., 2006) and a vast body of mood/sentiment/opinion analysis work (Mishne, 2005; Mihalcea and Liu, 2006; Tong and Snuffin, 2006). The latter group consists, for example, of those interested in story propagation through weblogs (Lloyd et al., 2006; Wu and Tseng, 2006) or novel applications such as friend recommendation (Hsu et al., 2006).

2.6 CMC and Language

In section 2.4 weblogs were briefly discussed as a genre. One of the aims of this thesis is to explore the linguist nature of the blog genre. So far this chapter has reported on studies looking at language relating to personality and gender. While this touched upon

[21] The term used to refer to academics studying weblogs.

email (Gill, 2005) and internet discussion (Herring, 2000), there has also been work looking at language in CMC more generally. This is reported here before discussing the language of weblogs.

CMC is traditionally a written medium. It does, however, approach becoming spoken to varying degrees. Static webpages might be purely written, but instant messengers create conversations as close to a spoken form as is perhaps possible for the written word. Yates (1996) found e-mail communication to display properties which precluded it from being categorised strictly as either written or spoken language. E-mail has been found to be a written form since interlocutors are physically separated, it is durable, and authors often use complex linguistic constructions; however, e-mail is often unedited, uses first- and second-person pronouns, present tense and contractions, and is generally informal (Bälter, 1998; Baron, 2001). Gruber (2000) found that scholarly e-mail discussion lists had properties in common with both oral communication and academic letter writing.

Colley and Todd (2002) referred to a number stylistic features not often seen outside of e-mail. These include trailing dots, capitalisation, and excessive use of exclamation and question marks. Studying a corpus of postings to a bulletin board, Collot and Belmore (1996) found that the language was most like that of '*public interviews and letters, personal as well as professional.*'

As highlighted above, many studies of personality and language have been carried out on spoken text. Computer-mediated communication, like most writing, is less rich than face-to-face communication (Panteli, 2002). CMC is not strictly a written medium however, and information is communicated by alternative means. Werry (1996) points out that in internet relay chat (IRC), linguistic strategies have been adopted to replace the missing intonational and paralinguistic cues of face-to-face discourse. This finding is reflected in the use of coordination devices in Hancock and Dunham's (2001) study of computer-mediated task-based interactions.

Despite the lacks of cues offered by CMC, interactions can still provide information on the interlocutors. In a study of text-based communication within organisations, Panteli (2002) found social cues suggesting status.

An interesting observation is that most studies of CMC, be they synchronous or

asynchronous, concern interaction. Discussion groups, e-mail and internet chat all concern communication between two or more individuals; they are forms of dialogue. Despite the findings that certain types of weblog contribute to social networks (cf. Efimova & de Moor, 2005, discussed in the next section), personal blogs are by-and-large monologues. This fact distinguishes this current work from those that come before, and provides an argument for the uniqueness of blogs as a genre for study.

2.6.1 Language and Weblogs

Perhaps the most relevant work to this thesis is that which looks specifically at language use in weblogs. Huffaker (2004) studied gender difference in weblogs kept by teenagers. He found very little surface difference (word count, word length), but males tended to use more 'active' language. Cohn, Mehl and Pennebaker (2004) used the LIWC tool to investigate changes in language surrounding the events of September 11, 2001. They found that in the short term authors expressed more negative emotions, were more cognitively and socially engaged, and wrote with greater psychological distance. Over time, these features slowly returned to their baseline levels. Though not studying language explicitly, Krishnamurthy (2002) also found that following 9/11, the number of daily posts to Metafilter, a community news weblog, increased while the number of links decreased.

Language in weblogs is also studied for more commercial purposes. BlogPulse[22] (Glance et al., 2004) is a tool for data mining weblogs. It is used to discover key phrases and names being talked about in the world of weblogs.

Nilsson (2003) looked at the language used in a community of researchers' weblogs. She found that there was a much higher use of *in-group* terms (I, me, my, we, us and our) than *out-group* term (they, them and their), which is to be expected of the personal nature of the texts. She also found posts to be written in 'short, paratactic sentences' employing 'informal, non-standard constructions and slang.' A further linguistic feature she identified as common to blogs was the use of frames (as per Brown and Yule, 1983), which allows the author to assume that their audience has background knowledge in the concepts they discuss.

[22] http://www.blogpulse.com

Previous findings suggest that language in CMC displays many of the properties of both spoken and written language. The same can be said specifically of weblogs. Not only does this give a potential insight into the language this study will work with, but it gives a perspective on the situation of weblogs as a genre (as discussed in section 2.4.2).

Herring et al. (2005), in their analysis of weblogs as a 'bridging genre' describe weblogs as lying on a continuum between standard HTML documents, and asynchronous CMC such as newsgroups. Indeed, during their research on blog search technology, Glance et al. (2004, p6) noted that:

> If we believe the metaphor that blogging is like publishing while posting [to news groups] is more like chatting, it's not surprising that weblog entries tend to be more polished pieces of writing, with fewer grammatical errors and tighter diction.

So weblogs are seen as closer to written language than the more conversational newsgroups. However, not all weblogs are one sided, as most written forms can be. Work on the social networks that bloggers can form has shown that weblogs can take the form of both monologue and dialogue (Efimova & de Moor, 2005). The increasing use of commenting technology allows readers to leave feedback on what they have read. Bloggers, if they so choose can respond to this on their blog. Blogs can therefore be thought of as *simultaneously self-reflective thoughts presented publicly, and continuous conversations* (Nilsson, 2003; page 31).

Crystal (2001) claims that online language is neither written nor spoken, but is multifaceted and has aspects of both writing and speech. This claim prompted Nilsson to ask if CMC is of a different nature to less mediated forms of communication or if it can be explained by differences in genre and activity. In answering this, Nilsson applied Crystal's list of differences between written and spoken language to blogs.

- **Boundedness and Dynamicity** Blogs, like writing, are space bound because text is bound to the space it occupies and because there is an accepted range for post lengths. The dynamicity is governed by the nature of blogs to consist of a front page with the most recent posts and an archive containing all those that

came before. Whilst posts can be edited, those that fall into archives rarely are, they remain static. New posts are added all the time however, the front page changing all the time. This makes blogs time bound like speech, since whilst posts will always be present in some form, they are limited in how long they will exist in the main context of the blog.

- **Synchronicity** The time lag between creation and reading of a blog post can be great, which put blogs far from the synchronous nature of conversation. However, traditional published material is well thought out and carefully edited, while the nature of blogging promotes immediacy, and posts tend to resemble quickly jotted notes. The time-bounded nature of posts also encourages quick responses from readers if a commenting function is present.

- **Paralinguistic Cues** There can be no more extralinguistic cues than in face-to-face conversation. Blogs are increasingly becoming multimedia. While including pictures with text may be no different from say newspaper writing, audioblogs (recorded spoken journal entries) are becoming increasingly common. Traditional writing cannot rely upon context to make meaning clear, but the use of hypertext links allows bloggers to add context to their text. Also the less formal nature of writing in blogs allows for attempts to replicate speech nuances with the use of emoticons (e.g. the wink smilie — ;-) — to suggest the author is not being serious).

- **Constructions** There are constructions which are very characteristic of both speech (informal slang) and writing (complicated legal terms or chemical names). Blogs employ constructions from both forms of language: their written nature allows the discussion of long not often spoken terms; their informality allows use of contractions, nonsense words, and abbreviations like LOL (laughing out loud).

- **Communicative Functions** Speech suits social functions; it expresses social relationships, and it can be used for opinions and attitudes. Writing suits the recording of facts and the communication of ideas. Social networks can easily

form around blogs (Marlow, 2004), but their textual basis makes them equally suitable for recording thoughts and ideas.

- **Ability to be Revised** The written word can be revised as much as required before a reader ever sees it, but speech can only be revised after it has been heard. Blogs can be revised as often as possible, both before and after they are published, but it is not common to do so. Even if a post is deleted after it is published, it can still be found on internet archives.

- **Unique Communicative Features** Prosody is a unique feature of speech that cannot be written down efficiently. Writing however has formatting and structural organisation that cannot be applied to speech. That they can form social networks (Marlow, 2004) is generally considered to be an important affordance of blogs. The posting order allows readers to read the latest news first, without re-reading information they already know.

By looking at the aspects of spoken and written language above, Nilsson showed that the language of blogs has much in common with both. Blogs do not seem to fall completely under either medium, written or spoken, but instead fall somewhere in between. This echoes the work described in the previous section looking at language in CMC in general, and provides an argument for the representativeness of blogs as a genre for study.

2.7 Approaches to Linguistic Analysis

This section provides background on the analysis techniques that will be used for investigations in this thesis. Introduced first are those methods based on external dictionaries: approaches where the features to be examined are pre-defined. Secondly, more data-driven methods are discussed; where features of interest are derived purely from the raw data at hand.

2.7.1 Top-down approaches

Top-down, or dictionary-based approaches are defined as those providing external feature sets with which to examine data. The two examples of this to be discussed are the Linguistic Inquiry and Word Count (LIWC; Pennebaker & Francis, 1999) and the MRC Psycholinguistic Database (Coltheart, 1981; Wilson, 1987).

2.7.1.1 LIWC

The Linguistic Inquiry and Word Count (LIWC; Pennebaker & Francis, 1999) is a dictionary based approach to content analysis. The LIWC is a text analysis program which was originally designed to investigate the relationship between disclosure and language use features and health and well being (Pennebaker, 1997; Pennebaker et al., 1997; Graybeal et al., 2002). However, this method has since been applied to investigate linguistic behaviours in a variety of genres, including perceived character changes in politicians (Pennebaker & Lay, 2002) and reactions to a traumatic event (Cohn et al., 2004). This method is adopted here since it has been used previously to investigate individual differences (Pennebaker & King, 1999; Gill, 2004).

LIWC is essentially a word count approach, searching for all word contained in any of its dictionaries. The output expresses the frequency of each category as a percentage of the whole text. There are also some purely statistical measures: words count and words per sentence are raw counts; use of words greater than six letters and sentences ending with question marks are percentages.

The dictionary categories are divided into four main categories: Standard linguistic dimensions, Psychological processes, Relativity and Personal concerns. The first category contains those statistical categories discussed above along with basic linguistic categories such as Pronouns (broken into First-, Second- and Third-person categories), Articles, Numbers and Negations. Note that while Pronouns for example of part-of-speech categories they are derived purely from the existing dictionary of pronouns, and not by tagging of any kind. The remaining three categories are largely concerned with traditional content and analysis concepts, derived from theoretical sources. They are are further subdivided into groups of dictionaries: Affective or emotional processes (eg. Positive feelings, Anger), Cognitive Processes (eg. Certainty, Discrepancy), Sen-

sory and perceptual processes (Seeing, Hearing and Feeling), Social processes (eg. Communication, Family); Time (Past, Present or Future tense verbs), Space (eg. Up, Inclusive), Motion; Occupation (School, Job or work and Achievement), Leisure activity (eg. Sports, Music), Money and financial issues, Metaphysical issues (Religion and Death & dying), Physical states and functions (eg. Eating, drinking, dieting, Grooming).[23]

The product of this is that the LIWC contains around 70 dictionary categories, containing over 2000 word or word stems. Pennebaker and colleagues distinguish the LIWC from other text analysis programs by pointing out that both the dictionaries and the analysis approach have been independently rated and validated by judges (Pennebaker & Francis, 1999; Pennebaker & King, 1999).

2.7.1.2 MRC

The MRC Psycholinguistic Database (MRC) is a machine readable resource containing psycholinguistic information on a large number of words (Coltheart, 1981; Wilson, 1987). However, following Gill, it has been adapted to become a form of content analysis which measures the psycholinguistic properties of texts.

The categories for which texts are scored includes: Number of Letters, Number of Phonemes, Number of Syllables, Kucera and Francis Frequencies (includes written with category and sample counts), Thorndike and Lorge Frequency, Brown Verbal Frequency, Familiarity, Concreteness, Imagability, Meaningfullness, Age of Acquisition and Dolby's word status categories (e.g. Standard, Archaic, Poetic, Dialect) as derived from the Shorter Oxford English Dictionary in 1963. For each of these categories both the mean and standard deviation are calculated, and for many the percentage and number captured by the dictionaries. In additional to the psycholinguistic data, the program is also designed to calculate statistical measure such as the number and percentage of words captured by the database, the total number of strings in the text, and the number and percentage of groups of numbers and non-alphanumeric characters.

The dictionary lookup approach is similar to that of the LIWC, but beyond the dif-

[23]There is an additional 'experimental' dimension consisting of Swear words, Nonfluencies and Fillers.

ference in focus, they differ in further ways. Firstly, the LIWC consists of pre-defined dictionaries based on human judgement of linguistic terms, while the MRC is built upon empirically derived data collated from several studies. Secondly, the MRC also includes part-of-speech information. This allows for disambiguation of word senses which in turn results in more accurate data processing. Finally, the resources differ greatly in size and therefore linguistic coverage: the MRC contains around 150,000 words, with psycholinguistic information for around 40,000; the LIWC on the other hand contains just 2000 word stems.

2.7.2 Bottom-up approaches

Bottom-up or data-driven approaches are characterised by their reliance upon the data to provide the theory, rather than using only specific features, which impose their own set of theories. Working with raw data, manipulating it as required, does leave room for over-fitting to prove theories. However, with clear transparent decisions made in the process, results are less arguable, and methodologies are easily reproduced.

2.7.2.1 Collocation

Collocation, as it is used here, relates to combinations of two or more linguistic units together in sequence. Units can be words, parts-of-speech tags or punctuation markers for example. This is a more general use of the term than is often adopted. What is referred to here as *collocation* is often termed *co-occurrence.* Co-occurrence is normally more specifically sub-categorised into *collocation*, concerned with words only, and *colligation*, more grammatically oriented. Since this thesis is concerned with language patterns generally, rather than solely identifying key words, the term is used in the more general sense.

A related area is that of concordancing, which is the viewing of a target word in the context of its occurrence. They are both approaches aimed at reflecting more context of language than individual words can allow. The study of word collocations, or n-grams (sequences of *n* length, although typically 2 or 3), has been used to identify domain-specific vocabulary (Damerau, 1993) and differences between native and non-native speakers (Milton, 1998). Collocation is of course not limited to words alone: it

is possible to apply such contextual approaches to grammatical tags such as parts-of-speech (cf. Koppel et al., 2002; Argamon et al., 2003).

It is also possible to test the statistical significance of pairings of words with collocations. For example, 'Mary Poppins' will be a significant collocation since 'Poppins' will most likely not appear in that position in many other bigrams; conversely 'and the' is not statistically significant since both words will appear in many other contexts. As has been the case for the analogous problem of corpus frequency comparison (Kilgarriff, 2001), there has been much debate as to a suitable measure for determining the patterning of collocations. In smaller samples, the G^2 statistic is regarded as better approximating the χ^2 distribution than the X^2 statistic (Dunning, 1993). However this approximation may be violated in sparse n-gram data (Pedersen, 1996), and so Pedersen et al. (1996) proposed the the use of Fisher's exact test. In an evaluation of statistical tests, Daille (1995) found that the overall 'best statistical model—that is to say, the one which gives a correct list of terms with the lowest rates of noise and silence—turns out to be one based on likelihood ratio—in which frequency is taken into account.'

There are further considerations, such as which words or features should be included or excluded from an analysis, whether upper or lower limits should be put on n-gram frequency, and how long the n-gram should be. A description of n-gram calculation software, and the options involved can be found in Banerjee and Pedersen (2003).

2.7.2.2 Contextuality of language

Previous work on the relationship between parts-of-speech and personality was explored in section 2.2.1.2. This section goes into more detail of the F-measure, Heylighen and Dewaele's measure of contextuality/formality (2002). Furnham (1990) originally proposed the following description of Extravert language: it is less formal; has a less restricted code; it uses vocabulary more loosely, and uses more verbs, adverbs and pronouns. By using factor analysis of syntactic tokens produced by L2 speakers, Dewaele and Furnham (2000) subsequently described *implicit* language as a preference for pronouns, adverbs and verbs, while *explicit* language involves a prefer-

ence for nouns, modifiers and prepositions.

Heylighen and Dewaele (2002) explored the notion of implicitness in greater detail and developed a measure of a text's relative contextuality (implicitness), as opposed to its formality (explicitness). Briefly, they considered a notion of *deixis* following Levelt (1989), and identified a group of expressions that must be anchored to some part of the spatio-temporal context of an utterance in order to be properly interpreted. That is, further information is required in order to disambiguate to what it refers: for example pronouns ('she', 'there' or 'it') refer to people, places and objects that are not made clear from the expressions alone. Context is required to interpret their meaning; 'she got it there' is easier to interpret if what precedes it is known, 'Mrs Smith wore her blue hat to shop at Harrods; she got it there'.

Greater use of these expressions leads to higher levels of *contextuality*, while greater use of non-deictic expressions leads to higher *formality*. Heylighen and Dewaele proposed that certain parts of speech (such as verbs) are generally (although not invariably) deictic in nature, while others (such as nouns) are generally non-deictic. They defined the F-measure as a single measure of a text's contextuality versus formality: a low score indicates contextuality, represented by a greater relative use of pronouns, verbs, adverbs, and interjections; a higher score indicates formality, represented by greater relative use of nouns, adjectives, prepositions and articles. F is defined as follows:

$$\begin{aligned} F = \quad & 0.5 * [(noun\,frq + adj\,frq + prep\,frq + art\,frq) \\ & -(pron\,frq + verb\,frq + adv\,frq + int\,frq) + 100)] \end{aligned}$$

Heylighen and Dewaele validated their measure via factor analysis of part-of-speech data and found that over 50% of the variance was accounted for by a factor very similar to the definition of the F-measure. They used it to explore corpus data derived from Dutch, Italian, and English sources. The results in all languages were consistent: written language scored higher than spoken language, implying the former to be more formal; newspapers were found to be more formal than works of fiction; interview data was more formal than casual conversation.

At this point it is worth clearing up confusion regarding the term ‘formal’ as it is used here. The word ‘formal’ is traditionally used in opposition to ‘informal’. It is certainly attractive to describe high F-scores as relating to ‘formality’, when written genres such as newspapers score highly. However, whilst it may therefore be intuitive to say that spoken genres are less formal — more *in*formal — this is not the result of the F-score. A lower F-score only implies greater contextuality. From discussions with Jean-Marc Dewaele it is clear that this ambiguity causes a problem when interpreting results and understanding subsequent analysis. Previous findings have been reported using the original context[24] in which they were found. However, from this point forward as the F-measure is used in this thesis it shall be considered solely a measure of contextuality — high and low.

In this thesis, the F-measure has a number of applications: it will first be used to place blogs in a larger context of other textual genres (see chapter 4); it will then be used to investigate individual differences in both personality and gender. There are of course other factors which could be used to distinguish between genres. Following an extensive factor analysis of 67 linguistic features, Biber (1988) found a number of significant factors. One of these factors, known as ‘involved versus informational production’ concerned amongst others, most of the variables in the F-measure. Loewerse, McCarthy, McNamara and Graesser (2004) followed Biber, repeating his analysis with a new set of 236 language and cohesion features, including a number of LSA-based metrics. They also found several factors that readily highlight differences in genres.

However, it is Heylighen and Dewaele’s F-measure which has been used specifically to investigate individual differences between writers *within* a genre. Therefore, it is the measure adopted for this thesis.

2.7.3 Methodological issues

There are a number of concerns with aspects of the above approaches that are worth outlining.

[24] An equally tricky though ultimately less ambiguous word to use when discussing the F-measure.

2.7.3.1 Concerns in an on-line environment

There are many concerns to be addressed when conducting an experiment online: the validity of the personality measure, the representativeness of the sample population, and ownership of the data.

The simplest point to address is the ownership of the data. In her study of weblogs, Nilsson (2003) discussed ethics in the context of ownership of weblog text. The main problem researchers face regards the nature of blogs: do they belong in the private or public domain? One fear is that researchers do not need to identify themselves or their intentions when collecting data. This must be a great temptation when there is so much weblog data freely available. The American Association for the Advancement of Science (AAAS) drew up guidelines to aid researchers in distinguishing which resources could be collected and which should be left alone. These guidelines are well suited to general research of this kind, but they do not apply to this specific project. The approach of this work, as described above, required complicitness on the part of the weblog author. By completing the questionnaire and taking part in the experiment, they were volunteering their data and giving explicit permission for it to be used. Anonymity was assured, and no data is made publicly available.

With regards to sampling, Gosling et al. (2004) addressed a number of preconceptions concerning internet questionnaires, one of which was that *internet samples are not demographically diverse*. Their conclusions were mixed: while they found internet samples to be more diverse than traditional samples in some senses (like gender) they were still not completely representative of the population. For example, people of lower socioeconomic status may not have access to the internet as readily as others. Still, many laboratory studies rely on samples purely drawn from student populations, with 85% of the traditional samples compared drawing on students as their subjects.

This study was aimed at a very specific population, that of bloggers. With regards to any matters affected by use or not of the internet, blogging by its very nature requires subjects to have ready internet access. Therefore, the presentation of the questionnaire on-line should result in few or no sampling effects.

Buchanan and Reips (2001) argued that the specific technology used in internet questionnaires, both by the experimenter and the subject, could also affect the sam-

pling. Certain higher levels of technology used on a website may exclude potential subjects. One could argue that this becomes less relevant as technology becomes more pervasive, supposing of course that experimenters stick to simple approaches. The questionnaire for this study was constructed using nothing more complicated than simple Javascript, on which most blogging software relies. Therefore, the use of technology should also not affect sampling.

There can however be technological biases within the sample. Buchanan and Reips found that those participants who had used Apple computers scored significantly higher on Openness than those who used Windows-based machines. This effect was not considered in this study. There was no question built in to highlight it. However, as personality scores show (see section 3.4.3), blogging has its own relation to Openness.

Another potential bias is that of self selection (Buchanan and Smith, 1999). Voluntary online studies require that the subject be motivated enough to both take part in and complete the experiment. Web experiments are subject to relatively high dropout rates (Musch and Reips, 2000; Reips 2000), and it is easy to imagine how these effects relate to personality. Are more Open individuals more prepared to undertake an online personality test? Are more Conscientious individuals more likely to finish a test? These issues mean that without large comparison samples, it will be difficult to draw conclusions from the personality score distributions in which a test results (section 3.4.3 will discuss this in respect to subjects of this study).

Finally, the validity of the Inventory chosen for this study is considered. There is concern that traditional instruments, though simple enough to encode for online presentation, may not in that form validly assess the constructs they were originally designed to assess. To address this, the 50 item IPIP representation of the Five Factor Model was encoded so as to be administered online (Buchanan et al., 2005). After analysing the results it was found that some items no longer loaded on the expected factors as strongly as they did previously, or in fact loaded upon others more strongly. This allowed the number of items to be reduced to 41.

These revised scales were also implemented in an online test and more results were gathered. The reduced scales proved to be as internally consistent as both the online and traditional implementation of the original scales. Factor scores were also

correlated with reports of behavioural acts, and comparable results were found across measures (Buchanan et al., 2005).

2.7.3.2 Dictionary-based approaches

Section 2.7.1 described the Dictionary-based approaches to be adopted in this thesis following Gill (2004). This section reviews his concerns with such approaches (cf. Gill, 2004; Oberlander & Gill 2005). The first concern with content analysis is simply the lack of coverage, the limiting nature of the content of the dictionaries. This is particularly an issue for the LIWC because its dictionaries have been selected for their psychological relevance, and therefore may not generalise well across genres or topics.

Secondly, Ball (1994) noted that 'recall' is a problem for approaches like this, reflecting the technique's success in identifying and counting features. This is also a problem very much particular to the LIWC, due to its size. Despite including word stems to broaden potential matches, there are still only around 2000 words, compared to the 40,000 of the MRC.

The final limitation Gill highlights reflects the lack of context of a word. In relation to the LIWC this flaw is directly acknowledged by Pennebaker and King (1999). When a word's context is not taken into account, content analysis cannot say *how* it is used, merely that it is used. Hazards includes 'context, irony, sarcasm, or [...] multiple meanings' [p1297]. Disambiguation of word senses is less of a problem for the MRC due to its use of indexing by part-of-speech, but there is no further context.

2.7.3.3 Parts-of-speech taggers

A number of the approaches outlined in the previous sections require that files be tagged for parts-of-speech. Tagging is different from parsing, since categories are assigned at word-level with no concern for clause- or sentence-level. There are very many different approaches to part-of-speech tagging, most of which are generally regarded as having at least 95% accuracy. However, this can vary according to individual features of the corpus and tagger (Manning & Schütze, 1999).

One problem that taggers face is when the training and test data is of a completely different nature. A tagger trained on the Wall Street Journal will do well tagging the

Financial Times, but poorly on speech transcription for example.

Errors with taggers tend to be systematic; that is to say there is not one specific error that the tagger is most likely to make, such as modal verbs always tagged incorrectly as prepositions. Therefore despite specific word level concerns, the overall performance is suitable for most analyses.

2.8 Summary

This chapter has surveyed the fields upon which this thesis draws. The chapter opened with a brief introduction to personality theories and explained the choice of the five-factor model as the model of personality for this thesis. It then discussed facets of the five-factors and suggested how these might relate to language. It then reviewed previous work in personality and language. There are many aspects of language which have been investigated, but very few studies have looked beyond Extraversion and, to a lesser extent, Neuroticism. This thesis is also concerned with gender and language, so work in this field was also reviewed. Firstly, various approaches to gender difference were discussed, before reporting results from those general studies which treat men and women as whole homogeneous groups.

Genre was then introduced, first with a brief discussion attempting to define genre, followed by work looking at genres within CMC. Weblogs are generally considered an identifiable genre. After an in depth explanation of weblogs, and a review of their place both in society and academic study, there was a review of language and CMC. CMC in general, and weblogs in particular seem to be neither exactly like written or spoken text, but reflect aspects of both.

An introduction was then given to the methodologies and specific tools to be used in this thesis. This was briefly followed by a discussion of the methodological issues that might arise from adopting these approaches.

There has been much work in the fields of study of this thesis. This will be used to both inform the research, and form the basis of hypotheses at each stage of analysis. It is hoped that by exploring these specific hypotheses, the general questions of the thesis can be answered.

Chapter 3

Collection, Preparation and Profile of Data

This thesis investigates the linguistic properties of personal weblogs, or blogs, paying particular attention to relationships with individual differences. Blogs were chosen for the reasons outlined in section 2.5.1. There are two main aspects to the data required for this study: data pertaining to individual differences such as age and personality, and samples of blog text. This chapter begins by looking at the methodology used for collecting and preparing the data required. It then discusses some basic data analyses such as mean age and the distribution of personality scores.

3.1 Data collection method

3.1.1 Materials

The collection of the sociobiographic and blogging habit data was conducted on-line, via the author's departmental webpage using an HTML form which subjects filled in and submitted. Blog texts were collected by asking the subjects to e-mail the author with the data. Failing that, it was collected directly from the URL they provided.

Subjects were initially directed to an introductory page explaining what data was being collected and the reasons behind the study. This informed subjects how long the form should take to fill in, stated that all responses would be treated confidentially, and

provided contact details should they have any questions.

The questionnaire itself was divided into three sections and contained instructions relating to each section: basic demographic data, information on blogging habits, and personality information. Upon completing the questionnaire the subjects were taken to a further webpage which thanked them for their cooperation and gave them details for submitting their blog data.

Preliminary versions of the materials were piloted to evaluate ease of use, coverage of questions and to identify bugs in the coding.

3.1.2 Participants

Seventy one subjects met the requirements for further analysis (see section 3.2 for selection details), of which there were 24 males and 47 females. The mean age of subjects was 28.4 with a range of 15-50. All participants were native speakers of English.

A sociobiographic questionnaire and an on-line implementation of an IPIP Five Factor Personality Inventory (Buchanan, 2001) were administered to give information about subjects' background and scores on the personality dimensions of Neuroticism (mean 22.4, SD 6.3), Extraversion (mean 30.5, SD 6.5), Openness (mean 29.3, SD 4.7), Agreeableness (mean 26.3, SD 3.7) and Conscientiousness (mean 31.8, SD 6.1). Optional questions regarding blogging habits were also asked.

Each subject also provided all the text for their blog from one whole month. This provided a corpus of 71 blogs, consisting of 1854 individual posts (mean 26.1, SD 20.3) and 411843 words (mean 5801, SD 5829).

3.1.3 Collection of sociobiographic information

The first part of the form concerned basic demographic data. Most importantly, since the study required native English speakers, the first question required subjects to check a box confirming the they were 'a Native Speaker of English.' The rest of the section title *Questions about you* comprised questions on: 'E-mail' and a check box if the subject wished to hear of the results of the study; 'Name'; 'Age'; 'Gender'; 'Nationality';

'Place of Birth'; 'Place where you grew up'; 'Place where you live now'; 'Level of Education'; 'Occupation'; 'What technology do you have access to'. There was also a question relating to how the subject found the experiment (see section 3.1.5.1 for more details).

The second section of the form contained *Questions about your blog* and the questions were optional, with the exception of the name and URL of the subject's blog. The form asked for details of: 'How frequently do you blog' (a choice of *rarely, occasionally, frequently* or *all the time*); 'How often do you blog' (a more quantifiable choice of *more than once a day, at least once a day, every day, a few times a week, once a week* or *less often*); 'Who do you write for, who is your intended audience'; 'Where do you blog from'; 'When do you blog'; 'Which of these do you use regularly' asking about the use of links, quotes, pictures, and other web-based content; 'What is your definition of a blog' and 'Why do you blog', two questions that provided text boxes for the subjects to write their thoughts on the two matters; 'How revealing are you' and 'How realistic/honest are you' two questions aimed at determining how much of their lives the subjects felt they included in their blogs, and whether or not they felt they told the truth in their blogs.

The third and final part of the questionnaire asked for *More about you*, and contained an on-line implementation of an IPIP Five Factor Personality Inventory (Buchanan, 2001). This was chosen as it addressed many of the concerns raised in the section 2.7.3.1 regarding online assessments: it is relatively short and should not adversely effect dropout rates; it was designed for use on the internet and has been shown to be a valid measure. Since the inventory was designed with this use in mind, it could be placed within the questionnaire exactly as it was designed. It consists of 41 items, each with a 5-point rating scale, which subjects are instructed to use to describe how accurately each statement describes them. The items of this inventory, along with the factors and directions they load can be found in the appendix (table A.1).

Note that this thesis is conducted with an inventory and personality model reflecting five-factors, when previous work on e-mails (Gill, 2004) used a three-factor model. The reasons for this are outlined in section 2.1.1

3.1.4 Collection of linguistic data

As described in section 2.5.1 one reason for using weblogs was the volume of data. To take advantage of this, rather than ask for a number of posts from each subject, they were asked to provide a month's worth of their blog. They were specifically asked to submit all their text from May 2003. This was done for 3 reasons:

1. The assumption was made that a random month would be as representative of the subjects' blogging habits as any other. If subjects were given a choice, they may choose an unrepresentative month, such as the one they wrote the most in, their happiest month, or one in which they liked the writing. It was not appropriate to allow the choice of month to lie with the subject.

2. Putting all subjects within the same time frame, means they all experience the same world context. On a fine grained day-to-day personal level, their lives may differ. However, with respect to television, movies and news, topics frequently discussed in blogs, individuals *can* all discuss the same subjects. This allows analysis to be less constrained by topic.

3. Since the experiment was first 'launched' in June, May was the closest complete month prior to that date, allowing the sociobiographic information to be most relevant.

Since subjects can maintain their blog in a variety of different ways, either by themselves or via third party providers, there was no easy way for submission of blog data. Therefore, on completion of the questionnaire, subjects were requested to e-mail the author with the appropriate data. Instructions were included on the page as to some of the most common methods for completing this task. It was stated that the HTML version of the data was preferred but the plain text would suffice. The HTML version was preferred because of the elements it encodes, such as weblinks and use of images that could be studied in relation to personal differences in the genre, alongside the linguistic data. If subjects failed to send their data after they had submitted, it was collected from the URL they had provided.

3.1.5 Procedure

3.1.5.1 Recruitment of subjects

The first stage of recruitment was to define criteria to which potential subjects could compare themselves. It was decided that bloggers must:

- be the only author of a personal weblog.
- be native speakers of English.
- have written in May 2003.

The next stage was to attract subjects, to advertise the experiment. There are a number of ways of attracting subject to an online experiment:

- Newsgroups. When one intends to attract subjects from a very specific community, there will often be a newsgroup or online bulletin board on which an advert can be placed. From an extensive search of the internet at the time of the study, there did not appear to be a newsgroup for bloggers.
- Direct contact. Again, for a specific community, there may be mailing lists that one can e-mail with an advertisement. There does not appear to be any such list for bloggers. However, most bloggers put their e-mail addresses somewhere on their blog page, so they are easily contactable, albeit individually.
- Search engines. There are various techniques one can use, such as using meta-tags, to make a website more easily found by search engines. In this instance it did not seem worth the effort given the specific subject group required and the perceived likelihood that they would be searching for experiments to do.
- Word of mouth. As with face-to-face experiments, researchers can often rely on the fact that subjects will tell their friends. It is safe to say that bloggers do not exclusively make friends only with other bloggers, but there are more positive arguments for this approach. It is clear from observing blogs to bloggers read and link to other blogs. Even if they don't explicitly make reference to it, they still often maintain a list of blogs they read.[1] Bloggers read other blogs.

[1] Often referred to as a blogroll.

> Theoretically, if a blogger posted a link to the experiment then other bloggers would read about it, and so word should spread. There is of course no guarantee that a blogger will post about it, but since blogs are about daily events, some bloggers may well consider doing an experiment about blogs interesting enough to write about.

Two approaches were taken to the recruitment of subjects: direct mailing and word-of-mouth. For direct mailing a number of bloggers were found through search engines, blog hosting sites, and via links. They were chosen if they matched the criterion above. The word-of-mouth approach was implemented by asking bloggers who completed the experiment, as well as those mailed directly, if they could provide a link within their blog.

3.1.5.2 Presentation of on-line materials

In order to make the process transparent, both the introductory page and questionnaire were made as clear as possible. These pages were both piloted to check clarity and adjusted according to feedback. Standard text was used, and the format was kept as simple as possible, with obvious headings. In the questionnaire only one question was presented per line on the page.

3.1.5.3 Submission and debriefing

Upon completing the questionnaire, the subjects were taken to a page which thanked them for their participation, as well as giving them further instruction for submitting their blog data. The author's e-mail address was present, and subjects were offered the opportunity of contact if they had any queries. If the subject had not fully completed all required sections of the questionnaire, they were taken to a webpage highlighting which sections were mandatory and asked to return and complete those sections before submitting once more.

3.2 Preparation of Corpus

Approximately 100 subjects submitted sociobiographic information via the online submission form. Most of those who submitted their blog data submitted HTML files, with only a small number being text. There were a number of subjects who failed to follow their submission with their text. This situation informed the first stage of processing the corpus:

1. *Collection of absent files* For those subjects who did not send their weblog text, the data had to be collected manually.

2. *Application of selection criteria* A number of submissions did not meet various criteria: There was data from non-English speaking subjects; there were weblogs with multiple authors; the data for May 2003 was unavailable; the weblogs were not of a personal nature (see section 2.5.2 for more details), or they appeared to be considerably less than 50% personal text. This resulted in a pool of 71 subjects, each with a file of sociobiographic data and one containing their blog text.

3. *Processing of sociobiographic data* The 71 data files were processed in order to produce one file on sociobiographic data containing 71 individual entries. Each subject was given a unique identification code so their data could be processed anonymously. Numerically encoded responses to the items on the personality inventory were used to calculate personality scores for the five factors.

4. *Corpus tagging* In order to be able to extract just the personal style text from the entire HTML-encoded file, it was decided to markup the submissions with XML. Marking up consisted of high level text identification tags such as *Personal* and *Commentary*, and low level web-specific features such as *Quote* and *Link* (The schema used, represented as an XML DTD file, can be seen in the appendix (figure A.1). Many of the low level tags could be made automatically from the original HTML code; the rest of the corpus was tagged by hand. This left 71 XML files encoding all blog data submitted.

5. *Spell checking* Stylistic editing of text was kept to a minimum in order to retain as much individuality as possible (for example, non-standard words and informalism). A basic spell-check was carried out using a simple tool developed by the author based on the Wintertree Java toolkit.[2] This allowed for common mistakes to be automatically corrected and a dictionary of the aforementioned informal non-standard words to be maintained throughout. Since many of the dictionary based tools are derived from either American or British English, the spelling of words such as *colour/color* was standardised in the corpus to be consistent throughout. This stage of processing resulted in 71 spell-checked XML files.

6. *Extraction of text* All the data in a weblog was tagged and spell-checked. However, only the text that was of a personal nature, written by the author themselves, was required for linguistic analysis. Therefore, text tagged as *Personal* was extracted from the blogs and text tagged as *Quote* was removed from that text. Each chunk of personal text[3] was saved in a separate file, labelled by author identification code.

The result of this processing was that there were 71 subjects, contributing 1854 personal text files. Simple frequency statistics showed that personal chunks accounted for an average of 73.9% (SD 17.4%) of all encoded chunks. Simple word counts however revealed that they accounted for 86.8% (SD 15.5%) of all the author-written (ie. not quotes) words.

[2]Sentry Spelling Checker Engine. Wintertree Software, Nepean, Ontario, Canada K2J 3N4.

[3]More often than not a chunk is a whole post, as each post could easily be classed as one category. This was not always the case however. For example, a post may begin and end with personal text, but there may be a break in the middle to report a quiz result, which has its own category. The average ratio of chunks to post was 1.08 so this is not a common occurrence.

3.3 Demographic data report

3.3.1 Gender and age

Diaries are most predominantly kept by females (Thompson, 1982; Burt, 1994), and it has previously been suggested that the majority of bloggers are teenage girls (Orlowski, 2003). A previous study (Herring et al., 2004b) found that while each gender accounts for about half of all weblogs, blogs are in fact dominated by females of teen age and preferred by females in general. It is therefore expected to see a greater number of both females and younger subjects in our corpus. Previous studies of CMC (Herring, 2000; discussed in section 2.3) have found that women made shorter posts than men to discussion lists and news groups. However, studies finding this tend to focus on work-based groups, and findings also suggest that women are less confidant posting, often feeling intimidated by their male colleagues. Blogs on the other hand are personal and individually written. They are similar to the situations in which it was reported that women were more likely to participate. So, whilst there should be a greater proportion of females, it is not necessarily clear who would write more. Consider the following examples, common to internet humour sites:

> **GIRL'S DIARY** *Sunday 11th May 2003* - Saw Andy in the evening and he was acting really strangely. I went shopping in the afternoon with the girls and I did turn up a bit late so I thought it might be that. The bar was really crowded and loud so I suggested we go somewhere quieter to talk. He was still very subdued and distracted so I suggested we go somewhere nice to eat.
>
> All through dinner he just didn't seem himself; he hardly laughed, and didn't seem to be paying any attention to me or to what I was saying. I just knew that something was wrong. He dropped me back home and I wondered if he was going to come in; he hesitated, but followed. I asked him again if there was something the matter but he just half shook his head and turned the television on. After about 10 minutes of silence, I said I was going upstairs to bed. I put my arms around him and told him that I loved him deeply. He just gave a sigh, and a sad sort of smile.
>
> He didn't follow me up, but later he did, and I was surprised when we made love. He still seemed distant and a bit cold, and I started to think that he was going to leave me, and that he had found someone else.
>
> I cried myself to sleep.

> **BOY'S DIARY** *Sunday 11th May 2003* - West Ham were relegated today. Gutted. Got a sh*g though.

While the above samples are fictional, they represent preconceived expectations about diaries of men and women. So, not only do more women keep diaries, but it is expected that they write more. They are also expected to discuss their feelings and thoughts more than males, who prefer to discuss impersonal, external topics.

The blog corpus consists of text from 71 subjects, providing over 410,000 words (mean 5801, SD 5829). There are 47 females and 24 males (approximately 320,000 and 90,000 words respectively). This means that while the men wrote an average of about 3700 words in the month, females wrote 6800 each, almost double (despite large standard deviations, this is a significant difference: $t = 2.315$, DF=69, $p < .05$). This is reflected not only in a longer posts, but more frequent postings: women wrote on average 30 personal chunks in a month (SD 22.7) while men only 19 (SD 11.6; this is a significant difference: $t = 2.196$, DF=69, $p < .05$); women wrote an average of 251 words in each personal chunk (SD 186), while men wrote 194 words (SD 117; a non-significant difference).

The average age of the females is 27.8, and 29.4 for the males. These are similar averages but the distribution of subjects within age ranges differs between genders. Figure 3.1 shows the number of both male and female subjects, along with the total, that lie within each age range.

Whilst teenagers are far from the majority group here, younger subject, under 30, clearly dominate. These younger groups also consist of many more females than they do males. This would seem to concur that personal blogs are more likely to be kept by younger females.

3.4 Personality data report

3.4.1 Hypotheses

Before profiling the personality breakdown of subjects, this section discusses expectations for each personality trait as concerns blogging behaviour. These are drawn from

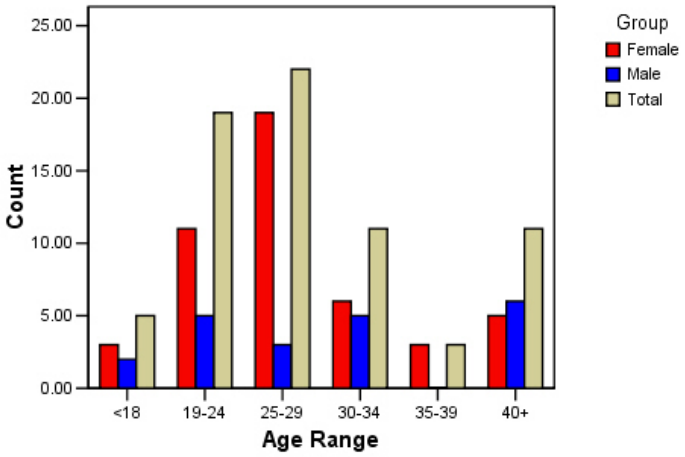

Figure 3.1: The number of subjects within each age range, by gender and in total

common perception, from previous findings (see section 2.2), and from the facets of the traits introduced in section 2.1.2.

The main intuition concerning the personality of bloggers relates to their level of Extraversion. There is much written on the web that discusses the idea of keeping a blog as 'exhibitionism' and 'mental masturbation.'[4] This suggests that bloggers are perceived as Extraverts. This is plausible, since bloggers are confidently describing details of their lives to anybody who is interested. This is supported by Extraverts' greater desire to communicate (McCroskey & Richmond, 1990).

A plausible counter argument however concerns the perceived anonymity of the online world. Bloggers write at a distance from their readers, choosing to conceal their identity if they so wish. This suggests that bloggers may be introverted by nature. This too seems plausible since it is easy to imagine Extraverts confiding in their friends directly about their activities and thoughts while Introverts would choose to communicate through a potentially anonymous written medium. In addition to this, a recent study has found that significant motivation for blogging is the author's desire to be a writer (Li, 2005): writers are generally perceived to be Introverts, due to the solitary nature of the working environment.

Likewise, similar arguments could be made for the amount people post: do Ex-

[4]See `http://jilltxt.net/?p=85` for discussion and rebuttal.

traverts write more, in the same way that they say more in the real world (Gifford & Hine, 1994), or do Introverts, because they have just as much to say, but are more comfortable in a written register?

There are less clear intuitions with regards to the remaining personality traits. As an aspect of Openness, individuals more open to experience are possibly more likely to have adopted blog technology, leading to a skewness in the distribution. Highly Conscientious individuals would seem more likely to keep their blog up-to-date, suggesting a much higher post count than people with low scores.

The impulsive facet of Neuroticism suggests that high scorers may post more frequently, more reactive as events occur or thoughts come to them. The considerate nature of highly Agreeable individuals possibly suggests shorter average post counts, as they are more aware that their readers may not want to read very long posts.

3.4.2 Scores

Subjects took a 41 item, online variant of the IPIP Five Factor Personality Inventory. This results in a numerical score on the factors of Neuroticism (**N**), Extraversion (**E**), Openness (**O**), Agreeableness (**A**) and Conscientiousness (**C**). As described above, each item loaded on one trait and was scored from 1 to 5.

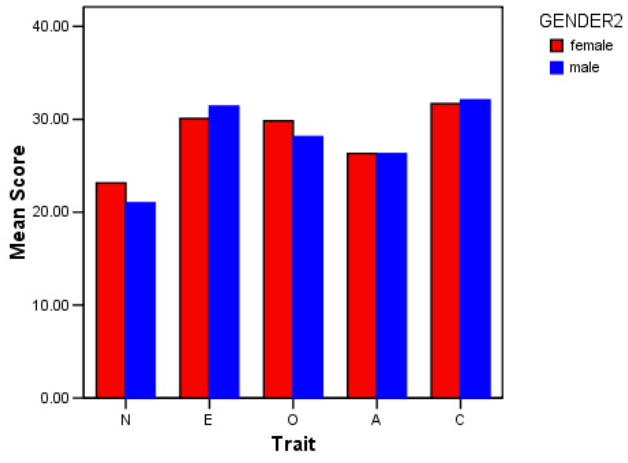

Figure 3.2: Mean personality scores for females and males

Figure 3.2 shows the means for the genders. Females score slightly higher on Openness and Neuroticism, while men score higher on Extraversion. None of these differences are significant however; there appears to be no significant link between gender and any of the personality traits. This is of great importance in the context of this thesis as both gender and personality are to be studied, and so their independence makes results clearer to interpret.

3.4.3 Score distribution

A simple method for checking personality results is to look at the distribution of each factor and compare it to a normal curve. This presents an easy way of spotting interesting aspects or abnormalities in the data. Of course, with only 71 subjects, and no comparison results, it is hard to quantify anything found as a definitive result. However, hypotheses to explain any findings will still be presented here.

Note that the range of each graph reflects the minimum and maximum possible scores based on the number of items in the questionnaire. Each graph also includes the normal distribution given the data, as predicted by SPSS.

Figure 3.3 shows the distribution on the Neuroticism scale. With variations expected from 71 subjects, the distribution is a reasonable approximation of the normal curve. Note that there were 8 items in the questionnaire that related to Neuroticism, hence scores range from 8 to 40.

Figure 3.4 shows the Extraversion distribution. Possible scores range from 9 to 45, since there were 9 Extraversion items in the questionnaire. With the exception of the two individuals scoring the two lowest possible scores, there appears to be a slight bias in favour of high Extraversion. This suggests that while Introverts do blog, it is slightly more common for Extraverts to do so. This could of course be an artifact of the online data gathering approach, as discussed previously. The distribution again is a fair approximation of the normal curve.

The Agreeableness distribution and normal curve can be seen in figure 3.5, with possible scores ranging from 7 to 35. Perhaps more so than Extraversion it can again be seen that there is a slight bias in favour of the higher scores. This suggests that bloggers are more likely to be Agreeable individuals, although again this too could be

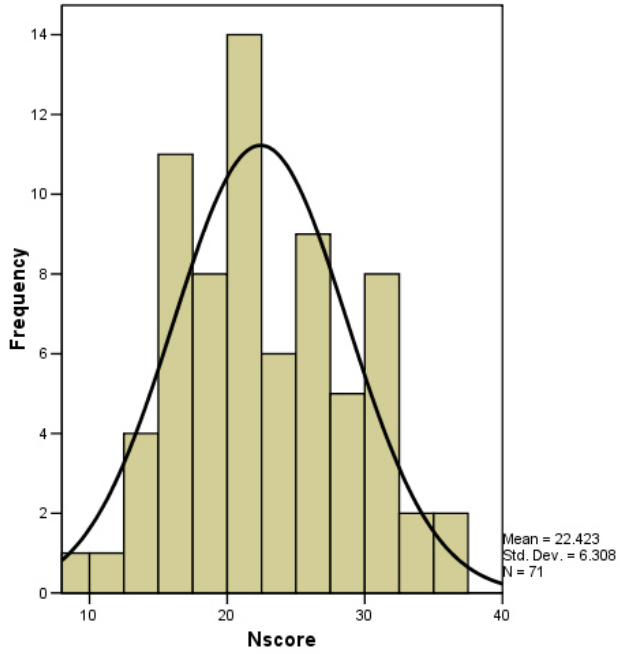

Figure 3.3: Distribution of Neuroticism scores

an artifact. Again, the normal distribution is approximated well.

The distribution of Conscientiousness (figure 3.6) also approximates the normal curve, scores ranging from 10 to 50. There appears to be a lack of individuals at either end of the distribution. This suggests that possibly highly Conscientious individuals do not like to spend time blogging, or perhaps just doing experiments about them; also people who score low on Conscientiousness never get around to blogging, or just filling out questionnaires.

Allowing for the small number of subjects, the distribution plots for the four traits detailed above suggest that the data is well within norms expected for statistical analysis. Openness (figure 3.7) tells a completely different story. The two main observations that can be made are that the data is both unevenly distributed and is significantly not normal. The lowest possible score on the Openness scale is 7, yet the lowest score from the blog data is 18, which one subject scored. Conversely, 10 subjects scored the maximum 35 points, with 8 scoring 34.

Ignoring the unevenness, it is clear there is a very heavy bias toward scoring high

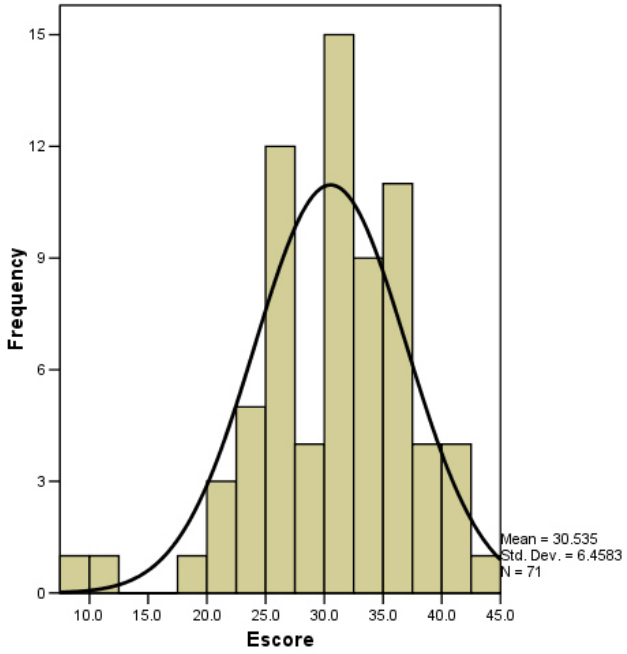

Figure 3.4: Distribution of Extraversion scores

on Openness. This may be indicative of the nature of bloggers, that they are very Open individuals. It certainly seems plausible that people who post details of their lives online could be described as open. Of course, it is possible that this result only reflects the bias among those who chose to submit data. Participation required subjects to submit personal details, and it is easy to imagine that only the most open of individuals are prepared to do so.

From personal correspondence with Tom Buchanan, one of the authors of inventory used here, it is clear that the 'Open blogger' hypothesis will remain just that. Whilst it is plausible, without a comparison sample such as non-bloggers recruited and tested in the same way, the exact cause of this extreme anomaly cannot be determined.

Regardless of the cause, this result has implications for future analysis. Statistical techniques for parametric data rely on the assumption that a variable is normally distributed. Though there are transformations for skewed data, Openness is beyond such measures. An alternative is to use equivalent non-parametric tests, though these can produce similar results.

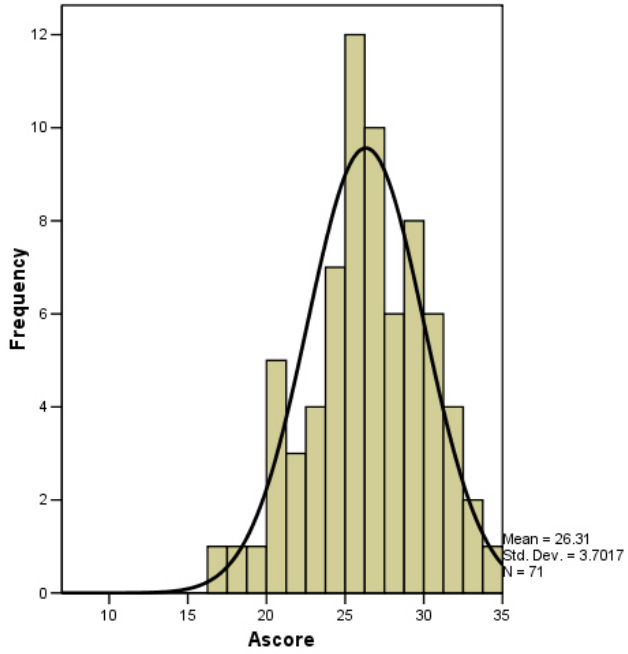

Figure 3.5: Distribution of Agreeableness scores

A simple correlation study (of Openness with a number of random variables) was carried out using both parametric and non-parametric approaches. The results of Pearson's *r* (parametric), Spearman's *rho* and Kendall's *tau* (both non-parametric) were compared. Whilst the strengths varied to a certain degree, the significance levels were similar across the three approaches. Therefore, given these results, the (theoretically) parametric nature of the variable in question (Openness) and because the intention is to compare all five personality variables in a uniform manner, Openness shall be evaluated in the same parametric manner as the 4 other traits.

3.4.4 Correlations

Theoretically, the five dimensions of personality should be independent of one another. However, it is worth investigating the possibility of any correlations between factors for our sample blogging population. Table 3.1 shows the correlations between the five factors.

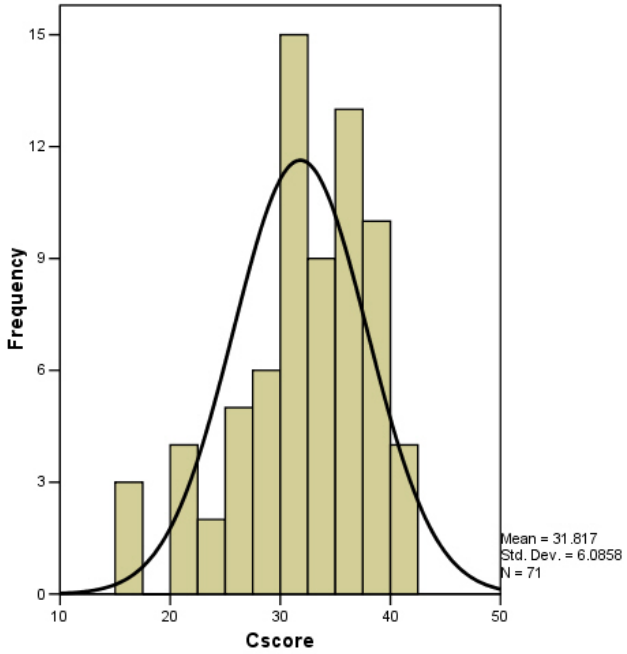

Figure 3.6: Distribution of Conscientiousness scores

It is clear that Neuroticism correlates negatively with both Agreeableness and Extraversion, most strongly with the later. While this is not desirable, it is not unexpected (cf. Matthews et al., 2003). In an evaluation of the online instrument used for personality scoring in this study, Buchanan (2005) investigated correlations between the same factors. From 2148 subjects, Neuroticism was found to correlate negatively with Extraversion, Agreeableness and Conscientiousness. Agreeableness and Conscientiousness also correlate, though positively. In the context of this study it will therefore not be unexpected for opposing effects to be found for Neuroticism and Extraversion (and to a lesser extent Neuroticism and Agreeableness).

3.4.5 Personality classes

Following Gill (2004; Oberlander & Gill, 2005) when stratifying the corpus for certain analyses, for each personality trait the extremes are considered as one standard deviation above or below the mean. Scores greater than one standard deviation above the

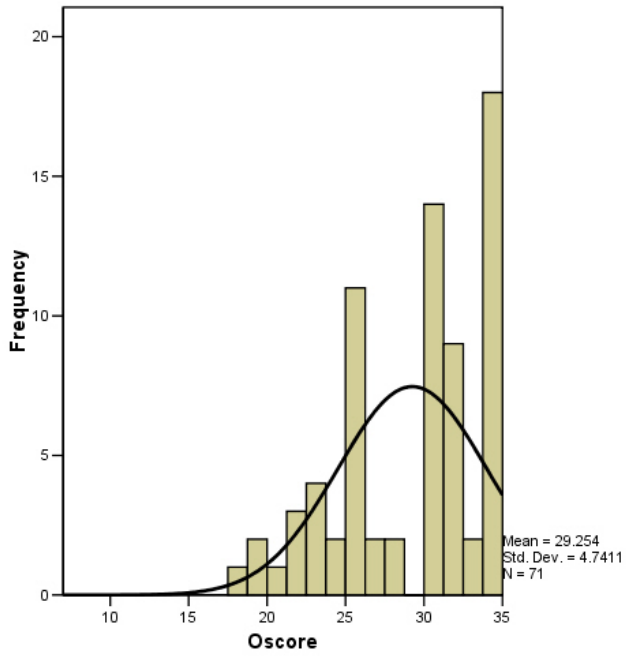

Figure 3.7: Distribution of Openness scores

mean are considered the High group. Those lower than one standard deviation below the mean are the Low group. Those in between are the Mid group.

It is clear however that this approach is not suitable for Openness. Due to the distribution, one standard deviation above the mean leaves only those scoring 34 or 35 out of a possible 35, the absolute highest Openness degree, in the high group. To stratify across Openness, consider again the distribution discussed previously (figure 3.7). There is a clear score that zero individuals achieved of 29, which appears on a plausible position on the graph. Anyone scoring higher than 29 is to be considered High, and anyone lower Mid. Since the lowest score is 18, when the lowest possible is 7, it is plausible to say that there is nobody of Low Openness in the blog corpus.

A further complication arose in early analysis of word frequencies. Certain phrases were appearing significantly more in some sub-groups than others, with no apparent personality-based reason for this. Familiarity with the data led swiftly to the conclusion that it was merely the common subject matter of one individual with one of the largest samples of text. To address this, it was decided to place a cap on text size for those

	N	E	O	A	C
N		–0.406**	0.122	–0.286*	–0.161
E	–0.406**		0.211	0.003	0.079
O	0.122	0.211		0.116	0.205
A	–0.286*	0.003	0.116		0.216
C	–0.161	0.079	0.205	0.216	

Table 3.1: Correlations between the five personality factors

Note: two tailed, *$p < 0.05$, **$p < 0.01$

analyses. The cap level was 2 standard deviations above the mean, 17459 words. This affected 3 individuals, resulting is 390,000 words in the corpus.

Table 3.2 shows the number of subjects and approximate word count of each sub-group. Alongside each figure is the percentage of the total it represents. With the exception of Openness, the Low and High subject groups are similar in size for each trait, varying between 15 and 20% of the subjects. However, the word split is not so straightforward.

For both Openness and Agreeableness, it is clear that the ratio of words in the sub-groups is similar to the subject split. There are slight differences in the other trait splits though. High neurotics, as well as low extroverts seem to write less than the rest of the subjects. These in fact are the only sub-corpora than contain fewer than 50,000 words. Low conscientious subjects seem to account for more words than expected, though this is balanced not by the high but the mid group.

That Introverts write less is understandable in the accepted features of the Extraversion trait. The other differences are less straightforward. Of course, the same limitations discussed earlier apply here. These differences are small, and with such a small subject base, it could simply be coincidence that the few individuals who have written the most fall into the same personality sub-groups.

At certain levels of analysis, Gill employed a much stricter technique when stratifying his corpus: a subject was only considered in the High/Low group of a trait if they were in the mid group of the remaining traits. This was possible for Gill because not

		Num	(perc)	Words	(perc)
	low	12	(17%)	82K	(21%)
N	mid	46	(65%)	261K	(67%)
	high	13	(18%)	48K	(12%)
	low	11	(15%)	41K	(11%)
E	mid	48	(68%)	279K	(71%)
	high	12	(17%)	71K	(18%)
	low				
O	mid	28	(39%)	160K	(41%)
	high	43	(61%)	231K	(59%)
	low	11	(15%)	69K	(18%)
A	mid	47	(66%)	250K	(64%)
	high	13	(18%)	73K	(19%)
	low	11	(15%)	78K	(20%)
C	mid	46	(65%)	230K	(59%)
	high	14	(20%)	84K	(21%)

Table 3.2: Number (and percentage) of subjects in each personality class.

only did he have more subjects ($n = 105$) but each was scored on only three personality traits.

Cross stratification of the blog corpus like this results in only 1-5 subjects in each group. Of course, the word count would still be higher than that of Gill (on average, approximately 8000 words per group) but subject numbers are far too small; groups are tied to a few individuals.

This approach would also be confused by the way in which Openness has been stratified. Rather than adopt Gill's approach, it was decide to create a further neutral sub-group. It was decided that this group should consist of those subjects who were in the mid group on the four more normally distributed personality traits. This resulted in a sub group of 11 subjects (15%) and approximately 84,000 (21%) words, comparable in size to the High and Low subgroups described above.

3.4.5.1 Text length and frequency

There appear to be large differences in the number of words for certain personality sub-groups (see table 3.2) such as the high and low Extraversion groups. It appears, as expected, that Extraverts do indeed write more than Introverts. There is much length and frequency data that can be derived from the corpus: total word count for one month, average word count per chunk, number of chunks in month, average number of chunks written per day of writing.[5]

However, when correlating personality scores with these measures, as well as using tests of statistical difference between sub-groups, there are no significant effects. So within the group at least, there is no relationship between any personality type and text length or frequency.

3.5 Summary

This chapter has discussed the creation and processing of the blog corpus to be used for this thesis. It has also described some of the features of the subjects.

Blog text and sociobiographic data was collected via an online experiment. This data was processed to extract just the personal text from each blog, and provide five-factor personality scores for each subject. The resulting corpus consists of 71 subjects, and approximately 410,000 words.

The are more female subjects than male, the majority being under 30, as is the expected demographic of personal diary weblog authors. Though there are small differences, there is no significant effect for gender differences and personality. There are also no effects for personality and text length, although women write more than men. Personality scores are reasonably well distributed, given the limited sample size, the exception being Openness. While lack of comparison figures prevent the determining of a clear explanation for this, it has some effect on work to come.

[5] *Chunks per day* is equivalent to chunk count, since every subject wrote in the same month. *Chunks per day of writing* is to eliminate the possibility that bloggers were physically unable to post on some days, by only considering those days they did post.

Chapter 4

Linguistic Profile of Blogs

Chapter 2 introduced the genre of blogs, and reviewed previous work in which they have been studied. The last chapter discussed the collection of the blog corpus for this thesis, as well as profiling aspects of the authors in the corpus. Before further exploring individual differences within blogs, this thesis intends to explore aspects of the linguistics of the genre as a whole.

Section 2.4.2 discussed the work of Herring et al. (2004a) who investigated the legitimacy of the claim that blogs are an independent genre. There has been much work on automatic genre analysis (Karlgren & Cutting, 1994; Kessler et al., 1997; Finn and Kushmerick, 2005), with attention turning to genres in CMC (Santini, 2005). However, being able to distinguish strictly between genres is not the intention of this thesis; this work merely intends to delineate the language of the blog genre. The intention is to show that while blogs are distinct, they are not so distinct as to not be representative of language in general. From previous work (cf. sections 2.4.2 and 2.6), it is clear that blogs are considered a genre unto themselves with many unique features. However, it is also clear that the nature of the language within is similar to CMC in general in that it shares properties of both written and spoken language.

This section will further explore this idea by comparing blogs to other text genres. Much work in this field has developed measures that can be used to compare large corpora (Kilgarriff 1997). Here the blog genre is profiled prior to investigating individual differences within it. Therefore, measures have been chosen that can be applied to studying individuals as easily as whole corpora.

The first measure is one of contextuality (Heylighen & Dewaele, 2002) which has previously been used to investigate both genre and individual differences. The second is a unitary measure developed to approximate the frequency of language used in a text. These measures will be used to compare the blog corpus to a number of genres drawn from the British National Corpus (BNC). The BNC consists of over 4000 files, containing over 100 million words of both spoken and written English. Files are classified by a number of categories, including register, domain and genre.

4.1 Contextuality

In 2.7.2.2 Heylighen and Dewaele's F-measure (2002) was introduced. This is a measure based on the notion of deixis, or the contextual nature of certain parts of speech. The F-measure is calculated as a positive summation of the relative frequency of nouns, adjectives, prepositions and articles, together with a negative summation of pronouns, verbs, adverbs and interjections:

$$\begin{aligned} F = \quad & 0.5 * [(noun\,frq + adj\,frq + prep\,frq + art\,frq) \\ & -(pron\,frq + verb\,frq + adv\,frq + int\,frq) + 100)] \end{aligned}$$

High scoring texts were those considered most formal in style, while those that score lower are more contextual in nature. Note from the discussion in section 2.7.2.2 that the use of 'formal' can be ambiguous. After discussion with Dewaele, within the work of this thesis, the F-measure is considered solely as a measure of contextuality.

Calculating the F-score of a number of genres from the BNC allows blogs to be placed on a scale and furnishes an opportunity to test the face validity of Heylighen and Dewaele's F-measure by examining the plausibility of that scale. The F-score of Gill's e-mail corpus can also be calculated, and included in the placement.[1]

[1] This work has previously been reported in Nowson et al. (2005).

4.1.1 Method

Using Lee's BNC World Edition Index[2] (2001), 17 genres were selected from the BNC. These included both spoken ($n = 4$) and written ($n = 13$) material, ranging from sermons and fiction writing, to text taken from newspapers and academic works. Only files dating from 1985 to 1994 and (for speech) only spoken files with a single speaker were included. Altogether there were 837 files comprised of 23 million words. The original release of the BNC comes pre-tagged using the CLAWS tagset. These tags are algorithmically reduced to the set needed for calculating the F-score of each file. These scores are then averaged to give the F-score of each genre.

Both the blog and e-mail corpora have also been tagged using the MXPOST tagger (Ratnaparkhi, 1996) and the PENN tagset. Note that errors introduced by the tagging process are not a concern here: the systematic nature of tagger errors means there will be little effect on a general parts-of-speech measure such as this. That is to say errors are likely to occur across each POS evenly, and so any one tag is no more incorrect than any other (see section 2.7.3.3 for more details). These tags were mapped down to the same set as required by the F-measure calculation for comparison. In the e-mail corpus, each file contained 2 messages from the same writer ($n = 105$). In the blog corpus each file contained all the text for an author from one month ($n = 71$). The corpora contain approximately 60,000 and 400,000 words respectively.

4.1.2 Results

When the F-score calculation was completed on the BNC genres selected, they ranked as in Table 4.1. As predicted by Heylighen and Dewaele (2002), spoken genres are on the whole more contextual than written, with Sermons, Lectures, and Unscripted Speeches scoring the lowest. Scripted Speeches are less contextual than Unscripted and also less contextual than those written genres considered most contextual: Fiction, Personal Letters and E-Mails.

Many of the results are intuitive: Academic writing is less contextual than Non-Academic; Professional Letters are less contextual than Personal; University-level Es-

[2] Available at http://clix.to/davidlee00

Genre	Ave F	*SD*
Sermons	42.4	*2.6*
Lectures on Social Science	44.3	*2.8*
Unscripted Speeches	44.4	*4.4*
Fiction Prose	46.3	*4.0*
Personal Letters	49.7	*3.3*
Sports Mailing List E-Mails	50.0	*0.6*
Scripted Speeches	53.0	*2.9*
School Essay	53.2	*2.7*
Biography	56.3	*6.4*
Non Academic Social Science	56.9	*6.0*
Nat Broadsheet Social	57.5	*3.9*
Professional Letters	57.5	*4.2*
Nat Broadsheet Editorial	58.1	*1.4*
Nat Broadsheet Science	60.0	*3.2*
University Essays	60.3	*0.6*
Academic Social Science	60.6	*3.3*
Nat Broadsheet Reportage	62.2	*1.3*

Table 4.1: Average F-score of selected genres from BNC

says are less contextual than School level. There are also degrees of similarity: Personal Letters are close to the BNC's E-Mails (which come from a mailing list; cf. Collot and Belmore, 1996).

The F-score was calculated for the new blog corpus along with Gill's existing e-mail corpus (SD 5.1, 4.0 respectively). The results are displayed, along with those of the closest genres selected from the BNC, in Table 4.2. As one might expect, the e-mail corpus is very similar to the E-Mails taken from the BNC;[3] proximity to Personal Letters follows from this. It can be seen that the blogs are scored as being significantly less contextual than the e-mails (t=3.54, DF=174, $p<.001$).

[3]This finding can be taken as evidence of the lack of effect introduced by part-of-speech tagger errors. If these genres scored differently, it would be expected to be due to such errors.

Genre	Ave F
Personal Letters	49.7
Sports Mailing List E-Mails	50.0
E-Mail Corpus	*50.8*
Scripted speeches	53.0
School Essay	53.2
Blog Corpus	*53.3*
Biography	56.3
Non academic Social Science	56.9

Table 4.2: Average F-score of E-Mail and Blog corpora as situated in the BNC genre ranking

4.1.3 Discussion

There are two key observations that can be made concerning the position of blogs with respect to the other genres. The first is with respect to the division of spoken and written genres. While it cannot be said that the F-measure delineates the two registers distinctly, blogs fall very close to both Scripted Speeches and School Essays. This adds weight to the argument that blogs share properties of both spoken and written language. The placing of both E-mail corpora similarly reflects this tendency of language in CMC.

However, despite the similar placing, the second observation to be made concerns the significant *difference* between the scores of the blog and e-mail corpus. This difference can be explained by considering some of the situational factors involved in deixis. Heylighen and Dewaele draw on four categories: the *persons* involved, the *space* of the communication, the *time*, and the prior *discourse*. When collecting e-mail data, subjects were instructed to imagine they were writing to a friend—a single *person* who knew them. The blog data however, was collected from web-published blogs. These can be read by *persons* unknown to the writer; hence, to some extent, they are written with such readers in mind. So bloggers cannot assume as large a shared context with their readers as writers of e-mails composed for friends.

Not knowing the reader means the writer can assume less about any knowledge of places, or *spaces* that are discussed. Similarly, since one cannot know when a blog post will be read, or whether any previous posts have been read, the writer can assume less about the *time* and *discourse* contexts. Interestingly, that the situation of the communication should have such a clear effect on genre difference supports the theory that genre is in part defined by purpose (see section 2.4.1).

In sum, it appears that the F-measure is a suitable method for detecting a level of difference between genres. In fact, the ordering of genres is not only intuitively plausible, but it is very similar to that found by Biber (1988) when ranking via his involved/informational factor (figure 4.1; cf. section 2.3). Biber found spoken classes at one end, with personal letters close to speeches. The other extreme end saw academic writing and press reportage.

4.2 Word Frequency

Word frequency is an important dimension in language. Genre and corpus comparison approaches are often based on frequency measures (Kilgarriff, 1997; Stamatatos, 2000b). Word frequency data is cheap to compute, offers a large number of data points, and can be applied in many way. Here, a unitary measure of average word rank is used to reflect the general language frequency level of a text.

It is not entirely clear as to what the relationship between frequency and the F-measure should be. On one hand, one would expect contextuality to be linked with more frequent words, particularly pronouns. However, the less contextual side of the equation contains both articles and prepositions, many of which are generally considered among the most frequent words.

4.2.1 Method

The basis for the reference word frequency list was the written section of the BNC.[4] Using a reference list from a general source such as the BNC has previously proved

[4] Available at the website of Adam Kilgarriff:
http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

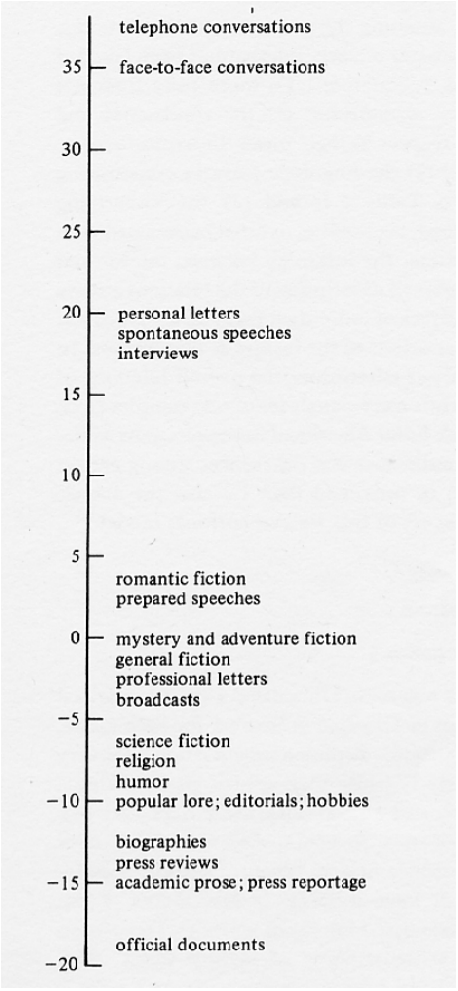

Figure 4.1: Mean scores of Biber's Dimension 1: 'Involved versus Informational Production'

to be more effective than a list generated from within a specific corpus (Stamatatos, 2000b) The reference list consists of all the words with a frequency of greater than five, which resulted in 154,759 words. While this would discount interesting low frequency linguistic phenomena such as hapax legomena, it was felt it was more computationally efficient than using a file of 921,074 words.

The list was originally ordered by frequency so each word could easily be assigned rank: starting with the most frequent word scoring the highest rank, 1. Rank was decreased with each decrease in frequency; items with the same frequency scored the same rank.[5]

Average rank usage was calculated by first calculating the frequency of each word in a file or genre. Each word would then only need to be looked up in the rank list once, the rank sum for that word being the frequency multiplied by the rank. Each word rank sum was added to produce a total rank sum for a file (or genre). The average rank was calculated by dividing the total rank sum by the total number of words. A higher score is a lower average rank which suggests greater use of low frequency uncommon words; conversely a low score is a higher average rank implying greater use of high frequency common words.

The same genres from the BNC as those in the contextuality study detailed above were used. Since the original release of the BNC comes pre-tagged using the CLAWS5 tagset, there was no need to transform the tags. In order to make transformation easier in the blog (and e-mail) corpus, the tags were derived from the processing of the WMatrix tool (Rayson, 2001, 2003). WMatrix uses the CLAWS7 tagset, so the tags were transformed into CLAWS5 for better match with the rank list reference file.

4.2.2 Results

The mean percentage of words for which rank information was found was 97.9% (SD = .53%). The highest was Sermons, with 98.8%, while the lowest was the Sports Mailing List (SML) data at 97%. The similarity of these scores suggest that no genre uses a significantly large amount of very infrequent words. The results, ordered by average

[5]Frequencies of 100, 50, 35, 35, 25 scored ranks 1, 2, 3, 3, 5.

word rank, can be found in table 4.3.[6]

Genre	Ave Freq Rank
Scripted Speeches	2266
Unscripted Speeches	2335
Sermons	2451
E-mails	*2608*
Lectures on Social Science	2850
Personal Letters	3227
Non Academic Social Science	3422
Blogs	*3495*
University Essays	3688
School Essays	3782
Academic Social Science	3904
Professional Letter	3905
Fiction Prose	3956
Nat Broadsheet Editorial	3975
Nat Broadsheet Social	4168
Nat Broadsheet Reportage	4241
Nat Broadsheet Science	4501
Biography	4538
Sports Mailing List E-mails	4602

Table 4.3: Average word frequency rank of selected genres

The results are in parts very similar to those found for the F-measure in the previous section. Spoken genres use more frequent words over all, while Gill's e-mail corpus has a similar average rank. Personal letters are lower placed than Professional letters, and Non-academic writing is lower than Academic, though both levels of essays are similar. Fiction prose is higher than might be expected, while academic writing is lower. Writing from newspapers uses lower frequency words than most, while blogs

[6]Due to method of calculation, which treated each genre as one file, it was not possible to calculate standard deviation.

are again roughly in the middle of the list. In fact, the average rank calculated here correlates significantly with the F-measure ($r = .620$, $p < .01$). However, there are some significant differences: most surprising is that Biographies, and particularly SML e-mails have such a high score (low average rank). They clearly use far less common words than would have been imagined from their high contextuality score.

On close inspection of the SML frequency data one possible reason for its low average rank presents itself. The data is very specifically taken from a mailing list for Leeds football club. Therefore, among the low frequency words are names and nicknames of Leeds players and staff, along with other teams, players and stadium. Considering the small contribution this genre makes to the overall BNC, these words have very low ranks (for example, 'Dorigo', is ranked 42758 in the BNC, but in this genre occurred 96 times and ranks 289). Since these words are used with relatively high frequency within the genre, the multiplicative effect leads a significant portion of the overall rank sum.

In fact, while names (proper nouns, tagged as *np0*) account for only 6.1% of all words in the corpus, they account for 26.4% of the total rank sum of the genre. Biographies similarly: 6.0% of words and 26.5% of the rank sum. This once again would be down to the very specific nature of biographies, containing low frequency people and place names.

In order to investigate the disproportionate effect the proper nouns had, it was decided to recalculate the average word frequency rank while ignoring all words tagged *np0*. The results can be seen in table 4.4. Also included is the percentage of proper nouns in each genre along with the percentage of the total rank for which those proper nouns account. The least that proper nouns contribute to the rank sum is 4.9% of Unscripted speeches; the most is 32.7% of the total rank sum of words in Professional letters.

The results are similar to before, excepting that those genres with the greatest percentage of nouns have moved the most. The SML texts are still, however, relatively high on the list, although Broadsheet reporting now ranks higher than blogs. Figure 4.2 plots the average ranks of the genres after proper nouns have been removed against the original average word rank. There is a clear trend, but some genres have moved

Genre	Ave Freq Rank	% words	% rank sum
Scripted Speeches	2124	1.8	6.3
Sermon	2147	2.9	12.4
Unscripted Speeches	2250	1.0	4.9
E-mail corpus	*2257*	*2.7*	*13.5*
Professional Letters	2626	6.6	32.8
Lectures on Social Science	2703	1.0	5.1
Personal Letters	2745	3.3	14.9
Non Academic Social Science	2971	2.2	13.2
School Essay	2996	3.4	20.8
Nat Broadsheet Reportage	3049	7.0	28.1
Blog Corpus	*3162*	*1.6*	*9.5*
Fiction Prose	3231	3.7	18.3
University Essay	3266	1.7	11.4
Nat Broadsheet Editorial	3281	4.3	17.5
Biography	3339	6.0	26.4
Sports Mailing List E-mails	3389	6.1	26.4
Nat Broadsheet Social	3413	3.9	18.1
Academic Social Science	3422	1.8	12.3
Nat Broadsheet Science	3762	3.5	16.4

Table 4.4: Average word frequency rank of selected genres (discounting proper nouns), percentage of words, and percentage of rank sum, contributed by proper nouns

more than others, as table 4.4 has shown.

Interestingly, the percentage of words that are proper nouns correlate strongly with the percentage those proper nouns contribute to the total rank sum ($r = .962$). This relationship can clearly be seen in figure 4.3. This suggests that while some genres use more proper nouns than others, they are no more or less common than those used in any other genre. On average nouns account for 5 times more of the rank sum than they do the word count (SD = 0.96).

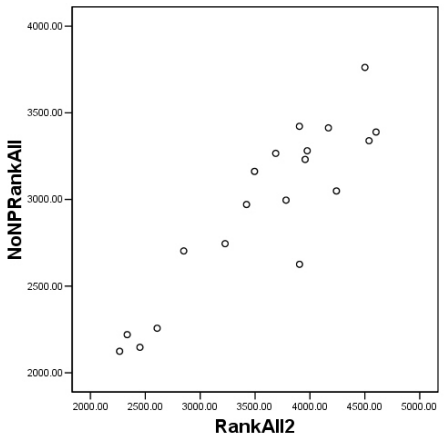

Figure 4.2: Scatter-plot of raw rank against rank without proper nouns

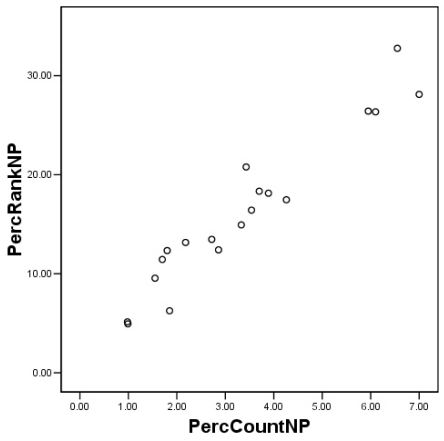

Figure 4.3: Scatter-plot of percentage of proper nouns against percentage of total rank they contribute

One interesting observation is the variability of percentage of proper nouns within the genres. Figure 4.4 is a scatter plot of the original average word rank against the percentage of proper nouns. This shows that only the lowest ranking (highest scoring) genres use a high frequency of proper nouns. They have a broad distribution however because they may also use few proper nouns. Higher ranking genres on the other hand use only a relatively low frequency of proper nouns.

Low ranking genres such as Academic writing use few proper nouns because they are impersonal. Higher ranking genres such as Blogs and E-mails have low use of proper nouns because they are overly personal and are more likely to use pronominal references to others. On the other hand, low ranking genres such as Broadsheet reportage and Biographies require by their nature a high incidence of proper nouns: places, people, companies etc. Higher ranking genres, such as the spoken genres, clearly require fewer proper nouns.

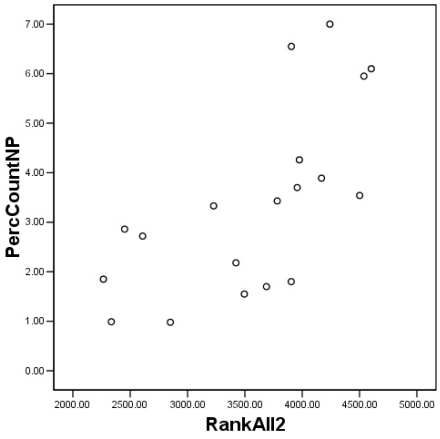

Figure 4.4: Scatter-plot of raw rank against percentage of proper nouns

Returning to the relationship with the F-measure, it is possible to investigate the relationship between average rank and the broad part-of-speech categories used to calculate the F-score. Table 4.5 shows the correlation of average rank with the eight parts-of-speech of the F-measure. Most of the categories correlate significantly with average rank, in the same direction as the load on the F-measure. However, following the earlier intuitions concerning proper nouns, it is indeed nouns which correlate most strongly.

Clearly, despite the strong correlation, there are significant differences between the end products of the F-measure and average rank. It only remains to look briefly at how exactly they relate, with a scatterplot of F-score against rank (figure 4.5). The pattern is similar to that observed in figure 4.4, which shows rank against proper noun frequency. It appears that while the more contextual genres can use both frequent and infrequent

Trait	*r*
Noun	.724**
Adjective	.490*
Preposition	.525*
Article	.523*
Pronoun	–.557*
Verb	–.487*
Adverb	–.348
Interjection	–.456

Table 4.5: Correlation between POS frequency and average rank with the BNC genres

Note: two-tailed, *$p<0.05$, **$p<0.01$

words, the less contextual genres appear only to use less common words.

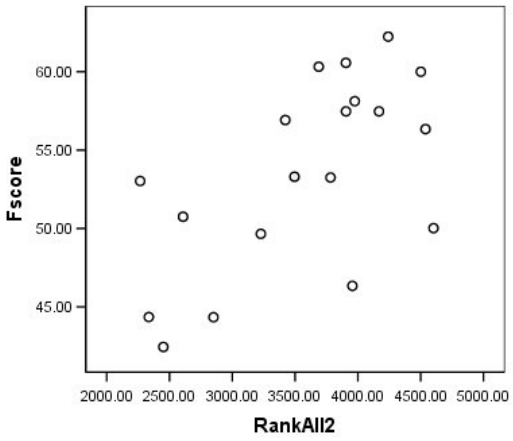

Figure 4.5: Scatter-plot of rankagainst F-score

4.2.3 Discussion

Ordering genres by average word frequency rank did not produce the overall order hypothesised. Despite this however, the relative positions of the blog and e-mail corpus are consistent with the previous finding for contextuality: that they are both situated

between spoken and written genres; and that the e-mail corpus contains more frequent words than the blog corpus.

There are a number of possible explanations as to why rank may not be related to contextuality. Firstly, some genres are simply more specific than others. Personal Letters were found to be of a similar contextuality to the Sports Mailing List data. However, it is imagined that personal letters may be about anything, whilst the sports mailing list mainly concerns one team from one sport. This specificity increases the chance that words normally infrequent in language contribute disproportionately to calculations of rank.

Secondly, as much as it is only an intuitive connection, contextuality does indeed appear to be in some way related to formality in the more traditional sense (as it relates to tone and informality). There is a strong argument that formality should also have a non-linear effect on word frequency. While formal texts may contain longer more unusual words, informal texts will contain more slang and nicknames, which will also be low frequency when considering language as a whole.

Perhaps the greatest flaw in this approach is that text from the BNC is compared against language from the BNC. Comparing data to itself can result in a degree of over-fitting. The frequency and rank of words specific to particular genres are wholly dependent on their degree of contribution to the whole corpus. To illustrate this, consider again the proper nouns of the Sports Mailing List (SML) genre.

Proper nouns constitute 6.1% of the words in the SML. Imagine that in turn the SML accounts for just 0.01% of all the words in the BNC. This means that the proper nouns of the SML account for 0.00061% of the words of the BNC, which would make them relatively infrequent. Consider the SML making a much larger contribution to the BNC, say 10%. This would result it the SML proper nouns contributing 0.61% of the words in the BNC, greatly increasing their relative frequency and thus their rank.

There are two ways to deal with this. The first approach is to balance the contribution of specific genres to the corpus when constructing it; either by taking the same amount of text from each, or by including examples from different sources (ie. mailing lists on more than one topic). The second more practical approach is to discount not only the proper nouns, but all overly specific content words. Perhaps the most obvious

way of doing this is to only consider function words (cf. Argamon et al., 2003; Koppel et al., 2002).

4.3 Summary

The aim of this chapter was to explore the distinct yet conforming linguistic nature of blogs as a genre before exploring individual differences within them. To do this, two simple unitary measures were adopted; measures which can just as easily be used to investigate individual differences.

In measuring the degree of contextuality, the F-measure has also provided evidence for both the conformity and distinctiveness of the blog genre. While the e-mail corpus proved similar to texts from the Sports Mailing List genre within the BNC, blogs proved significantly different. However, all three CMC-based genres placed between spoken and written genres. More generally, the contextuality findings suggest the F-measure will be useful for investigating individual differences.

The frequency measure adopted to further support these findings provides less concrete results, yet they are still as expected. The reasons for the lack of support have been discussed, and a possible alternative method suggested.

Chapter 5

Top-down Approaches to Personality Differences

The previous chapter examined blogs as a whole, in order to add to the discussion regarding their distinctiveness as a genre. It is now time to turn the attention of the thesis to the second of its hypotheses: personality is linguistically projected in blogs. This chapter, and the chapter to follow, explore how personality differences can affect language. As explained previously, there are two kinds of analysis used in this thesis: top-down and bottom-up. The next chapter deals with the latter; the former is considered here.

Top-down approaches as used here concern those based on pre-existing dictionaries. Dictionaries of words can either divide them into categories, or list properties associated with each word. The two dictionaries used here are: the Linguistic Inquiry and Word Count (LIWC; Pennebaker & Francis, 1999[1]) which stores words in psychologically derived categories; and the MRC Psycholinguistic Database (Coltheart, 1981; Wilson, 1987), which associates a number of properties with the words it contains.

This chapter contains five main sections. It begins with a replication of a factor analysis of 15 LIWC variables (Pennebaker & King, 1999; cf. Gill, 2004), and then explores the correlation that these factors have with personality. Following this, the

[1] As previously mentioned, there is a more recent version, LIWC2001 (Pennebaker, Francis & Booth, 2001), but remaining consistent to the two studies to be examined closely requires use of the older version.

full range of LIWC categories are explored, in order to see which categories can be associated with each trait. The final analysis is a replication of the previous stage, the correlation and subsequent multiple regression analysis, using instead the variables provided by the MRC. The chapter concludes by discussing findings in light of earlier criticisms made of dictionary-based approaches (see section 2.7.3.2).

5.1 Factor Analysis of LIWC data

Previously, Biber (1988) used factor analysis of language use to distinguish writing styles across genres, but Pennebaker and King (1999) used it to study structure within comparable texts. Using a large corpus of student essays that had been passed through the LIWC tool, Pennebaker and King chose 15 variables to use in their analysis (see section 5.1.1.1 for details of their selection criteria). This produced four distinct factors: 'Immediacy', 'Making distinctions', 'the Social past' and 'Rationalization' (for more details on these factors see section 2.2.1.4). Amongst other aspects of individuality, they investigated the relationship between their factors and the personality of their subjects, as modelled with five-factors.

There are a number of reasons for attempting a replication of this study. Firstly, this is one of the only studies looking directly at language differences and the five-factor personality model. Secondly, in his study of e-mail, Gill (2004) replicated the factor structure, suggesting their stability across genres of writing. Replication of these factors in the blog corpus would not only further validate this suggestion, but also provide a set of stable features with which to investigate personality.

5.1.1 Method

Scores on each of the LIWC's categories were required for each author. Following Pennebaker and King, each subject's scores are an average of the scores for each of their pieces of writing. For this, all 1854 individual personal texts were analysed with the LIWC, and mean scores calculated for each subject.

5.1.1.1 Variable selection

In their study Pennebaker and King (1999) outlined a number of considerations for selecting which of the 72 LIWC variables would be retained for factor analysis. Firstly, only those variables which showed reliability of .60 or greater in earlier validation studies were included. Secondly, categories were required not to overlap; for example, Prepositions were not included, since many inclusive and exclusive words are prepositions. Thirdly, categories that did not refer to the meaning or features of specific words, for example word count, were excluded. Similarly, current concern words, with their topic specific nature were also excluded. Finally, only variables that had a mean usage level of at least 1% were included.

The 15 variables to be included by Pennebaker and King in their factor analysis were: Words of more than six letters, First-person singular, Negations, Articles, Positive Emotions, Negative Emotions, Causation, Insight, Discrepancy, Tentative, Social Processes, Past Tense, Present Tense, Inclusive and Exclusive. Mean frequencies (and standard deviations) for these variables within the blog corpus can be found in table 5.1.

Table 5.2 shows these means ordered by rank, alongside the means and ranks of both Pennebaker and King's original study and those from Gill's e-mail corpus. There appears to be a basic underlying pattern; in fact the rank of means from the blog and e-mail corpora are almost identical bar the slightly lower frequency of Articles. Across all three studies, no variables are placed more than three places higher or lower; the difference is mostly only one place.

Perhaps the biggest differences are: the blog corpus uses a greater frequency of Words greater than six letters, but fewer examples of the Present tense; The original corpus contains more First-person singular pronouns, but few Articles; Causation in this study, along with Causation and Negative emotions in the e-mail corpus actually fall below the criterion of minimum 1% usage.

To ensure compatibility with the factor analysis of Pennebaker and King, as replicated by Gill, the same 15 variables were selected from the current data. However, as highlighted above (see table 5.2), the fourth of Pennebaker and King's selection criterion is not met by all variables. The criterion says that variables must have a mean

Dimension	Examples	M	SD
Words $>$ 6 letters	n/a	15.30	2.67
First-person singular	*I, me, my*	6.81	1.66
Negations	*no, never, not*	1.83	0.53
Articles	*a, an, the*	6.84	1.51
Positive emotions	*happy, pretty, good*	2.86	0.69
Negative emotions	*hate, worthless, enemy*	1.66	0.78
Causation	*because, effect, hence*	0.73	0.30
Insight	*think, know, consider*	1.71	0.45
Discrepancy	*should, wish, want*	1.94	0.62
Tentative	*maybe, perhaps, guess*	2.43	0.65
Social processes	*talk, us, friend*	5.90	1.75
Past tense	*walked, were, had*	4.06	1.00
Present tense	*walk, is, be*	9.96	1.95
Inclusive	*with, and, include*	5.77	0.76
Exclusive	*but, except, without*	3.61	0.82

Table 5.1: Mean relative frequencies for LIWC variables selected for factor analysis

Dimension	Nowson		P&K		Gill	
Words $>$ 6 letters	15.3	(1)	13.06	(2)	12.69	(1)
Present tense	9.96	(2)	13.95	(1)	11.12	(2)
Articles	6.84	(3)	4.73	(6)	6.17	(6)
First-person sing.	6.81	(4)	10.63	(3)	6.51	(3)
Social processes	5.9	(5)	6.51	(4)	6.34	(4)
Inclusive	5.77	(6)	5.95	(5)	6.32	(5)
Past tense	4.06	(7)	3.79	(8)	4.56	(7)
Exclusive	3.61	(8)	4.21	(7)	3.55	(8)
Positive emotions	2.86	(9)	3.38	(9)	3.1	(9)
Tentative	2.43	(10)	2.84	(10)	2.62	(10)
Discrepancy	1.94	(11)	2.84	(11)	2.18	(11)
Negations	1.83	(12)	2.18	(13)	1.69	(12)
Insight	1.71	(13)	2.47	(12)	1.65	(13)
Negative emotions	1.66	(14)	1.8	(14)	0.99	(14)
Causation	0.73	(15)	1.1	(15)	0.68	(15)

Table 5.2: Means (and ranks) of 15 LIWC variable scores for three studies

Note: ordered by rank in this study

Study	No Var	Bartlett's	KMO
Pennebaker & King	15	2831	.633
Gill	15	333	.580
Nowson	15	340	.714
Nowson	14	318	.702
Gill	13	278	.600
Nowson	13	290	.714

Table 5.3: Bartlett's test of sphericity and KMO scores for the six samples

Note: Bartlett scores $p < 0.001$

usage of at least 1%, so with a score of only .73%, Causation should be excluded. To allow for this discrepancy, the factor analysis was carried out a second time with the remaining 14 variables. Similarly, Gill's data revealed two variables not meeting the criterion: Negative Emotions and Causation, .99% and .68% respectively. Therefore, to check compatibility with Gill, a third factor analysis was also carried out, with the 13 variables that Gill used.

Exploratory factor analysis was carried out on the means of each subject's texts, in much the same manner as Pennebaker and King and Gill. Diagnostic tests (Bartlett's test of sphericity,[2] and Kaiser-Meyer-Olkin's measurement of sampling adequacy,[3] KMO) reveal similar suitability to the previous studies. The test results for the three samples of this study, the two of Gill, and Pennebaker and King's original are shown in table 5.3.

[2]Factor analysis requires that variables be independent from one another. If the obtained Bartlett's score is significant, this has been shown.

[3]KMO indicates the proportion of variance in the variables which is common variance, which might be caused by underlying factors. Scores range from 0 (very bad) to 1 (perfect), so the higher the score the better.

5.1.2 Results

5.1.2.1 Analysis using 15 LIWC variables

The first stage is a factor analysis using all of Pennebaker and King's 15 selected LIWC variables. Five eigenvalues are over 1, and examining the scree plot indicates that a four factor solution would best fit the data. Principal-components analysis was used to determine the four factors, and varimax rotation was applied to aid interpretation. All 15 variables had communalities[4] greater than .40.

Rotated factor loadings[5] are shown in table 5.4. Note that only loadings greater than .4 are shown to further aid interpretation. The results of Pennebaker and King, and Gill can be found in the appendix (tables B.1 and B.2). In order to aid comparison table 5.5 shows which variables loaded on which factors and in which direction for the three studies. Factors 1 and 2 of both Pennebaker and King, and Gill, have been switched, as there is greater similarity between them and factors 2 and 1 of this study.

The most striking observation about these results is that where Gill's factors 1 and 2 match those of Pennebaker and King, in this study the factors appear to be reversed. Variables loading onto factor 2 (eigenvalue 1.55) of Pennebaker and King, which they call 'Making distinctions', are exclusive words, discrepancy words, negations and tentative words, all positively loaded, and inclusive words with a negative loading. Gill's factor 2 (eigenvalue 1.91) had the same loadings with the exception of discrepancy words which did not load on this factor. Factor 1 of this study (eigenvalue = 4.79) also contains positive loadings of exclusive words, discrepancy words, negations and tentative words. However, also loading on factor one are words relating to insights, causation and the present tense.

Factor 1, termed 'Immediacy', in the previous studies (Pennebaker and King eigenvalue = 3.35, Gill eigenvalue = 2.92) included positive loadings for discrepancy words, words relating to the present tense, and use of the first-person singular. There were also negative weightings for the use of articles, and words of length greater than 6. In addition to these five, Gill found a loading for insight words. Factor 2 of this study

[4] Intuitively: variables with high communality share more in common with the rest of the variables.

[5] As in Pennebaker and King, and Gill, the factor loadings are rotated. However, in Gill (2004) the percentage of variance reported for each factor is that for the unrotated factor solution. Therefore, while in the text unrotated variances are compared, both are reported in the results tables.

	Factor 1:	Factor 2:	Factor 3:	Factor 4:
original	(29.9% var)	(10.0% var)	(9.1% var)	(8.5% var)
rotated	(20.6% var)	(15.5% var)	(12.3% var)	(8.9% var)
Exclusive	.767			
Discrepancies	.753	.421		
Tentative	.731			
Causation	.608			
Present tense	.548	.486		
Negations	.538	.532		
Insight	.478		.436	
Words $>$ 6 letters		–.823		
Articles		–.737		
First-person sing		.525		
Positive emotions			.722	
Social			.694	
Past tense				.651
Inclusive				.513
Negative emotions				–.424

Table 5.4: Rotated factor loadings for exploratory analysis of 15 LIWC variables

	1N	2PK	2G	2N	1PK	1G	3N	3PK	3G	4N	4PK	4G
Exclusive	+	+	+									
Discrepancies	+	+		+	+	+						
Tentative	+	+	+									
Causation	+										+	+
Present tense	+			+	+	+		+				
Negations	+	+	+	+								
Insight	+					+	+				+	
Words>6 letts				–	–	–						
Articles				–	–	–						–
First-pers sing				+	+	+						
Pos emotions							+	–	+			
Social							+	+	+			
Past tense								+	+	+		
Inclusive		–	–						–	+		
Neg emotions										–	–	+

Table 5.5: Direction of loading found in the three studies using 15 LIWC variables

Note: factors 1 and 2 of both Pennebaker & King and Gill are switched.

(eigenvalue = 1.50) has similar loadings for the five variables discussed, in addition to negations being present.

This reversal of factors is further highlighted by the amount of variance that they each account for. For Pennebaker and King, the first two factors account for 22.4% and 10.3% of the variance respectively. Likewise, Gill's factors account for 19.4% and 12.8%. The equivalent factors in this study account for 10% and 29.9%. The most obvious explanation for this is that the frequency of the types of words included in the texts is different: words associated with the previous factor 1 are not as prevalent in the current study as those previously associated with the less important factor 2.

The relationship between the remaining factors is less clear. There are few loadings common to the current study and the previous ones. Factor three, following the previous studies, has a positive loading for social words, but in agreement with Gill, there is a positive loading for expressions of positive emotion where Pennebaker and King found it to be negative. Both previous studies found loadings for past tense words, Pennebaker and King found a positive association with the present tense, and Gill negative for inclusive words. The only other loading this study found was insight words.

Insight words are the one category on which there is no agreement: they appear in this study's factors 1 and 3; Gill finds them in his factor 1 (this study's factor 2) and Pennebaker and King find them loading on factor 4. In terms of agreement on factor 4, both previous studies find a positive loading for causation, which this study placed solely in factor 1. This study did agree with Pennebaker and King with a negative loading for negative emotions, though Gill found a positive loading. Factor 4 was also loaded with both words of an inclusive and past tense nature, which the other studies were not. Gill was also alone in finding a loading for articles.

Though the factors differ slightly, there are obvious similarities, particularly in the first two factors of each study. Likewise, the overall variance accounted for by the 4 factors was similar: 51.1% for Pennebaker and King, 53.3% for Gill, and 57.4% here.

5.1.2.2 Analysis using 14 LIWC variables

Since one of the variables in the current study did not meet Pennebaker and King's criterion that mean usage should be greater than 1%, the factor analysis was carried out

again using the remaining 14 variables. The scree plot again suggested a four factor solution, since five had an eigenvalue of greater than one. Principal-components analysis was once more used to extract four factors, and varimax rotation used to ease interpretation. With the exception of insight and past tense words (.35 and .36 respectively) all variables had a communality greater than .47.

The rotated factor loadings are shown in table 5.6. Again, to aid comparison table 5.7 shows which variables loaded on which factors and in which direction for the three studies. The 15th variable in the two previous studies is included in italics. The first two factors of previous studies are again switched.

The first two factors found in this study are still the opposite of Pennebaker and King's first two, but they are closer matches than previously. Factor 1 here (eigenvalue = 4.32) more closely matches Pennebaker and King's factor 2 now, due to the removal of causation words, and the failure of insight words to load on this factor. The only difference is that the present tense loads here, but not in the original factor 2.

Factor 2 (eigenvalue = 1.45) matches Pennebaker and King's first factor as well as it did previously, matching all loadings. Extra to those of the original study, in the previous stage of this work negations loaded on factor 2, but here insight words load.

Factors 3 and 4 (eigenvalues = 1.36 and 1.25 respectively) match those of Pennebaker and King as well as they did at the last stage, which is to say poorly. Factor 4 only matches with the negative loading of negative emotions. Factor 3 has a similar loading of social words, but an opposing loading of positive emotions. Negative emotions also load on factor 3, in place of insight words at the last stage.

This factor model is perhaps the strongest so far, as it accounts for 59.8% of the total variance.

5.1.2.3 Analysis using 13 LIWC variables

In Gill's replication of Pennebaker and King's work, he found two variables that did not meet the mean usage of at least 1% criterion: causation and negative emotion words, .68% and .99% respectively. To make a better comparison with the results of Gill, a further factor analysis was carried out with the same two variables removed. Causation words, with a mean usage of .73% are suitable for removal, and while in

	Factor 1:	Factor 2:	Factor 3:	Factor 4:
original	(30.8% var)	(10.4% var)	(9.7% var)	(8.9% var)
rotated	(20.1% var)	(17.7% var)	(12.4% var)	(9.7% var)
Exclusive	.818			
Tentative	.766			
Discrepancies	.746	.434		
Negations	.653			
Articles		–.779		
Words $>$ 6 letters		–.771		
First-person sing		.581		
Present tense	.500	.573		
Insight		.420		
Positive emotions			.750	
Social			.648	
Negative emotions			.613	–.474
Inclusive				.652
Past tense				.555

Table 5.6: Rotated factor loadings for exploratory analysis of 14 LIWC variables

	1N	2PK	2G	2N	1PK	1G	3N	3PK	3G	4N	4PK	4G
Exclusive	+	+	+									
Discrepancies	+	+		+	+	+						
Tentative	+	+	+									
Causation											+	+
Present tense	+			+	+	+		+				
Negations	+	+	+									
Insight				+		+					+	
Words>6 letts				–	–	–						
Articles				–	–	–						
First-pers sing				+	+	+						
Pos emotions							+	–	+			
Social							+	+	+			
Past tense								+	+	+		
Inclusive		–	–						–	+		
Neg emotions							+			–	–	+

Table 5.7: Direction of loading found here with 14 and in previous studies using 15 LIWC variables

Note: factors 1 and 2 of both Pennebaker & King and Gill are switched. Italicised variables are those excluded from the current study.

this study negative emotion words received 1.66% usage, this was the second lowest mean of the 15 variables after causation. See table 5.2 for details.

The scree plot for this data suggested a three factor model, as there were four factors that had an eigenvalue greater than 1. Once again, principal-component analysis extracted the factors, and varimax rotation was used to enable interpretation. All variables had communality greater than .33. The rotated factor loadings are shown in table 5.8. Once more, table 5.9 shows the direction of the factor loadings for the 13 variables of both this study and Gill, and the loadings for Pennebaker and King, with the additional variables in italics. In this comparison, Factors 1 and 2 of Gill's study are not switched, as in his study they switch themselves.

The ordering of the factors remains similar to the previous analysis, and this is more in line with Gill's findings at this stage, since upon reducing the variables from 15 to 13, he found the ordering of the first two factors reversed.

Factor 1 (eigenvalue = 4.17), like Gill's factor 1, includes positive loadings for negation and exclusive, discrepancy, and tentative words. This study also found a positive loading for present tense words, whilst Gill found a negative loading for the past tense. Factor 2 (eigenvalue = 1.44) had similar loadings to Gill with the exception of insight words. As with the previous analyses insight loaded onto factor 2, but Gill found that it did not load strongly to any of his factors.

This study also found the same variables loaded on factor 3 (eigenvalue = 1.36) as Gill. However, as with its loading on factor 1, he found that inclusive words loaded negatively, whereas they have always been positive in this study. Overall the variance covered by the 3 factors is 53.6% compared to Gill's 48.2%, and, discounting their fourth factor, Pennebaker and King's 42.5%.

As Gill noted in his study, the three factors derived at this stage more closely match the first three factors of Pennebaker and King than at any other stage. Their 'Making distinctions' factor is identical to Gill's first factor, and so the differences with this study and the original are as described above. For 'Immediacy' the only differences are that, like Gill, insight is not found to load, but unlike Gill, discrepancies are, as they did in the previous stages of this work and Gill's. There is also a more clear match for Pennebaker and King's 'Social past' factor. The past tense and social words are both

	Factor 1:	Factor 2:	Factor 3:
original	(32.1% var)	(11.1% var)	(10.4% var)
rotated	(21.0% var)	(20.2% var)	(12.4% var)
Exclusive	.833		
Discrepancies	.762		
Tentative	.728		
Negations	.674		
Articles		–.830	
First-person sing		.667	
Present tense	.488	.638	
Words $>$ 6 letters		–.608	
Insight		.444	
Positive emotions			.670
Social			.622
Inclusive			.600
Past tense			.453

Table 5.8: Rotated factor loadings for exploratory analysis of 13 LIWC variables

positively loaded in both studies, and positive emotion words load here, but they were found to be negatively loaded. Present tense words also loaded on ‘the Social past’ in the original study.

5.1.3 Discussion

The overall similarity between the factor analyses done here, and those previously performed by Pennebaker and King, and Gill, appears to be reasonable. The closest agreement found with Pennebaker and King’s original 15 variable analysis appears to be that found in the latter models of both Gill and this study. These were the analyses conducted with the 13 variables that Gill found to match Pennebaker and King’s original criterion for inclusion.

The similarity of the factor analyses of the three studies has two main implications:

	1N	1G	2PK	2N	2G	1PK	3N	3G	3PK	4PK
Exclusive	+	+	+							
Discrepancies	+	+	+			+				
Tentative	+	+	+							
Causation										+
Negations	+	+	+							
Articles				–	–	–				
First-person sing				+	+	+				
Present tense	+			+	+	+			+	
Words $>$ 6 letters				–	–	–				
Insight				+						+
Positive emotions							+	+	–	
Social							+	+	+	
Inclusive		–	–				+	–		
Past tense							+	+	+	
Negative emotions										–

Table 5.9: Direction of loading found here and by Gill with 13, and in the original study using 15 LIWC variables.

Note: factors 1 and 2 of both Pennebaker & King, but not Gill are switched. Italicised variables are those excluded from the current study.

The first is that the factors derived from those selected with Pennebaker and King's criterion appear fairly robust. The second is that the text used in this study, that of on-line diaries, is similar, and therefore comparable to the e-mail texts of Gill and the written texts of Pennebaker and King. This second implication is perhaps a more expected result, since half of Gill's texts were e-mails "*about what has happened...in the past week*" and half of Pennebaker and King's texts were essays about "*what it has been like for you coming to college*" so that both, it could be argued, are diary-like in style.

There are three main difference between the models: the lack of a fourth factor; the reversal of the first two factors; and differences in variable loadings. With respect to the first difference, from personal correspondence with James Pennebaker it seems that the third and fourth factors have always proved the hardest to replicate. It is interesting that the two variables dropped by Gill for having a mean usage of less than 1% are both in Pennebaker and King's final factor, along with insight words. Not only are causation and negative emotion words the lowest ranked of all fifteen variables in all three studies (see table 5.2), but insight is the next lowest in both this study and Gill's. It makes sense then that tightening the restrictions on which variables are included in the model, would result in a more reliable, robust, and crucially replicable analysis.

Looking at it this way, since it could be argued that the three factor model of this study most closely replicates Pennebaker and King's first three factors, the factors can be said to represent the same ideas. That is to say the first factor is about 'Making distinctions,' the second is about 'Immediacy' of language, and third relates to 'the Social past.'

With respect to the second difference it may be that tightening the model is what produces the reversal of importance of factors. However, since this occurred from the first analysis of this study, it may be that 'Making distinctions' is more important in some corpora than 'Immediacy', and vice versa.

The third difference mostly boils down to minor variations in variable loadings. However, most significant and interesting is the loading of Positive Emotion words on the third factor, 'the Social past,' which was negative in the original study, but positive in the two subsequent analyses. The mean usage of Positive Emotions words was

similar between all studies, and the ranking identical (see table 5.2), so it cannot be an effect of frequency. Pennebaker and King's finding means that talking of 'the Social past' involves a distinct lack of positive emotion words, whilst they are positively expected in the subsequent studies.

Another way to describe this, is that whilst overall there was comparable use of positive emotion words in all corpora, individuals who used social words and past tense words in Pennebaker and King's essays are not the same as those who used positive emotion words in the task, and vice versa. In the blog and e-mail corpora however, the same individuals used all three.

Gill proposed that this was due to Pennebaker and King's overtly requesting emotional writing ("*thoughts and feelings*") whereas he requested a more fact based approach to the subjects' activities ("*what has happened to you or what have you done.*") This explanation, however, cannot explicitly hold for this study, since the subjects' texts were written without any knowledge of the study, and so they were given no instructions about what to write. They were free to write their diary as simply a discussion of the days' activities, or a lengthy diatribe on how they felt at any given moment. Aside from that, the equivalent relative frequency of positive emotion words in Gill's corpus does not suggest a lack of emotion words.

Despite Gill's explanation not proving adequate, an explanation must lie within the differences between the data sets. Pennebaker and King asked subjects to write essays on two topics: the first was a train of thought exercise which involved writing feelings and thoughts; the second was to write about what coming to college for the first time was like, and how it compares to the subject's life previously. One can imagine that the first of those would certainly illicit emotion words, but not necessarily those concerned with the past, or social functions. Conversely, the second task would most certainly illicit words concerning the past, and one could easily imagine social words ('*I made many new friends*' or '*I went to lots of parties*'). With regards to positive emotion words, one could easily imagine subjects describing their feelings about coming to University, a significant and daunting step in life, with more negative words than positive ('*I was really worried about where I would be living*' or '*it was quite*

intimidating and scary').[6] This suggests that there would be minimal co-occurrences of positive emotion words with social and past tense words.

In Gill's study, he requested subjects to write two 'e-mails': the first concerned "*what has happened to you, or what have you done in the past week*"; the second asked "*what your plans are for the next week*." The use of the past tense is most likely to occur in the former e-mail. It is easy to see how writing about things that have already happened might lead to writing about any emotions felt at the time. It is less likely that writing about events yet to come would elicit much mention of emotions. So there is an obvious relationship between discussions of the past and positive emotion words, so they are both going to load positively on any factor relating to 'the Social past.'

In this study the majority of the text, since it is essentially a diary style, is discussing prior events and emotions. So, like Gill's 'past' e-mails, there is a strong relationship between positive emotion words and the past.

This is of course merely a theory, as it is not possible to access Pennebaker and King's original data, and can only suggest plausible differences between the different corpora of these studies. However, this result does suggest that whilst there may be patterns to text of similar nature, in this case personal writing, differences in the context of the writings can affect those patterns.

5.2 Correlation of LIWC factors with personality

Following Pennebaker and King, this section investigates correlates of personality. It uses the factors derived in the previous section, and the related variables in order to investigate language differences within personality dimensions.

5.2.1 Method

Since the factor analyses of this study led to three variations of the factors found by Pennebaker and King, while Gill had two, for ease of understanding not all results

[6] These are not quotes from the texts of Pennebaker and King, but merely examples of what one might imagine being said in the respective contexts.

will be compared. The previous section, particularly table 5.9 shows that the factor analyses with *Causation* and *Negative emotion* words removed most closely resembled the original factors of Pennebaker and King. Therefore only the 13 variable solution, the similar solution of Gill, and that of the original study will be compared. This also means that there will not be a comparison of the fourth factor in any detail, since it is the factor on which there is least agreement. It should be noted that due to vastly different population sizes (this study's 71 and Gill's 105 contrast strongly with Pennebaker and King's 841 subjects) it is hard to make comparisons about correlation strengths. What can be compared are the correlations that reach significance, as well as differences in direction.

5.2.2 Results

For completeness, scores for the 15 variables (table C.1), and the 14 variables (table C.2) can be found in the appendix. The correlation results of the 13 variable factor analysis can be seen in table 5.10. Scores for the original study can be seen in table B.4 while Gill's results are in tables B.5 and B.6. Comparison will be performed by personality trait for each factor, beginning with Neuroticism and Extraversion, since these were in all three studies, followed by Openness, Agreeableness and Conscientiousness.

5.2.2.1 Neuroticism

In Pennebaker and King's study, neither 'Making distinctions' (factor 1 of this study) nor any of its constituent variables correlate significantly with Neuroticism (r=.05). By contrast, while still not significant, Gill found a negative correlation for the factor (r=–.11). In this study the relationship remains positive but reaches significance (r=.245). The strongest and most significant correlation is that of discrepancy words, unlike in the previous studies (r=.339). The only variable of Gill's to stand out is Inclusive words which have a significant positive relationship, as to be expected with his finding that they negatively load onto the factor (r=.26). Though Inclusive words are not in the first factor of this study, they also do not correlate significantly with Neuroticism.

Dimension	N	E	O	A	C
Factor 1	**.245***	**–.190**	**–.055**	**–.252***	**.019**
Exclusive	.133	–.079	.143	–.189	–.061
Discrepancies	.339**	–.251*	–.118	–.290*	–.035
Tentative	.140	–.144	–.107	–.198	.060
Negations	.163	.020	–.222	–.245*	.098
Present tense	.159	.200	.009	–.092	.102
Factor 2	**.003**	**.166**	**–.180**	**–.139**	**.068**
– Articles	–.072	.031	.136	.255*	–.054
First-person singular	–.017	.175	–.098	–.081	–.060
Present tense	*.159*	*.200*	*.009*	*–.092*	*.102*
– Words $>$ 6 letters	.020	–.055	.290*	.262*	.034
Insight	–.111	.063	.106	.094	.075
Factor 3	**–.079**	**.179**	**.295***	**.133**	**–.154**
Positive emotions	–.043	.162	.127	.069	–.060
Social	–.035	.238*	.195	.037	–.109
Inclusive	–.012	.015	.249*	.094	–.091
Past tense	.011	–.116	–.028	–.125	–.157

Table 5.10: Correlation of LIWC factors (13 variables) with personality scores

Note: $n = 71$, two tailed, *$p<0.05$, **$p<0.01$. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading.

Factor 2, or ‘Immediacy’ in the previous studies, has only a very weak correlation in this study (r=.003). However, both previous studies found at least some relationship, and Pennebaker and King found it to be significant (r=.10). Pennebaker and King also found significant correlations with the use of First-Person Singular and Articles (r=.13, and r=–.09 respectively) in the directions expected given their loading. Given the variable loadings for factor 2 here, more than half correlate in a counter intuitive direction, but given the weakness of the factor correlation, that is understandable.

In this study factor 3, or ‘the Social past’ correlates more strongly than factor 2, but is still a non-significant negative score (r=–.079). Pennebaker and King also found only a weak correlation, though it was positive (r=.04). Gill however found a significantly negative one (r=–.24). Likewise, most of the associated variables seem to have strong leanings in the right directions, including Inclusive words significantly. Pennebaker and King found a significant negative correlation for positive emotions (r=–.13), which is interesting because despite the factor correlation disagreeing between the studies, the direction corresponds with their unique finding of a negative loading. The remaining correlations from this study remain weak, yet almost consistently negative, in agreement with Gill.

5.2.2.2 Extraversion

Pennebaker and King found significant negative correlations between their ‘Making distinctions’ factor, factor 1 here, and Extraversion (r=–.14). This is supported by significant correlations for the variables of this factor. Though the correlations of this study are not significant, they are of reasonable enough strength as to be in agreement (r=–.19). That sole exception to this is the Present Tense, which correlates positively (r=.200). This does agree with Pennebaker and King’s findings (r=.01), but not with the direction of the factor. Gill also found a negative correlation, though it was much weaker (r=–.03), and there was less consistency within the variable. In this study, Discrepancies were once again the only factor of this variable to correlate significantly (r=–.251).

Very little correlated significantly with Extraversion from factor 2. The biggest difference is the positive direction found in this study and the original (r=.166 and

r=.04 respectively) while Gill found it to be negative in his data (*r*=–.08).

Whilst the correlation found with factor 3 was only as great as that of factor 2, (*r*=.179), it produced a significant correlation in Social words (*r*=.238). Pennebaker and King also found significance for Social words (*r*=.12) but also Positive Emotions, (*r*=.15). This positive direction for Positive emotion words went against Pennebaker and King's finding for loading direction, but was in agreement with this study and that of Gill (*r*=.162 and *r*=.15 respectively). Gill also found a positive correlation for the factor (*r*=.11) but Pennebaker and King found it to be of no relation (*r*=.00).

5.2.2.3 Openness

Factor 1, or 'Making distinctions,' in both studies[7] is not significantly correlated with Openness (*r*=–.055 for this study and *r*=.06 for the original). The biggest difference is that here the factor, and variables within, correlate mostly negatively, whereas Pennebaker and King found it to be positive. They found the use of Tentative words the only significant correlation (*r*=.11), but Negations and Inclusive words barely correlated at all (*r*=.00 and *r*=–.01 respectively).

With 'Immediacy' and Openness, Pennebaker and King found their strongest correlation: both the factor and all but one of the corresponding variables (Discrepancy words) reached a strong level of significance (*r*=–.16 for factor, and similar scores for variables). This study found a similar, yet non-significant negative correlation trend (*r*=–.180). The biggest inconsistency is that Words of Greater than Six Letters, since it loaded negatively onto factor 2 of both studies, should correlate positively with Openness. Indeed it is the only variable to significantly correlate in this study (*r*=.290) but Pennebaker and King found a strong negative correlation (*r*=–.16).

Both studies found a significant positive correlation with factor 3, 'the Social past' and Openness, and in fact in this study it was the strongest factor correlation (*r*=.295 here and *r*=.08 previously). This study appears to have stronger correlations than were previously found. In particular Inclusive words, which as mentioned earlier Pennebaker and King only found to correlate a very small amount, significantly correlated

[7] For Openness, Agreeableness and Conscientiousness, comparison focuses on Pennebaker and King, since these factors were not in Gill's study.

with Openness (*r*=.249).

5.2.2.4 Agreeableness

Pennebaker and King found a negative relationship between 'Making distinctions' and Agreeableness, but no scores were significant (*r*=–.05). In this study, the correlation was still negative, but was much stronger, with the factor and two of its variables (*r*=–.252 for the factor itself, *r*=–.245 for Negations and *r*=–.290 for Discrepancy words), correlating significantly with personality scores for the third time.

Factor 2 provides the most significant disagreement in this study. Pennebaker and King find a positive correlation with Agreeableness (*r*=.07), while here it is found to be negative (*r*=–.139). This is a more distinct disagreement than that of factor 1 and Openness because the relative strength levels of the factor variable correlations are greater, many of them proving here to be significant in their own right (*r*=.07 for First-Person Singular and *r*=–.15 for articles in the original study and *r*=.255 for Articles and *r*=.262 for Words of Greater than Six Letters in this study).

'The Social Past,' while less significant, also appears to have provided differing results. Though their factor correlation is weak (*r*=–.02), Pennebaker and King's negative loading for Positive Emotion words and their significant positive correlation (*r*=.07) suggests that there is a negative correlation for this factor with Agreeableness. In contrast, this study found a positive correlation (*r*=.133).

5.2.2.5 Conscientiousness

Correlations with Conscientiousness on the whole tend to be weaker than with previous personality traits, while variable scores seem less consistent within factors. There do however, still appear to be noticeable differences between the studies.

Pennebaker and King's strongest results for Conscientiousness come from 'Making distinctions' (*r*=–.13) and the associated variables, with both Exclusive words and Negations reaching a significant level (*r*=–.08 and *r*=–.15 respectively). However, despite the correlations here being much weaker, certainly non-significant (*r*=.019 for the factor itself), and there being inconsistencies between the variable directions which should all be the same, the results appear to paint a picture of positive correlation.

Likewise, despite low scores and inconsistencies, it appears that Pennebaker and King found a negative correlation for 'Immediacy' (r=–.02) while here it is positive (r=.068). When it comes to the third factor however, the results are not only more consistent, but also in agreement: there is a negative correlation between Conscientiousness and 'The Social Past' (r=–.154 in this study and r=–.04 in the original study).

5.2.3 Discussion

The first observation that can be made when comparing the correlations across the three studies is that they are quite modest. The current study showed by far the strongest correlations, but Pennebaker and King found their's to be more significant. This is mainly due to the population size, with 841 subjects against Gill's 105 and this study's 71.

In summarising their results, Pennebaker and King found three patterns. First, they found almost all of the 'Immediacy' variables correlated negatively with Openness. Whilst in this study, the variables that Pennebaker and King found to load on their first factor almost all correlate negatively, there are minor variations, and weaker links. So, even though the factor itself does show a negative correlation of similar strength, it is not clear the data supports their finding that '*the more immediate and simple people's writing, the lower they rate themselves on Openness*.'

Secondly, they found individuals who are less extraverted, write more on 'Making distinctions.' In this study factor 1 does correlate somewhat negatively with Extraversion, as do many of its constituent variables. There are similar loadings here as for Pennebaker and King's constituents, though the most significant of these, Discrepancies, did not actually load on the original 'Making distinctions' factor.

Their third observation is that 'Making distinctions' has a negative correlation with Conscientiousness. Two of the variables that make up 'Making distinctions' do correlate negatively in this study, but on the whole, the factor correlates positively. However, this correlation is particularly weak, and the correlation directions of the variables is inconsistent with their factor loadings. In fact both this weakness and inconsistency is found throughout Conscientiousness.

Despite many agreements, there are also a number of significant differences in the

correlations found in the studies. The most obvious difference is that of Agreeableness and 'Immediacy.' Pennebaker and King found people who were highly Agreeable were more likely to be immediate in their writing, but in this study the opposite is true. Since both forms of writing are considered personal, one difference between them is their intended audience: Pennebaker and King's essays have no particular intended audience, certainly not beyond the course in which they were gathered; blogs however can be read by everybody.

Given the nature of Agreeable people to be more accommodating, one would imagine them to always be immediate in their writing style, and particularly with a larger audience. This is not the case here however. Whilst there seems no reason for this as yet, this point will be considered further when investigating with Heylighen and Dewaele's measure of contextuality later in the thesis.

There are other observations which can be made. In this study, the 'Making distinctions' factor correlates significantly with Neuroticism. Pennebaker and King found mostly positive correlations for the factor and its constituents, but these were all very weak. This study finds much stronger links, most significantly again with Discrepancies. Gill however found a stronger but negative link, suggesting to him that the writing of more stable individuals contains fewer features one would consider 'Making distinctions.' This study found that it should be the *less* stable individuals that do so. The results of this study also show a significant negative correlation between 'Making distinctions' and Agreeableness. All the variables within the factor correlate negatively, and two of them significantly. Pennebaker and King's results had a similar leaning, but with very little strength. This suggests that 'Making distinctions' is more important in blogs for distinguishing individuals than in the previous genres.

Along with more general factor observations, Pennebaker and King also highlighted correlations with specific LIWC variables. As one might expect they found Neuroticism to correlate positively with negative emotion words (r=.16) and negatively with positive emotion words (r=–.13). This study found similar results (r=.16 and r=–.04 respectively). Gill however found opposing values, for which he proposes that "*differences in topic assigned for the experimental task affects the expression of emotion*."

Pennebaker and King also found, as they expected, that Extraversion correlated with positive emotion words (r=.15) and total social processes (r=.12), while Agreeableness linked positively with positive emotions (r=.07) and negatively with negative emotions (r=–.07). The results of this study were similar, (r=.16, r=.24, r=.07, r=–.20 respectively).

5.3 LIWC and Content Differences

So far this chapter has only explored general factors derived broadly from text categories that were selected via a set of criteria. They do not necessarily describe which features are most important for detecting a particular personality trait. Neither can they reveal much of use for personality rich text generation. Therefore, this section will be looking at all LIWC categories individually. This will allow the words and categories that could lead to detection and projection of personality traits to be identified. Firstly, a correlation analysis was carried out to investigate which of the LIWC variables can be associated with which personality dimensions. Secondly, multiple regression was performed on those correlating variables to see how much, if any, of the variance in personality could be explained.

5.3.1 Correlation of the LIWC with personality traits

5.3.1.1 Method

Pearson correlation analysis was used to reveal any relationships between the mean LIWC category scores of each author (as calculated for the previous stage) and their personality scores. In reporting the results, they will be compared to the findings of Gill - as much as is possible given the different personality models.

5.3.1.2 Results

As per Gill, the results of the analyses which showed a significant relationship at the $p<.1$ level are shown for Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness (tables 5.11, 5.12, 5.13, 5.14, and 5.15 respectively).

LIWC Variable	Example words	*r*	*p*
Discrepancies	*should, would, could*	.339	.004
Job/work	*employ, boss, career*	.275	.020
Anxiety	*nervous, afraid, tense*	.255	.032
Future tense	*will, might, shall*	.232	.052
Eating/drinking	*eat, swallow, taste*	.230	.054
Humans	*boy, woman, group*	–.219	.066
Physical states	*ache, breast, sleep*	.198	.098

Table 5.11: Correlation of Neuroticism scores with LIWC variables

It is clear from the tables that Agreeableness, Extraversion and Openness correlate with at least 10 variables, with Neuroticism not far behind. More than half of those variables that correlate with Extraversion and Openness are significant at the $p<.05$ level. However, Agreeableness and Neuroticism have fewer variables reaching significance (4 and 3 respectively) while Conscientiousness only correlates with 3 variables, and only one of those at $p<.05$.

Neuroticism High neurotics, as one would intuitively expect, worry a lot, and their language shows a strong correlation with Anxiety words. This could also tie in with their increased use of words relating to their Job, their Physical states — particularly their dietary habits — and their tendency to talk in the Future tense. They also talk less about other people in an abstract sense. Most significantly, they greatly use Discrepancy words, which includes the terms of *needs*, *wants* and *wishes*.

In terms of individual categories none that Gill determined to correlate with Neuroticism correlated here. However, the categories imply similar ideas. Gill found a negative correlation with the Past tense, which suggests that more neurotic individuals talk less about the past. He also found they use more First-person and less Second-person pronouns, suggesting as previously hypothesised (section 2.1.2.1) that they talk about themselves far more than they talk about other people. Gill also found they used more Inclusive words, talked about Grooming, but tended not to Swear.

LIWC Variable	Example words	*r*	*p*
Occupation	*work, class, boss*	–.333	.004
Achievements	*try, goal, win*	–.264	.026
Discrepancies	*should, would, could*	–.251	.035
School	*class, student, college*	–.247	.037
Humans	*boy, woman, group*	.242	.042
TV	*TV, sitcom, cinema*	–.239	.045
Social Processes	*talk, us, friend*	.238	.046
Communication	*talk, share, conversation*	.234	.050
Grooming	*wash, bath, clean*	.219	.066
Present tense	*walk, is, be*	.200	.095

Table 5.12: Correlation of Extraversion scores with LIWC variables

Extraversion Correlations with Extraversion allows us to make two significant observations. Firstly there are strong negative correlations with the Occupation category and its sub-categories School and Achievements. This suggests that low Extraverts, or Introverts, are more likely to talk about their day in terms of work or successes than Extraverts, who appear less concerned with such things. Secondly Extraversion also shows a positive correlation with Social processes along with its subcategories Communication and Humans, suggesting that Extraverts are more social, or at least talk more about being social.

High Extraverts also talk more about Grooming, and less about TV. They also talk more in the present tense, and like low neurotics, use fewer Discrepancy terms.

When studying Extraversion, Gill only found one of the categories found here to correlate significantly, although in the opposite direction. He found that it was low Extraverts who talk more about Grooming. Perhaps his most intuitive finding was a positive correlation with word count, suggesting that Extraverts write more in their e-mails. He also found a positive correlation for Anxiety, Positive feelings, and the super-category of Affective processes, suggesting that the more Extravert individuals discuss emotions and feelings more frequently.

LIWC Variable	Example words	*r*	*p*
Words > 6 letters		.290	.014
Positive feelings	*happy, joy, love*	.262	.027
School	*class, student, college*	–.255	.032
Occupation	*work, class, boss*	–.251	.035
Grooming	*wash, bath, clean*	.250	.036
Inclusive	*with, and, include*	.249	.036
Prepositions	*on, to, from*	.236	.047
Negations	*no, never, not*	–.222	.063
Assents	*yes, OK, mmhmm*	–.207	.083
Communication	*talk, share, conversation*	.200	.094

Table 5.13: Correlation of Openness scores with LIWC variables

Openness People who score higher for Openness, like high Extraverts, tend not to talk about their Occupation or School based activities, and do talk about Grooming. They also use Communication words along with words related to Positive feelings and Inclusion. They also have a higher tendency to use long words, but clearly also short words with their propensity for using Prepositions. They also show a negative correlation with the use of Assenting or Negating terms.

Agreeableness The strongest correlation found for Agreeableness was, once again, Discrepancy words. Combined with previous results this suggests that more Discrepancy words are used by people who score low on Agreeableness and Extraversion, but high on Neuroticism. Along with Discrepancy terms, other Cognitive process subcategories correlate negatively. Both Tentative and Certainty terms are used more by people who are low on Agreeableness.

Since Gill's personality model was the EPQ-R, he does not have scores for Openness, Agreeableness and Conscientiousness. However, as discussed in section 2.1.1 Psychoticism is seen by many to have a negative relationship with Agreeableness and Conscientiousness, allowing a degree of comparison. This is first seen in his positive findings for Cognitive process terms along with, more specifically, Certainty words.

LIWC Variable	Example words	*r*	*p*
Discrepancies	*should, would, could*	–.290	.014
Words > 6 letters		.262	.027
Articles	*a, an, the*	.255	.032
Negations	*no, never, not*	–.245	.039
Motion	*walk, move, go*	–.219	.067
Swearing	*damn, fuck, piss*	–.210	.079
Anger	*hate, kill, pissed*	–.206	.085
Certainty	*always, never*	–.205	.087
Body states/symptoms	*ache, heart, cough*	–.205	.087
Grooming	*wash, bath, clean*	–.205	.087
Negative emotions	*hate, worthless, enemy*	–.202	.091
Tentative	*maybe, perhaps, guess*	–.198	.098

Table 5.14: Correlation of Agreeableness scores with LIWC variables

Another commonality is the use of Swearing, Anger words, and more generally talking about Negative emotions. While these are all positively correlated with Psychoticism, they are negatively linked with Agreeableness as previously anticipated (section 2.1.2.4). This study also found a negative correlation with Negations, further suggesting that more Agreeable individuals tend to be less negative. One point where the studies differ is in the pattern for Motion words, which was found to correlate negatively for both Agreeableness and Psychoticism, though there is no clear personality based reason for this.

Highly Agreeable individuals use less words relating to Bodily states, and more specifically Grooming. They also use more Articles, and longer words.

Conscientiousness Very little correlates with any significance with Conscientiousness. There is a negative link to words associated with Friends, but the strongest is a negative correlation with Death words. This is not as random as it appears since Gill found Death words made a positive correlation with Psychoticism in his study. Word count is also negatively correlated, which could suggest that highly Conscientious peo-

LIWC Variable	Example words	*r*	*p*
Death	*dead, burial, coffin*	–.323	.006
Word Count		–.216	.070
Friends	*pal, buddy, coworker*	–.215	.072

Table 5.15: Correlation of Conscientiousness scores with LIWC variables

ple spend less time blogging, or at least writing shorter posts.[8]

5.3.2 Multiple regression of the LIWC

Multiple regression is an analysis technique which can reveal the structure of a set of variables. Though it is similar to correlation, it shows the degree to which a dependent variable can be explained by one or more independent variables. Several methods can be used to select the dependent predictor variables: in this study, a stepwise analysis is considered most suitable since variables will be entered if they show a significant relationship with the independent variable, though they will be removed if they do not show a significant enough correlation.

As in Gill's study, personality factor is considered the dependent variable, though this does not mean that personality is caused by the linguistic features. This form of analysis was chosen for the following reasons:

1. Of greatest interest is the overall realisation of personality through language, and so the combination of linguistic variables that gives the best sense of particular personality dimensions is important. Therefore it is how these features contribute to the suggestion of personality, how use of these feature in language might lead a reader to make a hypothesis as to the writer's personality, that is of interest. How each personality dimension leads to the uses of certain linguistic features is not to be the outcome of this analysis.

2. Having the personality traits as the independent variables would mean an analysis for each linguistic variable. This use of multiple measures leads to greater

[8]Since LIWC scores are averaged across all texts of an author, Word count reflects the average personal text length.

occurrences of Type I errors and so is undesirable.

3. Causation or directionality of the analysis is not inherent to the technique, but is imposed in order to aid interpretation. Therefore, statistically speaking, the direction of the relationship is not important.

Therefore, in the multiple regression analyses here, the personality dimensions shall be the dependent variables and the linguistic features the independent.

5.3.2.1 Basis for Variable Selection

For each personalty dimension (Neuroticism, Extraversion, Openness, Agreeableness and Conscientiousness) a stepwise regression was performed on the LIWC variables that reached $p < .1$ significance in the correlations, with variables entered in order of strength (as can be seen in tables 5.11, 5.12, 5.13, 5.14, 5.15).

Multiple regression requires that the independent variables truly be independent of one another, so further pre-selection was required. Generally the most specific LIWC variable was chosen (eg. Anger and Sadness over Negative Emotions and Affective Processes). However, should the sub-category variable not be retained in the regression equation, the super-category variable would replace it in a re-run of the analysis.

In addition to these statistical considerations, in keeping with Gill's analyses, there are further selection criteria to apply to the linguistic variables.

- **Topic Independence** The most obvious limit to generalising texts is their topic. In their factor analysis, Pennebaker and King (1999) noted this and so excluded variables known as personal or current concerns. Similarly, in order to distance from topic and be concerned more with processes, such words shall be omitted.

- **Genre Independence** Exclusion of topic words allows for abstraction from the content of the text, but results may still be specific to the genre, in this case to blogs. Pennebaker and King also addressed this issue of linguistic reliability across genre. In their factor analysis they chose only those variables which occurred most consistently across the different texts in their validation studies: those variables scoring an average Cronbach's α greater than .60. This criterion shall also be adopted here.

- **Independence of Language Sparsity** Pennebaker and King required linguistic variables to occur with a frequency of at least 1%. By specifying a minimum usage level, this ensures more accurate characterisation across different texts, particularly shorter texts. It removes anomalies from having an impact on results.

  However, the use of such anomalies — those occurring only once being known as hapax legomena — may be characteristic of a particular personality type, and thresholding removes the possibility of discovering these items. However, for reliable comparability at this stage, it is still suitable to limit sparse items.

5.3.2.2 Method

As in the previous correlation studies, the mean LIWC dictionary scores across all an author's texts are used. The data is then subject to the following analyses:

1. **Regression of all variables** Those variables selected by the statistical criteria above are included. The results can be seen in table 5.16.

2. **Regression allowing for topic independence** Those variables considered to be concerned with topic are removed and the analysis re-run. The results are shown in table 5.17.

3. **Regression allowing for genre independence** Variables which did not reach a level of .60 for linguistic reliability in Pennebaker and King's validation study are dropped at this stage. The results of the further analysis are in table 5.18.

4. **Regression without sparse data** One further regression analysis was conducted with those only those variables remaining that had a mean usage of at least 1%. See table 5.19.

5.3.2.3 Results

Regression of all variables The results of entering all variables is shown in table 5.16.

Highly neurotic individuals use more terms relating to their Jobs and Physical states along with more Discrepancy words. Gill found that they used more Inclusive words

and more First-person pronouns. This last finding, along with the finding here that they are less likely to use terms relating to people in general, supports the theory that high neurotics tend to talk more about themselves. While Gill's equation accounts for 11% of the variance, this study's accounted for 30%.

High Extraverts are found to talk more in the Present tense and use more words relating to Communication processes, but they use less words relating to Achievements and Discrepancies. Gill found a positive use of Affective process words, words relating to feelings and emotions, and a negative use of Sporting words. He accounts again for 11% of the variance, while this study accounts for 41%.

People who scored high in Openness tend to use longer words, more Inclusive words, and talk more about Positive feelings. They also tend to use fewer Negations and talk less about School. The equation here explains 35% of the variance.

Highly Agreeable people use less terms relating to Discrepancies and Body states, while highly Conscientious people don't talk about Death. This is similar to Gill's finding that highly psychotic individuals do discuss matters of Death. Though fewer variables reach the equation of the last two personality dimensions, the variance accounted for is still reasonable: 17% and 11% respectively.

Regression allowing for topic independence Once all topic related categories are removed, the remaining equations can be found in table 5.17.

Whilst Discrepancies remains in the equation for Neuroticism, albeit slightly more strongly correlated, removing concerns of Jobs and Physical states has resulted in Human terms dropping from the equation and Anxiety terms being included, another aspect of high neurotics that is expected to be reflected in their language. These now account for 18% of the variance. Gill's results did not change at this stage.

Removing Achievements from the Extravert equation caused Communication words to drop out. The strength of the correlations changes slightly in the remaining variables, and they account for 28% of the variance. Gill's results changed entirely, with high word count and a low use of numbers being representative of high extraverts. These accounted for 8% of the variance.

The remaining variables found for Openness were found with slightly different correlation strengths, and the variance accounted for dropped to 29%. Dropping topic

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	Discrepancies	.32	.003		
	Job	.28	.008		
	Physical states	.29	.009		
	Humans	–.24	.032	**.30**	**.000**
E score	Achievements	–.31	.002		
	Discrepancies	–.68	.000		
	Present Tense	.53	.000		
	Communication	.23	.040	**.41**	**.000**
O score	Words > 6 letters	.30	.006		
	Positive feelings	.35	.002		
	School	–.25	.016		
	Negations	–.25	.027		
	Inclusive	.22	.033	**.35**	**.000**
A score	Discrepancies	–.37	.002		
	Body States	–.30	.011	**.17**	**.002**
C score	Death	–.32	.006	**.11**	**.006**

Table 5.16: LIWC multiple regression analysis (all variables) with personality scores

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	Discrepancies	.34	.003		
	Anxiety	.26	.020	**.18**	**.001**
E score	Discrepancies	–.63	.000		
	Present Tense	.60	.000	**.28**	**.000**
O score	Words > 6 letters	.28	.013		
	Positive feelings	.35	.002		
	Inclusive	.25	.019		
	Negations	–.26	.027	**.29**	**.000**
A score	Discrepancies	–.29	.014	**.08**	**.014**
C score	*none*				

Table 5.17: LIWC multiple regression analysis (topic controlled) with personality scores

concerns left just Discrepancy terms in the Agreeableness equation for 8%, but no variables were retained for Conscientiousness.

Regression allowing for genre independence Once all categories with reliability below .60 are removed, the resulting equations can be seen in table 5.18.

Removing Anxiety as an unreliable term, left just Discrepancy terms in the equation for Neuroticism, account for 12% of the variance. The remaining equations remained the same, as did those of Extraversion and Neuroticism in Gill's study.

Regression without sparse data Once all categories with frequency below 1% are removed, the final equations can be seen in table 5.19.

The equations remained the same with the exception of Openness. Positive feeling words were dropped due to low frequency, and this resulted in just long words and Inclusive terms appearing in the final equation though with stronger correlations than

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	Discrepancies	.34	.004	**.12**	**.004**
E score	Discrepancies	–.63	.000		
	Present Tense	.60	.000	**.28**	**.000**
O score	Words > 6 letters	.28	.013		
	Positive feelings	.35	.002		
	Inclusive	.25	.019		
	Negations	–.26	.027	**.29**	**.000**
A score	Discrepancies	–.29	.014	**.08**	**.014**
C score	*none*				

Table 5.18: LIWC multiple regression analysis (genre controlled) with personality scores

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	Discrepancies	.34	.004	**.12**	**.004**
E score	Discrepancies	–.63	.000		
	Present Tense	.60	.000	**.28**	**.000**
O score	Words > 6 letters	.33	.004		
	Inclusive	.29	.011	**.17**	**.002**
A score	Discrepancies	–.29	.014	**.08**	**.014**
C score	*none*				

Table 5.19: LIWC multiple regression analysis (sparsity controlled) with personality scores

before. They now account for 17% of the variance.

At this final stage of analysis, Gill found that no variables were retained for the equation of Extraversion.

5.3.2.4 Summary and discussion

The most obvious conclusion to be drawn, particularly when looking at the most strict multiple regression (table 5.19), is that Discrepancy words are strongly related to personality in blogs. Discrepancy words includes modal terms such as *could*, *would* and *should*, *if*, and variant terms relating to *wishes*, *wants*, *needs* and *hopes*. As related to blogs this could represent two things. First, it could be indicative of bloggers talking about their dreams, their aspirations for the future. Alternatively it could indicate people discussing their dissatisfaction with their life, pondering the things they should have done, considering all their life is lacking and what they need to fulfil it. Intuitively, judging by the personality types for which Discrepancy words are important — high neurotics, low Extraverts and people scoring low on Agreeableness — it is the

latter hypothesis which seems more likely to explain the result.

Extraversion is also linked strongly with the use of the Present tense. This seems intuitive since Extraversion is considered the action trait, so high use of present tense verbs in text concerning one's daily life is to be expected. The strongest results found in the final regression are for Extraversion, with 28% variance explained by just two linguistic dimensions.

That nothing explained Conscientiousness comes as no surprise from the weak correlations previously seen. Similarly, that using Words of greater than six letters is important for Openness, often considered the factor of intellect, is not at all surprising. That Inclusive terms are almost as strong, is perhaps less expected. Inclusive terms such as *with* and *include* suggest social functions, which one would more expect of the Extraversion dimension.

Though results may ultimately account for more variance than Gill, it is considerably less than the first stage of the multiple regression. Whilst Pennebaker and King's constraints for variable selection do make sense, they are particularly limiting, reducing considerably the set from which to choose. Excluding those variables considered 'sparse' with a mean usage of $<1\%$ not only rules out instances of hapax legomena as discussed earlier, but as Gill highlights use of certain emotion word categories. Expression of emotion could be a markedly important difference within personality type. Discrepancy words, seemingly so important in the blog genre, only have a mean usage of 1.96% (table 5.1). Likewise, many of these terms are ruled out for being unreliable across Pennebaker and King's various genres, but they will clearly be more important in some, personal diaries and e-mails for instance, than others.

5.4 MRC and Psycholinguistic Differences

The MRC Psycholinguistic Database was originally constructed to aid researchers in creating test materials, by providing information on a large collection of words (Coltheart, 1981; Wilson, 1987; for more background see section 2.7.1.2). This data was empirically derived, which differs from the human judgement of psychological categories that created the LIWC. This section adopts a similar methodology to the pre-

vious section (correlation followed by multiple regression) in order to investigate differences in the psycholinguistic properties of language used by different personality types.

5.4.1 General methodology for the MRC

As with the previous LIWC work, analysis is carried out with the spell checked version of the blog corpus. In addition, in order to disambiguate word senses the corpus must be tagged for parts-of-speech. This was carried out with Ratnaparkhi's MXPOST maximum entropy tagger (Ratnaparkhi, 1996). This tool was chosen for consistency with Gill (2004), who chose it after hand evaluation indicated it was the best for his data. Referring briefly to the concerns discussed in section 2.7.3.3, by using the same tagger as Gill, any errors it introduces into this study will be similar to those of the e-mail corpus, and results of the two studies should remain comparable.

Using Gill's suite of programs the POS-tags of MXPOST are first transformed down to the smaller set of ten used by the MRC. Each word-tag pair is then looked up, and the psycholinguistic information is calculated for each of the measures listed in section 2.7.1.2. As with the LIWC analysis, each of the 1854 personal files were analysed separately, and average scores calculated for each author.

5.4.2 Correlation of the MRC with personality traits

5.4.2.1 Method

Pearson correlation coefficients were calculated for each of the five personality scores with the psycholinguistics properties calculated for each author.

5.4.2.2 Results

The results of simple correlations (again, $p < .1$) can be found in tables 5.20, 5.21, 5.22, 5.23 and 5.24. Correlations will be reported by dimension and compared to Gill's results.

MRC Variable	*r*	*p*
Percentage of Digits	–.321	.006
Thorndike-Lorge freq. PercCapt.	.242	.042
Age of Acquisition StDev	.223	.061
Poetic word count	.218	.068

Table 5.20: Correlation of Neuroticism scores with MRC variables

MRC Variable	*r*	*p*
Archaic word count	.333	.005

Table 5.21: Correlation of Extraversion scores with MRC variables

Neuroticism The strongest relationship with Neuroticism (table 5.20) is the percentage of digits used in text. As strange as this may first sound, it ties with Gill's finding of both percentage and number of digits used correlating with Neuroticism. The Thorndike-Lorge frequency relates to a written frequency list derived by Thorndike and Lorge in 1944. While Gill found a positive correlation with mean verbal frequency, here only a correlation with the percentage of terms captured by the dictionary is found. This can only suggest that higher neurotics use more words than can be found in the Thorndike-Lorge dictionary, but says nothing about how (in)frequent those words are. The Age of acquisition result also suggests merely something statistical about the data, rather than the authors. A positive correlation with the standard deviation means higher Neurotics use more varied terms when it comes to their considered age of acquisition. There is also a positive correlation with the count of 'Poetic' words. That is, those words which were defined as being of use purely in *poetry or other contexts with romantic connotations*. Raw counts are less reliable than percentages for detecting patterns, because they do not take account of text length. Gill also found a positive link with not only the mean Concreteness of words used by neurotics but also the standard deviation therein.

Extraversion The only variable to correlate with Extraversion (table 5.21) is the count of 'Archaic' words: Extraverts use more word senses considered *restricted to special contexts such as legal or religious use, or used for special effect* than Introverts. Again however, since this is the count of words and not percentage, there can be a length effect, so the finding may not be reliable. Gill again found effects for Concreteness, the number of words captured and counts of words considered 'Standard' and 'Dialect' by the Dolby dictionary.

Openness As can be seen in table 5.22, there are a large number of factors with which Openness correlates. Many of them are the statistical side measures of the various psycholinguistic properties, though these can still be of interest: correlation with standard deviation can be interpreted as one end of a personality scale having more variability on a factor than the other.

Of most interest is the positive correlation of both mean and standard deviation scores on both the Kucera & Francis and the Thorndike-Lorge frequency scales. Not only do subjects scoring high on Openness use more words considered frequent in written language, but they use a great variety of frequent and infrequent words. They also use more words considered frequent in the Brown verbal list. The greater use of more frequent words lends itself easily to the idea that the more Open the individual, the shorter the words they use, since certainly the most common words are function words. However, a positive correlation with the mean number of letters, phonemes and syllables in a word suggests otherwise. Indeed, using the LIWC found Openness to correlate significantly with use of words greater than six letters. This length finding is much more in line with the consideration of Openness as the factor of intellect: higher scorers of Openness would be more expected to use longer, less frequent words.

On the other hand, the finding of correlations with the standard deviations again suggests that subjects toward the high end of the scale are simply more variable in the language they select. Indeed, the average coverage of the frequency measures across the blog corpus is between 80-90% (SD 6% for each measure). This suggests that the remaining 10-20% of words for which the MRC has no frequency data are perhaps the longest words, which would be expected to be of a much lower average frequency.

People of high Openness also use fewer words considered 'Nonce' (relative usage

MRC Variable	*r*	*p*
Pavio Meaningfullness PercCapt	.364	.002
Familiarity StDev	.355	.002
Age of Acquisition StDev	.337	.004
Pavio Meaningfullness StDev	.325	.006
Percentage of Digits	–.295	.013
Kucera & Francis written freq. Mean	.282	.017
Age of Acquisition Mean	.280	.018
Kucera & Francis written freq. StDev	.263	.027
Thorndike-Lorge freq. Mean	.258	.030
No. of letters Mean	.247	.038
No. of phonemes Mean	.247	.038
Thorndike-Lorge freq. StDev	.243	.041
Nonce word count	–.243	.042
No. of phonemes StDev	.240	.043
No. of letters StDev	.239	.045
Nonce word perc	–.229	.055
Brown verbal freq. Mean	.225	.059
Pavio Meaningfullness NosCapt	.222	.063
Kucera & Francis no. of samples StDev	.220	.065
No. of syllables Mean	.215	.072
No. of syllables StDev	.210	.079
Imagability StDev	.207	.084
Colloquial word count	–.204	.088

Table 5.22: Correlation of Openness scores with MRC variables

MRC Variable	*r*	*p*
No. of syllables StDev	.305	.010
Pavio Meaningfullness PercCapt	.272	.022
No. of phonemes StDev	.270	.023
Kucera & Francis written freq. Mean	.250	.035
No. of letters StDev	.248	.037
Thorndike-Lorge freq. Mean	.247	.038
No. of syllables Mean	.232	.051
No. of letters Mean	.226	.059
No. of phonemes Mean	.217	.069
Kucera & Francis written freq. StDev	.199	.096

Table 5.23: Correlation of Agreeableness scores with MRC variables

since the percentage correlates) and 'Colloquial' (*use that is normally restricted to informal (esp. spoken) English*), along with fewer digits. They show higher scores, along with greater variability when it comes to age of acquisition of language. Openness also correlates positively with all aspects of the Pavio Meaningfullness data, except the important mean.

Agreeableness Results for Agreeableness (table 5.23) are similar to those of Openness. People who score high on Agreeableness use not only longer more frequent words, but also words of more varied length and frequency.

Conscientiousness The results in table 5.24 are the most interesting. For almost all of the psycholinguistic categories, the only figure to correlate is the number captured by the dictionary and all with almost identical negative scores, despite their different sizes. The collective result of this is that highly Conscientious individuals use fewer of the words in any of the MRC dictionaries. This is corroborated by the negative correlation with the overall dictionary capture measure.

According to some of the more significant figures, highly Conscientious people not only use less 'Nonce', 'Dialect' and 'Poetic' words, but also less 'Standard' words. In

MRC Variable	*r*	*p*
Nonce word count	–.375	.001
Nonce word perc	–.304	.010
Dialect word count	–.246	.039
Age of Acquisition NosCapt	–.238	.046
Number of ASCII characters	–.237	.047
Poetic word count	–.231	.052
TotInString	–.226	.058
Number of Words	–.223	.061
Thorndike-Lorge freq. NosCapt	–.223	.060
Words captured by dictionary	–.223	.062
No. of phonemes NosCapt	–.222	.063
Age of Acquisition StDev	–.222	.063
Kucera & Francis written freq. NosCapt	–.221	.064
Kucera & Francis no. of categories NosCapt	–.221	.064
Kucera & Francis no. of samples NosCapt	–.221	.064
Pavio Meaningfullness NosCapt	–.221	.064
Brown verbal freq. NosCapt	–.220	.065
Familiarity NosCapt	–.219	.067
Concreteness NosCapt	–.219	.067
Imagability NosCapt	–.219	.067
No. of syllables NosCapt	–.218	.068
Toglia & Battig Meaningfullness NosCapt	–.218	.068
Standard word count	–.215	.072

Table 5.24: Correlation of Conscientiousness scores with MRC variables

fact they use fewer words and characters in total. This finding does relate to the lower word count found using the LIWC; however, these are only marginally significant results, which ties with the fact that no significant length effects were found previously (section 3.4.5.1).

5.4.2.3 Discussion

In this analysis the significance of some of the Dolby dictionary categories stands out, the strongest correlation in the study being 'Nonce' word count with Conscientiousness. Of Dolby's word categories, which define almost 90,000 words, 'Nonce' words account for just 0.04%, or 33 words. Not only is the number of words in that class low, but of the 71 subjects in the study, there were only 4 subjects and 7 occurences of 'Nonce' words (0.0005% of the words within the blog corpus). For such a significant result these are surprising figures.

On closer inspection of each occurrence, all 7 could be explained by tagging errors made by the part-of-speech tagger. With this in mind, a second manual analysis was carried out on the instances of the 183 'Poetic' words (0.2% of the Dolby dictionary). These again could wholly be explained as anomolous, resulting solely from errors. The same followed for those words classed as 'Archaic' and 'Colloquial.'

The reason these categories are so susceptible to tagging errors is in the specific words they each contain. While each category contains many unusual words that may well be 'Archaic' or 'Poetic', they also have more common words with an uncommon word sense. From this result, it can be concluded that the results from the Dolby dictionary analysis cannot be guaranteed. Therefore, these categories will be excluded from any further analysis.

Another interesting observation is the number of correlations with the purely statistical aspects of each category. This says a more perhaps about the coverage of each category than the individual differences.

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	Percentage of Digits	–.32	.006	**.10**	**.006**
E score				**.00**	
O score	Pavio Meaningfullness PercCapt	.42	.000		
	Percentage of Digits	–.36	.001	**.26**	**.000**
A score	No. of syllables StDev	.30	.010	**.09**	**.010**
C score	Age of Acq. Nos Capt	–.23	.046	**.06**	**.046**

Table 5.25: MRC multiple regression analysis with personality scores

Note: without Dolby categories

5.4.3 Multiple Regression of the MRC

5.4.3.1 Method

Properties derived from the MRC dictionary that showed a correlation with significance of $p < .1$ (with the exception of the excluded Dolby Categories) were entered into a stepwise multiple regression analysis for the corresponding personality traits. This will show which, if any, of the MRC variables best explain the variance in each of the personality dimensions.

5.4.3.2 Results

The resulting equations of the regression analysis can be seen in table 5.25.

While Gill found Extraverts to use less concrete language, and high neurotics to use more concrete and frequent language, there is little in these results as interesting. According to the results here, high Neurotics use less digits (10% of the variance), while due to the exclusion of Dolby's word categories no variables were retained for Extraversion. High Openness scorers also use less digits than their low counterparts,

along with a greater relative frequency of words that were in the Pavio Meaningfullness dictionary. These explain 26% of the variance within Openness. Highly Agreeable individuals have a greater range of number of syllables in their words (9%), while Conscientious individuals use fewer words for which Age of acquisition data was available (6%).

The results are disappointingly weak, with the exception of Openness, for which 26% of the variance is accounted for. However, it is hard to believe that the percentage of digits and the percentage of words which were found to have data in one specific dictionary are particulary meaningfull. There certainly seems to be no explanation for these results within the nature of the traits themselves.

5.5 Criticism of the Dictionary-Based Approach

Prior to the implementation of dictionary approaches, section 2.7.3.2 discussed a number of criticisms directed at them. This section returns to these in light of the current results.

The first criticism was a matter of coverage. This can most obviously be directed at the LIWC since it has only around 2000 words and word stems. However, this is also a fair criticism of the MRC, since data is not available in all categories for all words. For example, there is only Age of acquisition data for 3503 words, while Pavio Meaningfullness data is only held on 1504.

Many of the significant results found in the MRC concerned the standard deviation of a variable along with the percentage and number captured by the dictionaries. If the number of words being covered by dictionaries is the most interesting result, then there has not nearly been enough coverage: all that can be concluded is that some people use more words for which there is psycholinguistic information than others.

Another issue related to coverage is that dictionaries have to be made, and are often created with a specific use in mind. Using the LIWC for example, there can be no analysis of any words which are not included in the predetermined categories.

The second criticism made was one of recall, and reflects the accuracy with which dictionary searches can be made. Without a large scale hand analysis it is difficult

to know how robust the LIWC categories are. It is possible for mistakes to be made certainly: *managed* as in 'I managed to finish in time' is not in any dictionary; the word stem *manag** is in the Job dictionary however, fitting to words such as 'manager' and 'management.'

Criticism has also been made previously of the errors introduced by using part-of-speech taggers. Again, without a hand analysis it is unclear to what extent there are errors. However, dictionaries that use parts-of-speech to disambiguate word senses are clearly susceptible to inaccuracies in recall due to these errors. This was clearly illustrated by the errors in the Dolby figures reported, discussed in section 5.4.2.3.

A further criticism that can be levelled at dictionaries, and particularly some categories of the MRC, is one of age. The Kucera and Francis frequency data was derived in 1967, while Thorndike and Lorge derived theirs in 1944. Language use changes over time, and linguistic resources can become out of date. In a study of language change over 100 years of National Geographic magazine, in was shown that there was practically no change in language use in 40s, but the greatest change across the century occurred in the 50s (Juola, 2003). Despite the high specificity of this study, it is used here to illustrate the simple point that language changes; frequency lists derived in 1944 may not be totally accurate in 2005.

While these criticisms are made following the use of the LIWC and MRC, they can be applied to dictionary approaches in general. However, this should not be considered a criticism of the MRC itself however: it was not designed with content analysis in mind. Criticism is only levelled at this application of the MRC.

5.6 Summary

This chapter reported on top-down approaches to linguistic analysis for investigating personality differences. The results found here are derived by using dictionaries to determine significant relationships between individual traits and linguistic categories and properties.

Using preselected dictionary categories, it has proved possible to replicate a factor analysis of language (Pennebaker and King, 1999; cf. Gill, 2004). Despite differences

in text genre, and some minor loading variations, the factors of ‘Making distinctions’, ‘Immediacy’ and ‘the Social past’ were successfully found in blogs. This suggests that the factors are robust, and are certainly present in various forms of personal writing.

This study found differences in correlations between personality factors and the linguistic factors to those found in previous studies. Individuals who score lower on Openness, or Agreeableness or higher on Extraversion have a more immediate style of writing. High Extraverts, along with high scorers of Openness and low scorers of Conscientiousness talk more of ‘the Social past’. High neurotics, Introverts, and less Agreeable individuals use more language concerning ‘Making distinctions’.

The full list of LIWC categories was then used to explore individual differences. Upon regression, under perhaps overly strict conditions, very few of the categories that had shown relationships with traits remained in the equations which explained only a small degree of the variance: across the five dimensions results ranged from 0–28%. Discrepancy terms proved important in determining individual differences in blogs, with high neurotics and low scorers of Extraversion and Agreeableness using more Discrepancies.

Adapting the MRC Psycholinguistic database proved even less fruitful. The only categories retained for regression reflected nothing more than coverage statistics, along with the relative frequency use of digits. Further to this, one segment of the data proved susceptible to errors introduced by part-of-speech tagging, rendering it unusable for analysis.

As discussed at the end of this chapter, there are a number of criticisms that can be levelled at the specific tools used here, and dictionary-based approaches in general. In the next section, data-driven approaches are used: both to further investigate the relationship between personality and language, and to investigate if they can explain more variance than the poor overall performance reported here.

Chapter 6

Bottom-up Approaches to Personality Differences

This chapter follows the last by continuing the investigation into the relationship between personality and language. Where the last chapter used top-down approaches however, the methodology of this chapter is bottom-up. In the last chapter, dictionaries were used to look at the linguistic content of blogs: specific categories and features were the tools of study. Bottom-up approaches are those derived directly from data: they do not set out to investigate specific features; instead features are revealed by the investigations.

In this chapter a number of analytic techniques are used. Firstly, following Oberlander and Gill (2005) a frequency comparison technique is adopted to identify significant collocations for extreme personality sub-groups. Attention is then returned from groups to the individuals in a similar methodology adopted in the last chapter. After correlating these collocations directly with the raw personality scores, those that prove at least marginally significant are entered into a multiple regression analysis.

Subsequently, two unitary linguistic measures previously used to distinguish genres in chapter 3 are employed to investigate their ability at distinguishing between individuals: namely Heylighen and Dewaele's F-measure (2002) and the average rank approach to word frequently usage.

The chapter concludes by comparing the approaches of this chapter to those in

chapter 5. This is done by entering all variables that have been shown to correlate with the personality dimensions into a multiple regression analysis. This will not only show which of all the features used in this study are most useful for explaining variation in personality, but also which of the two methodologies is better.

6.1 Stratified Collocation Analysis

One common data-driven technique is n-gram analysis (Manning & Schütze, chapters 5 & 6, 1999). This has been used previously to determine distinctive collocations for individual differences (Gill, 2004; Oberlander and Gill, 2005). Findings from this work were discussed in section 2.2.1.5 of the literature review. Gill's e-mail corpus was stratified into groups scoring at the extreme ends of personality scales. Using log-likelihood comparisons on relative frequencies of n-grams within each sub-corpus, significant collocations for each trait were identified.

This is a robust approach that serves to submerge overly specific variation between individuals. However, while these collocations are representative of personality classes, it is also worth returning to individual differences to explore how these class-based results fare. This will be the topic of the next section. This section discusses in more detail the stratified log-likelihood comparison approach and reports the results.

6.1.1 Method

The N-gram analysis conducted here uses the stratified sub-corpora introduced in section 3.4.5. These were created by dividing the blog corpus into High and Low groups across the four personality dimensions of Neuroticism, Extraversion, Agreeableness and Conscientiousness. Due to the non-normal distribution of Openness there was no low group (see section 3.4.3 for more details). Due to the lack of two extreme groups, it is not possible to study Openness with this technique. Therefore, no collocations will be reported here for this trait.

There is also a neutral group consisting of those individuals who scored on the mid scale for the four personality dimensions. For each of the dimensions, the High and Low corpus will be compared to each other and the neutral group to investigate which

features are representative of each group.

Previous analyses (see section 4.2) lead to the conclusion that proper nouns can have overly disproportionate effects when comparing corpora. References to specific individuals are not reflective of a group; however, references to other people in general might be. To this end all proper nouns were replaced with a tag (NP1), and following Oberlander and Gill (2005), all references to days of the week were similarly replaced (NPD1). Proper nouns were identified from the CLAWS tagging of the WMatrix tool (Rayson, 2001, 2003) as used in section 4.2. In order to provide more general analysis, all punctuation was also collapsed into a single marker (<*p*>). Likewise, abstract tags were created to describe non-linguistic features of blogs: <*SOP*> and <*EOP*> were used to mark the start and end of individual blog posts; <*EMOT*> was used to mark the existence of an emoticon, or 'smilie'.

Collocations are calculated as two and three word n-grams, with the only cut-off being that features required a frequency $\geq$5 within a group. This is to ensure an accurate log-likelihood G^2 statistic (cf. Rayson, 2003). N-gram software (Banerjee & Pedersen, 2003) was used to identify and count collocations within a sub-corpus. Following this, G^2 is used to compute the significance of collocation use between corpora.

Robust collocations are identified by a three way comparison between the High and Low group of each personality dimension and the neutral group. For each feature found, its frequency and relative frequency are calculated. This in turn permits relative frequency ratios and log-likelihood calculations to be made between High-Low, High-Neutral and Low-Neutral (Oberlander and Gill, 2005).

A first pass at this analysis identified early problems. Compared to the analysis of the e-mail texts of Gill, blog sub-corpora contain similar or fewer authors, but between four and ten times as many words. The net result is that individuals can have a stronger influence over the subgroups. Even when the size of each file was capped at the mean word count plus two standard deviations (see section 3.4.5) the pervasive presence of verbose individuals is evident. Examining the output of the analysis described above allowed the author to identify n-grams which were recognised to be specific to certain individuals.[1] Excluding just those known specific n-grams is not a general solution,

[1]For example, there may be an author who ends all their 25 posts with 'That's all folks!'. It is likely that analysis of the sub-groups into which the author falls will provide significant results for the n-grams

since there may be others that are not recognised. Therefore a filter was applied to results requiring that each n-gram must be used by at least 50%[2] of the individuals within the subgroup of which it is reported to be representative. In limiting such individual influence, analysis is more robust.

6.1.2 Results

The results of the three-way stratified analysis are presented in two forms. The full tables of results can be found in the appendix (tables C.3 to C.13). Presented here are the condensed results, highlighting just which n-grams have been found to be representative of each sub-group. Due to the large size of the corpus, and despite the imposed limits of frequency and multiple author occurrence, many collocations have proved themselves to be significant. However, following Oberlander and Gill both those results with a critical value of 15.13 or greater, equivalent to reaching $p \leq 0.0001$, and those between 15.13 and 10.83 ($0.0001 < p \leq 0.001$) are reported.

Note that there are two kinds of features that can be associated with a High/Low subgroup: those which are over-used by the group and those which are under-used by the opposite group. However, for the neutral group it is noted which of the n-grams are underused to draw a distinction from those over-used. This is preferable to placing the underused n-grams in both high and low group lists.

There are many n-grams that appear significant which are merely different length representations of the same collocation. This is determined partly by subjective examination of n-grams. For example, it is highly likely that [*i love you*] and [*love you*], used with similar frequencies, are two instances of the same collocation. When this situation occurs, the more specific of the n-grams, the longest, will be discussed over the others. Of course, it is possible that shorter n-grams, such as [*i love*] can appear in other contexts than those of the longer n-gram. This will normally be clear if the n-grams place significantly apart in the order listed.

There are several types of distinctive collocation that can be identified and different

[*that's all folks*], [*all folks <p>*] and [*folks <p> <EOP>*], each with a count of 25. It is clear that those n-grams all come from the same stock phrase used by one individual and are not general to the group. Therefore they should be excluded.

[2] Conservatively rounded down in the case of an odd number of subjects.

ways in which they can be grouped. One of these is to report the LIWC categories into which the words fall. This will be done while keeping in mind the categories that have previously shown relationships with the personality traits (see section 5.3.1 for details.)

6.1.2.1 Neuroticism

The n-grams determined to be significant collocations for Neuroticism groups are shown in figure 6.1. There are considerably fewer level-1 n-grams for the high group than the low, especially considering that a number of those in the high group can be considered length-variants.

Consider first those LIWC categories associated with Neuroticism. The strongest correlation from section 5.3.1.2 was Discrepancies, higher use relating to high neurotics. The high group has two instances of n-grams with Discrepancy terms (‘if’ and ‘need’), while there are none in the low group. Words relating to Jobs, Eating and subsequently Physical states (also positively linked with Neuroticism) also appear in the high group but not the low (‘work’ and ‘eat’ respectively). There are no words in either group that reflect Anxiety or references to Humans.

Beyond the categories that showed correlations with Neuroticism, the LIWC can still be used to illustrate patterns within the distinctive collocations. High neurotics have a small number of collocations containing Pronouns — all First-person singular. Low Neurotics however, not only have more First-person but also Second and Third-person terms. The low group contains more instances of Prepositions but a similar number of Articles. The nominal marker ‘NP1’ also occurs several times in the low collocations though not at all in the high. This further supports the hypothesis that lower scoring neurotics are more likely to talk about other people, though not that high scorers talk more about themselves. Interestingly, the high collocations that contain ‘i’ are also the ones that contain Discrepancy terms. This pattern for pronouns and nominals was also found in e-mail (Oberlander and Gill, 2005).

Low neurotics have a high number of collocations containing Exclusive words. While ‘however’ is merely repeated in what appears to be the same context, ‘that’ is used in a different context each time. High neurotic collocations contain more verbs, reflected in Past and Present tense categories. They also use phrases relating to Time

High Neurotics

(1) [<p> <p>] [<p> <p> <p>] [<eop> <sop> so] [<sop> so] [*if i*] [*like a*]

(2) [*i need to*] [*at work*] [*instead <p>*] [*what a*] [*i need*] [<p> *still*] [*this year*] [*get a*] [*yesterday <p>*] [*need to*] [*to eat*] [*i remember*] [*write about*] [*and buy*] [*slowly <p>*]

Neutral

(1) [*ok <p>*] [<p> *ok*] [<p> *ok <p>*] [*i think*] [<p> *i*] [<p> *and*] [*and then*] [<p> *i want*] [<p> *and i*] [*all of*] [*fun <p>*]

Underuse: [*managed to*]

(2) [<p> *perhaps*] [*feel like*] [<p> *and then*] [<p> *NPD1*] [<p> *i'll*] [*going to*] [<p> *i think*] [*when i was*] [*i get*] [*about the*] [*go home*]

Underuse: [<p> *last*] [<p> *which*] [*and on*] [*find it*] [*this morning*]

Low Neurotics

(1) [*in NP1*] [*NP1 <p>*] [*that i*] [*and i*] [*a couple*] [*a couple of*] [*in NP1 <p>*] [<p> *as*] [<p> *NP1*] [*that he*] [*couple of*] [*NP1 <p> NP1*] [*to NP1*] [<p> *we*] [<eop> <sop> *i*] [<p> *however*] [<p> *you*]

(2) [<p> *NP1 <p>*] [*however <p>*] [<p> *however <p>*] [*mean <p>*] [*was that*] [*you see*] [*me <p> <eop>*] [<p> *i had*] [*the best*] [*is that*] [<sop> *i*] [*i mean <p>*] [*so <p>*] [<p> *now <p>*] [*i had*] [*to NP1 <p>*] [*see it*] [<p> *after*] [*a bit of*] [<p> *to*]

Figure 6.1: Distinctive collocations for Neuroticism sub-groups

Note: N-grams reaching (1) the 15.13 critical level ($p \leq 0.0001$) and (2) between 15.13 and 10.83 ($0.0001 < p \leq 0.001$)

in [*this year*] and [*yesterday <p>*].

Setting aside the LIWC, there are other patterns that emerge. Use of multiple punctuation marks seems exclusive to high neurotics, previously seen also in e-mails, as does post-initial 'so' (represented as [*<sop> so*]). Low neurotics on the other hand are more inclined to start their posts with 'i'. Low neurotics also appear to use more vague phrases: [*a couple of*] and [*a bit of*].

6.1.2.2 Extraversion

The n-grams determined to be significant collocations for Extraversion groups are shown in figure 6.2. There is a more even split of n-grams across high and low groups and both levels of significance than with Neuroticism.

Analysis again begins by considering those LIWC categories which have shown significant relationships with Extraversion. Use of Occupation and Achievement words were associated with Introverts, and the only word from those categories ('work') indeed occurs in the low collocations. The low group also contains the Discrepancy words 'but' and 'want'.

Both groups however contain words that relate to Social processes and the Present tense when these previously related only to high Extraversion. Note though, that the Social words in the High group are references to other people, while the Low collocation [*listen to*] could just as easily refer to music. There were no words found that fell in the remaining correlating LIWC categories.

Beyond those categories which previously correlated, a pattern once again arises following Pronouns. Introverts, use more collocations involving First-person pronouns, while Extraverts use more Third-person. The low group also contains more collocations using Prepositions.

Away from the LIWC, there is also a pattern for Nominals, with the high group containing several instances of 'NP1', while there are none in the low group. This was also observed in the e-mail corpus. There is also an interesting pattern for Introverts' use of clause initial first-person phrases with significant collocations of [*<p> but i*], [*<p> and i*].

There are other similarities between the blog and e-mail corpus: [*cool <p>*] is

High Extraverts

(1) [*ok <p>*] [*NP1 <p>*] [*<p> NP1*] [*in NP1*] [*NP1 and*] [*<p> as*] [*<p> <p>*] [*<p> however*] [*however <p>*] [*<p> however <p>*] [*NP1 and i*] [*money <p>*] [*in NP1 <p>*] [*was that*] [*<p> i am*]

(2) [*and he*] [*cool <p>*] [*<p> NP1 and*] [*<p> i also*] [*and NP1*] [*went to*] [*well <p>*] [*oh well*] [*1 <p>*] [*i have been*] [*and i went*] [*oh well <p>*] [*was the*] [*NP1 to*] [*<p> NPD1*] [*to her*]

Neutral

(1) [*my friends*] [*i am*] [*kind of*] [*going to*]

Underuse: [*to NP1*] [*last night*] [*<p> on*] [*a couple of*] [*went to the*]

(2) [*<p> i think*] [*not going to*] [*get to*] [*who i*]

Underuse: [*yet <p>*] [*this morning <p>*] [*a couple*] [*this morning*] [*for NP1 <p>*] [*<p> last*] [*<p> last night*] [*<p> to*]

Low Extraverts

(1) [*<p> and*] [*<p> and i*] [*<p> but*] [*<p> but i*] [*that i*] [*i think*] [*<p> ok <p>*] [*i want*] [*<p> ok*] [*<p> perhaps*] [*last night <p>*] [*<p> or*] [*is to*]

(2) [*myself <p>*] [*it the*] [*of it*] [*of it <p>*] [*<p> in fact*] [*all of*] [*in fact*] [*listen to*] [*point in*] [*but i*] [*<p> and the*] [*and i'm*] [*couple of*] [*from work*] [*got a*]

Figure 6.2: Distinctive collocations for Extraversion sub-groups

Note: N-grams reaching (1) the 15.13 critical level ($p \leq 0.0001$) and (2) between 15.13 and 10.83 ($0.0001 < p \leq 0.001$)

a significant collocation for both Extravert groups; [*well <p>*] appears for Extravert blog authors, and low neurotic e-mail subjects, bearing in mind that Extraversion has a negative relationship with Neuroticism.

An unexpected pattern within the Introvert group is the use of both collocations that appear certain and those that reflect tentativeness. Introvert collocations include [*<p> perhaps*] and [*couple of*] but also [*in fact*].

6.1.2.3 Agreeableness

The n-grams that proved to be significant collocations for the Agreeableness groups are reported in figure 6.3. There are a reasonably large number of collocations relating to the levels of Agreeableness, with 25 level-1 collocations for the high group. Although, there are many overlaps.

There were a large number of LIWC categories reported to correlate with Agreeableness, but only a few of them significantly. Discrepancies are perhaps the most interesting. There are collocations which contain the Discrepancy words 'if' and 'wanted'. However, there are also collocations which appear to suggest discrepancies, while not containing any of the words from the category: [*find out*] and [*tried to*]. In direct contrast to this however is the high collocation [*figure out*].

Motion words ('going') and Certainty words ('fact') can also be found in collocations of the appropriate group (both low). Negations and Articles appear in both groups, although 'the' appears in three distinct collocations in the high group.

The high group has a large number of collocations involving both Inclusive ('and') and Exclusive ('but', 'however', 'that') words. Both have many containing verbs, particular Present tense, though the low group has a great variety. Interestingly, both groups have distinctive collocations involving 'have': [*not have*], [*i have*] and [*have an*] in the high group; [*have to*], [*didn't have*] and [*have any*] in the low group.

The first non-LIWC based observation that can be made glancing at the distinctive collocations for Agreeableness is the clearly skewed use of contractions. Of the 38 low collocations, five contain contractions: [*they don't*], [*there's no*], [*i wouldn't*], [*didn't have*] and [*if i'm*]. While five out of 38 does not at first appear significant, consider that there are no contractions in the high collocations. Further more, there are a number

High Agreeableness

(1) [*<p> and*] [*that i*] [*i am*] [*<p> as*] [*and i*] [*i will*] [*so <p>*] [*<eop> <sop> i*] [*<sop> i*] [*is not*] [*<p> and i*] [*<p> so <p>*] [*<p> however*] [*<p> however <p>*] [*however <p>*] [*<p> after*] [*<p> it is*] [*<p> the*] [*<p> but <p>*] [*but <p>*] [*NP1 and i*] [*<p> i will*] [*figure out*] [*<p> more*] [*it is*]

(2) [*<p> we*] [*there are*] [*not have*] [*i have*] [*have an*] [*<p> there*] [*money <p>*] [*process <p>*] [*and that*] [*this is not*] [*i will be*] [*<p> and we*] [*NP1 <p>*] [*of the*] [*<p> and the*] [*did <p>*] [*i had*]

Neutral

(1) [*all of*] [*ok <p>*] [*and i'm*] [*<p> ok*] [*<p> ok <p>*]

Underuse: [*<p> <p>*] [*in NP1*] [*in NP1 <p>*] [*managed to*]

(2) [*is to*] [*my friends*] [*i think*] [*listen to*] [*<p> <eop> <sop>*] [*<p> because*] [*right now*] [*<eop> <sop> so*] [*<sop> so*] [*<p> i just*] [*<p> <eop>*] [*from work*] [*to talk to*] [*to talk*] [*people who*]

Underuse: [*yesterday <p>*] [*<p> other*]

Low Agreeableness

(1) [*going to*] [*<p> so*] [*they don't*] [*NPD1 <p>*] [*have to*] [*there's no*] [*the office*] [*yes <p>*] [*at me*] [*in fact*] [*far too*]

(2) [*i wouldn't*] [*find out*] [*<p> perhaps*] [*birthday <p>*] [*going to be*] [*didn't have*] [*of me*] [*tried to*] [*wanted to*] [*this weekend*] [*have any*] [*bank holiday*] [*turn up*] [*<p> either*] [*used to*] [*like a*] [*tomorrow <p>*] [*so i*] [*thank god*] [*see it*] [*if i'm*] [*<p> everyone*] [*<p> yes*] [*who i*] [*off <p>*] [*to bed*] [*listening to*]

Figure 6.3: Distinctive collocations for Agreeableness sub-groups

Note: N-grams reaching (1) the 15.13 critical level ($p \leq 0.0001$) and (2) between 15.13 and 10.83 ($0.0001 < p \leq 0.001$)

that could be contracted but are not: [*i am*], [*is not*], [*it is*], [*not have*] and [*i have*]. There are no similar non-contractions in the low set. The most significant of these would appear to be those containing 'have', since they are most similar to one another.

There are more collocations that involve punctuation ('<p>') in the high Agreeableness group, while the low has more Prepositions. The low group also has some very specific collocations: [*bank holiday*], [*thank god*] and [*to bed*].

6.1.2.4 Conscientiousness

Figure 6.4 highlights the n-grams determined to be significant collocations for Conscientiousness. As with Neuroticism, there are noticeably fewer distinct collocations for the high group than there are for the low group.

There was very little from the LIWC that correlated with Conscientiousness. The only related collocation is [*friends <p>*] which occurs, following the negative correlation with the Friends category, in the low group.

An interesting collocation that seems to relate very well to the careful attentive nature of highly Conscientious individuals is [*<p> <eop>*]. This suggests that those in the high group are more likely to ensure they end a post with punctuation.

The high Conscientiousness group has greater use for multiple punctuation. Neither the high nor low group contains any collocations involving nominals. However, the low group contains more collocations of Third-person pronouns.

The low group also contains collocations that could reflecting change of topic markers: 'anyway', 'actually' and 'however.' The low group also has many more collocations with verbs, while the high group only has one.

6.1.3 Discussion

There are many collocations that proved significant to one group or another across each personality dimension. Finding an appropriate manner to summarise the findings is difficult, but there have been a number of patterns emerge. Now while some of these have returned to individual words, there are often still contextual differences across those.

High Conscientiousness

(1) [*<p> <p> <p>*] [*<p> <p>*] [*ok <p>*] [*i will*] [*going to*] [*<p> <eop> <sop>*] [*<p> <eop>*] [*<eop> <sop>*] [*kind of*] [*the way <p>*] [*anymore <p>*] [*<p> i will*]

(2) [*<p> the*] [*by the way*] [*that <p> i*] [*how i*] [*<p> i hope*] [*next week*] [*the way*] [*him <p> i*] [*what i*] [*<p> etc*]

Neutral

(1) [*<p> ok*] [*<p> ok <p>*] [*i am*] [*to talk*]

Underuse: [*in NP1*]

(2) [*<p> damn*] [*my friends*] [*are going*] [*listen to*] [*i get*] [*not going to*] [*are going to*] [*and i*] [*and then*] [*i would*] [*to make*] [*<p> our*] [*fun <p>*] [*<eop> <sop> i*]

Underuse: [*<p> then <p>*] [*in NP1 <p>*] [*this morning*] [*anything <p>*] [*this morning <p>*] [*managed to*] [*the weekend <p>*] [*<p> other*] [*a bit*]

Low Conscientiousness

(1) [*<p> and*] [*<p> i*] [*that i*] [*<p> and i*] [*<p> however*] [*<p> one of*] [*<p> i hate*] [*<p> however <p>*] [*the game*] [*however <p>*] [*<p> last*] [*she was*] [*<p> to*] [*friends <p>*] [*last night*] [*<p> as*] [*i was*] [*that my*] [*anyway <p>*]

(2) [*<p> actually*] [*a few weeks*] [*<p> i should*] [*<p> despite*] [*do is*] [*thing to*] [*long as*] [*them <p> i*] [*<p> she*] [*sort of*] [*like a*] [*okay <p>*] [*<p> i was*] [*<p> i want*] [*that she*] [*in a*] [*<p> in*] [*<p> actually <p>*] [*case <p>*] [*is to*] [*<p> anyway <p>*]

Figure 6.4: Distinctive collocations for Conscientiousness sub-groups

Note: N-grams reaching (1) the 15.13 critical level ($p \leq 0.0001$) and (2) between 15.13 and 10.83 ($0.0001 < p \leq 0.001$)

One of the interesting patterns to emerge was the use of pronouns across the Neuroticism groups. It is expected that high scorers are more concerned with themselves than others. While the low scorers did show more significant collocations containing second- and third- person terms, they also used more first-person. Of course this does not mean they use more in general, since no correlation was found with the relevant LIWC categories. What these findings are more likely to reflect is that low scorers use first-person pronouns in a wider range of contexts.

Another interesting association with personality is that high Agreeableness scorers use less contractions, at least distinctively, than low scorers. This would seem to fit with Agreeableness reflecting an accommodating considerate nature, by using less informalisms.

There is one point worth noting regarding the [<*p*> *however* <*p*>] collocation. This collocation is distinctive in one group of each personality trait: low neurotics, high Extraverts, high Agreeableness and low Conscientiousness. There are also a number of very specific collocations that have proven distinctive. Note that this is not a criticism of the approach but of the raw blog data. Filters make sure that the data is not overly fit to any one individual. However, it is possible that results are very much fit to the groups of the study and as such results may not be generalisable.

6.2 Individual use of Collocations

The previous analysis was used because it abstracts away from the potential influence of any one individual. By treating the group as a whole unit and further considering only items used by half the group, it is possible to identify instances of language use more common to one group than another.

However, it is possible that this procedure abstracts too far from the variation within a personality trait. By assigning individuals with a range of scores to one group they are treated as equal when this might not be the case. So whilst results can prove useful in classifying a subject as a member of a group, they would be of no use in placing them on a more fine-grained scale.

Also, as noted in the previous discussion, it is possible that results are too closely

fit to the groups and will not generalise beyond. However, they are only fit to those individuals in the groups: due to the use of the neutral sub-group, there are individuals in the mid group of each trait who were not included in the analysis. This section investigates the direct relationship between personality trait and the n-grams identified as reflecting the extremes of that trait in the previous section.

6.2.1 Methodology

The starting point for this analysis is those n-grams determined in the previous section to be significant in distinguishing extreme personality groups. The relative frequency of each of these n-grams was calculated for every subject in the corpus.

The first step is to calculate the Pearson correlation of these relative frequencies with the appropriate personality trait. This will reveal which of the n-grams most strongly relate to each trait across the whole corpus.

The next step, following the general methodology of chapter 5, is to perform a multiple regression on the correlating n-grams. This is to see how much of the variance of each trait can be explained by individuals use of specific language. As with previous analyses, personality trait is the dependent variable, while n-grams as independent variables are entered into a step-wise multiple regression. Since variables must be independent, n-grams of different lengths representing the same collocation (as discussed in section 6.1.2) cannot be used together. Should this be the case, as in the multiple regression of LIWC variables (section 5.3.2), the more specific variable (longest n-gram) is entered first.

6.2.2 Correlation Results

6.2.2.1 Neuroticism

As with previous correlation analyses, following Gill, n-grams that show a relationship at the $p < .1$ level are reported (see table 6.1). In the course of the analysis, three n-grams actually correlated in the opposite direction to that expected ([*at work*], [*so* <*p*>] and [<*p*> *to*]) though none of these approached significance.

The first observation is a purely numerical one: of the 58 collocations that proved

N-Gram	*r*
[*was that*]	–.356**
[*NP1 <p> NP1*]	–.351**
[*this year*]	.343**
[*to eat*]	.321**
[*slowly <p>*]	.316**
[*and buy*]	.309**
[*<p> after*]	–.280*
[*is that*]	–.273*
[*<p> i had*]	–.270*
[*like a*]	.264*
[*if i*]	.263*
[*<eop> <sop> so*]	.262*
[*write about*]	.252*
[*<sop> so*]	.251*
[*that he*]	–.244*
[*a bit of*]	–.243*
[*me <p> <eop>*]	–.242*
[*and i*]	–.224
[*the best*]	–.218
[*<eop> <sop> i*]	–.212
[*to NP1*]	–.202
[*mean <p>*]	–.200

Table 6.1: Correlation of Neuroticism scores with N-Grams

N-Gram	*r*
[*point in*]	–.341**
[*and he*]	.322**
[*last night <p>*]	–.293*
[*it the*]	–.277*
[*cool <p>*]	.269*
[*1 <p>*]	.261*
[*<p> NP1*]	.245*
[*to her*]	.233
[*is to*]	–.232
[*<p> NP1 and*]	.204
[*<p> as*]	.201

Table 6.2: Correlation of Extraversion scores with N-Grams

distinctive to either extreme group of Neuroticism (figure 6.1), only 22 correlate with at least marginal significance. A by-product of this reduced set is that the patterns previously identified (section 6.1.2.1) are less apparent.

There is still evidence of previously correlating LIWC categories with the Discrepancy word 'if' and the Eating word 'eat.' Collocations involving references to other people all correlate negatively, this most strongly shown by what appears to be a list of names, [*NP1 <p> NP1*], reaching the $p < 0.01$ level of significance. Here again, we also see different collocations of the same word correlating in different directions. [*if i*] is used more by high neurotics while [*and i*] and [*<p> i had*] are used more by low scorers.

One collocation notable by its absence is the high neurotic use of multiple punctuation. This proved to be one of the most distinctive collocations of both this study and that of Oberlander and Gill (2005), yet it does not correlate with the raw score.

6.2.2.2 Extraversion

There are even less patterns evident in the correlation with Extraversion 6.2. One that remains is that the four references to other people, third-person pronouns and nominals, correlate positively as expected.

One result which is odd is the significance level of [*1 <p>*] as it relates to high Extraversion. Intuitively, this would be less surprising following a positive correlation with MRC's measure of digits used in a text. However, this correlated with Openness and Neuroticism, but not Extraversion.

6.2.2.3 Agreeableness

There were a large number of distinctive collocations for Agreeableness, and accordingly a large number that correlate with the raw trait. This means that a number of the patterns previously identified are still present. Collocation with contractions correlate negatively, while those without are positive. There are collocations containing 'have' that correlate in each direction. A number of the collocations suggesting discrepancies also correlate as expected, including the contradictory [*figure out*] (positive) and [*figure out*] (negative).

Perhaps more worryingly, is that a number of the seemingly specific phrases also correlate significantly, [*thank god*] and [*bank holiday*] for example.

6.2.2.4 Conscientiousness

The n-grams which proved to correlate directly with Conscientiousness are reported in table 6.4. There were fewer obvious patterns that could be identified in the collocations for the high and low Conscientiousness group, and subsequently few that can be seen here.

There are a number of collocations involving punctuation, correlating in both directions. Absent again however is the high use of multiple punctuation. This neither correlated with Neuroticism, nor here with Conscientiousness.

A number of the collocations also suggest over-specificity: [*case <p>*] and [*the game*] for example.

N-Gram	*r*
[*is not*]	.489**
[*this is not*]	.401**
[*thank god*]	–.378**
[*<p> it is*]	.355**
[*have any*]	–.342**
[*have to*]	–.332**
[*turn up*]	–.324**
[*wanted to*]	–.314**
[*<p> after*]	.310**
[*not have*]	.305**
[*going to be*]	–.290*
[*process <p>*]	.291*
[*figure out*]	.280*
[*at me*]	–.277*
[*<p> so <p>*]	.274*
[*i wouldn't*]	–.273*
[*did <p>*]	.266*
[*of me*]	–.262*
[*they don't*]	–.260*
[*yes <p>*]	–.258*
[*<p> and the*]	.258*
[*and that*]	.251*
[*so <p>*]	.251*
[*find out*]	–.250*
[*tried to*]	–.246*
[*<p> the*]	.242*
[*tomorrow <p>*]	–.240*
[*bank holiday*]	–.240*
[*<p> either*]	–.229
[*off <p>*]	–.227
[*of the*]	.227
[*see it*]	–.225
[*going to*]	–.211
[*have an*]	.203
[*<p> more*]	.203
[*the office*]	–.201

Table 6.3: Correlation of Agreeableness scores with N-Grams

N-Gram	*r*
[*a few weeks*]	–.405**
[*case <p>*]	–.418**
[*okay <p>*]	–.378**
[*the game*]	–.387**
[*by the way*]	.332**
[*in a*]	–.326**
[*do is*]	–.313**
[*that my*]	–.308**
[*<p> despite*]	–.291*
[*<p> one of*]	–.264*
[*<p> to*]	–.262*
[*<p> i hope*]	.263*
[*the way <p>*]	.253*
[*the way*]	.252*
[*how i*]	.250*
[*<p> i should*]	–.232
[*friends <p>*]	–.225
[*kind of*]	.210
[*<p> anyway <p>*]	–.208
[*i was*]	–.204
[*is to*]	–.203
[*<p> last*]	–.199

Table 6.4: Correlation of Conscientiousness scores with N-Grams

6.2.2.5 Discussion

The first point worth noting is that while the correlation analysis is informed by the log-likelihood comparison study, the two approaches and subsequent results should be considered distinct. That is to say, if an n-gram does not correlate with a trait variable, the first stage of analysis is not necessarily inaccurate. The variable may be highly used by one extreme subgroup, while completely underused by everybody else. Consider the example distribution of figure 6.5: an n-gram with this distribution is clearly representative of the high-subgroup. However, the frequency would not correlate strongly with the trait variable, since the distribution is distinctly uneven.

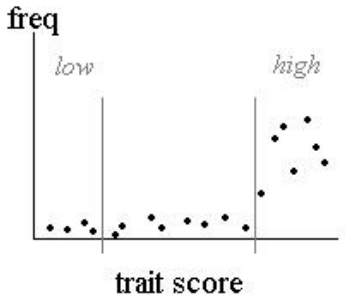

Figure 6.5: Theoretically possible distribution of a high sub-group n-gram

This could explain the lack of a full trait relationship with many of the more distinctive collocations from the stratified analysis: the high neurotic use of multiple punctuation for example.

An important observation of the results of the correlation analyses reported in this section is that there does not appear to be many obvious patterns to the collocations that correlate. This is one of the aims of the bottom-up approach. By using dictionaries, patterns are presented which must be searched for within the data. By starting direct from the data, it is hoped that patterns will emerge.

Despite some apparent specificity in the results, some patterns have remained from the group analysis intact: lower neurotics are more likely to refer to other people than

those on the high end of the scale; likewise Extraverts talk more distinctively about other people than Introverts; Agreeable writing appears to contain less contractions.

6.2.3 Multiple Regression Results

Correlation analysis only shows which of all features relate in some way to the personality dimensions. This section explores that relationship further to determine if those features can in fact explain any of the variance within a dimension. Those n-grams which showed a relationship at the $p < .1$ level of significance were entered into a stepwise regression for the appropriate personality trait. The four resulting equations can be seen in table 6.5.

Given the number of collocations that correlated significantly with the personality traits, there are surprisingly few that made it into the regression equations. Despite this however, the resulting level of variance explained by each appears high, certainly a higher level than is explained in the dictionary-based approaches of the previous chapter.

The equation for Neuroticism accounts for 59% of the variance. Of perhaps most interest are: the low neurotic preference for beginning posts with 'i'; high neurotic use of the discrepancy phrase [*if i*]; and the low scorers use of the affirming and positive [*the best*].

Extravert collocations retained in the regression equation continue to reflect their social nature. However, the interpretation of the presence of [*1 <p>*] is less clear. The equation accounts for 39% of the variance.

With only one more variable retained for the Agreeableness equation than for Extraversion, an impressive 57% of the variance is explained. Use or not of contractions still appears to be important, with the inclusion of [*is not*] and [*they don't*] in the equation.

The greatest amount of variance explained is 66% for Conscientiousness. Again, distinctive collocation containing similar terms (in this case First-person pronouns 'I' and 'my') but relating in opposite directions are retained in the equation: [*how i*] and [*<p> i hope*] are used more by high scorers while [*i was*] and [*that my*] are used more by low scorers.

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	[*was that*]	–.35	.000		
	[*this year*]	.26	.004		
	[*if i*]	.39	.000		
	[*the best*]	–.33	.000		
	[*<eop> <sop> i*]	–.24	.007		
	[*<p> i had*]	–.19	.028		
	[*and i*]	–.23	.008		
	[*is that*]	–.20	.022	**.60**	**.000**
E score	[*and he*]	.40	.000		
	[*1 <p>*]	.28	.008		
	[*last night <p>*]	–.28	.006		
	[*<p> as*]	.27	.009		
	[*<p> NP1 and*]	.26	.009	**.39**	**.000**
A score	[*is not*]	.44	.000		
	[*have to*]	–.34	.000		
	[*they don't*]	–.23	.008		
	[*bank holiday*]	–.26	.003		
	[*have any*]	–.24	.006		
	[*<p> more*]	.18	.039	**.57**	**.000**
C score	[*case <p>*]	–.39	.000		
	[*a few weeks*]	–.24	.005		
	[*that my*]	–.36	.000		
	[*how i*]	.30	.000		
	[*i was*]	–.29	.055		
	[*kind of*]	.26	.001		
	[*<p> i hope*]	.18	.019		
	[*do is*]	–.17	.034	**.66**	**.000**

Table 6.5: N-Gram relative frequency multiple regression analysis with personality scores

6.2.3.1 Discussion

The first observation to make concerning the multiple regression equations is the higher level of variance explained simply by the use of a handful of word n-grams. Admittedly, some of these do appear quite specific, and so regression equations appear to be the result of over-fitting. This cannot be the case entirely however, because n-grams were selected for multiple regression (via correlation) from the earlier stratified corpus comparison approach (section 6.1). Due to the adoption of a single neutral group for the entire corpus (see section 3.4.5 for details) rather than simply using the mid-group for each trait, only half of the potential subjects were used in each analysis. The data provided by this half of the corpus appears to fit well to all subjects.

One conclusion of the impressive regression results is that, compared to the low multiple regression scores reported by the dictionary-based approaches (sections 5.3.2 and 5.3.2), it appears that context — taking into consideration a word's surroundings — can be very important in linguistic analysis. Of course, it is worth noting that this is perhaps an inevitable outcome given the potential variable space. The LIWC contains only 2000 words and word stems, while the n-grams were drawn from over 400,000 unique bi- and tri-grams. Still, that the analyses carried out here ultimately reduces that to sets of five to eight n-grams shows that these are important findings.

In reporting the results of this n-gram analysis, patterns similar to those of the dictionary analyses have been identified. However, with context we can see individual differences that defy single word-level patterning.

No significant correlation has been previously found for first-person pronouns with Conscientiousness. However the regression equation reveals that one possible reason for this is the use of 'I' in different contexts: high scorers are more likely to use [*how i*] while low scorers will use [*i was*]. Similarly, only negatively related collocations containing 'have' made the regression equation for Agreeableness, but from previous stages (table 6.3 and figure 6.3) it is clear it is used in different contexts by both extremes.

Still, despite the impressive results, as has been noted, some collocations appear to be particularly specific: [*1 <p>*] and [*bank holiday*] for example. An obvious approach to deal with the first of these is to replace cardinal numbers with a more general

tag as was done for proper nouns. More generally however, there may be an issue of frequency distributions. While there are filters in place to ensure that any n-gram was used by half of any group, there was no balance of how many times each individual used a collocation. On top of a number of subjects filter, which ensures frequency per group, a frequency per individual filter may be required. This would perhaps further abstract away from particularly uncommon features such as hapax legomena. However, whilst there is no denying that such features are important in single author identification for example, it is hard to argue that such low frequency features can be representative of a group. Indeed, they must also be of little use when dealing with a continuous trait such as a personality dimension.

6.3 Contextuality

Following the n-gram approach of the last section, attention turns to unitary linguistic measures. The next section, average word frequency rank is used to explore individual differences, while this section adopts a measure of contextuality. In 2.7.2.2 Heylighen and Dewaele's F-measure (2002) was introduced. This measure is based on the notion of deixis, or the contextual nature of language. The F-measure was created based on certain parts-of-speech. As a reminder, the summation equation is included here:

$$\begin{aligned} F = \quad & 0.5 * [(noun\,frq + adj\,frq + prep\,frq + art\,frq) \\ & -(pron\,frq + verb\,frq + adv\,frq + int\,frq) + 100)] \end{aligned}$$

Low scoring texts are those considered most contextual in style; those that score higher are least contextual in nature. In section 4.1 this measure was used to place blogs in the context of genres selected from the BNC. The ordering the F-measure provided was plausible, suggesting that it does indeed measure that which it was designed to measure. Here the F-measure will be used to investigate differences in contextuality as they relate to personality.

Section 2.2.1.2 discussed previous work on personality and part-of-speech usage.

Dewaele and Furnham (2000) found that Extraverts preferred implicit language. However, using the F-measure — an extension of this earlier work — they found little effect for Extraversion. Still with implicitness, Oberlander and Gill (2004) hypothesised that high neurotics would have a similar preference and used n-gram analysis to provide some support of both preferences. Re-analysis has confirmed the presence of an effect (Oberlander & Gill, 2005).

It is therefore possible that both Extraverts and high Neurotics may use more contextual language; Introverts and low neurotics using less contextual language. However, if it is the case that there is only an effect under pressure, then it is unlikely one will be found in the blog corpus, a distinctly low-pressure task.

Heylighen and Dewaele (2002) also hypothesised that Openness might relate to contextuality. Though they had no data at the time, since Openness is often regarded as the factor of intellect they suggested that there should be a negative correlation: high scores in Openness would reflect less contextual language use. Note that this hypothesis was posited when the F-measure was considered to measure contextuality against formality (see section 2.7.2.2 for more details). The original hypothesis was that high scorers of Openness were more formal.

This section will investigate these hypotheses, as well as potential relationships with Agreeableness and Conscientiousness.[3] Further to this there is a small exploration of the nature of the F-measure. Extra-linguistic information derived from the blog corpus is used to investigate the relationship it has with data of a deictic nature.

6.3.1 Correlation Analysis

6.3.1.1 Method

In order to calculate the overall F-score of the blog corpus (as required for section 4.1), each author's F-score had already been computed. This used the part-of-speech tags as marked by the MXPOST tagger (Ratnaparkhi, 1996). As in the previous section, care was taken so as to preclude the influence of extreme individuals: in this case, outliers were excluded.

[3] A prior version of this study has been reported in Nowson et al. (2005). Though the approach and data have since been marginally revised, results all report similar directions.

Trait	*r*
Neuroticism	–.150
Extraversion	–.118
Openness	.170
Agreeableness	.260*
Conscientiousness	.076

Table 6.6: Correlation between F-score and personality trait

Note: two-tailed, *$p<0.05$

The average blog f-score was 53.3, (SD 4.08). Outliers are considered as those scoring above or below two standard deviations from the mean. This control removed three subjects: two scored higher than the maximum threshold, while one scored lower than the minimum.

6.3.1.2 Results

The results of the Pearson Correlation analysis for the remaining 68 files in the blog corpora along the Five Factor dimensions are displayed in Table 6.6.

Given previous hypotheses, a negative correlation with Neuroticism and Extraversion was expected. The correlations are in the expected direction but they are non-significant. This is perhaps more in line with Heylighen and Dewaele's findings, though there does appear to be some effect for at least Neuroticism, as Oberlander and Gill have found (2004, 2005). However, there is a stronger positive and significant correlation with Agreeableness. The correlation with Openness is also reasonably strong and positive, as predicted by Heylighen and Dewaele, though it also does not reach significance. Conscientiousness shows the smallest correlation of all.

Due to the constituent nature of the F-measure, it is also possible to investigate the frequencies of the individual parts-of-speech that define it. When there is an overall negative correlation between personality trait and the F-measure—as with Extraversion and Neuroticism—a negative correlation between trait score and frequencies for nouns, adjectives, prepositions and articles would be expected, while there should be

a positive correlation for pronouns, verbs, adverbs and interjections. The opposite should hold when there is positive correlation between trait score and the F-score—as with Agreeableness and Openness. Table 6.7 displays the results.

As might be expected from the correlations shown in Table 6.6, Agreeableness has the overall strongest correlations. However, Openness has the strongest individual correlations and the most that reach significance, the most significant being use of adjectives. None of the Extraversion and Neuroticism correlations reach significance, and once again there are only small correlations for Conscientiousness.

However, with only a few exceptions, the directions of the correlations are as expected. Neuroticism and Extraversion scores indeed for the most part correlate positively with the frequencies of more contextual parts-of-speech, and negatively with those parts-of-speech considered least contextual. Agreeableness and Openness scores correlate in the opposite directions, as does the Conscientiousness score.

6.3.2 Stratified corpus analysis

It therefore appears that there is some relationship between contextuality for the four personality dimensions of Neuroticism, Extraversion, Openness and Agreeableness. However, the relationship is stronger in some cases than in others. To take a closer look at each case, the average F-measure will be calculated for each of the high, medium and low scoring sub groups, as introduced in section 3.4.5.

6.3.2.1 Method

As explained earlier in the thesis (see section 3.4.5 for more details), the corpus is stratified using the mean and standard deviation of each personality trait score. High and Low personality sub-groups are created for each personality dimension by splitting off the groups at greater than 1 standard deviation above, and below, the mean score for each dimension. The remainder of the subjects are allocated into the Mid sub-group for that dimension. The exception is Openness, which due to the nature of its distribution, has only mid and high sub-groups. Note that in this analysis, the outliers have been removed, as described in section 6.3.1.1.

Trait	Noun	Adjective	Prep'n	Article	Pronoun	Verb	Adverb	Interj'n
Neuroticism	−.175	.105	−.114	−.093	.009	.217	.197	.038
Extraversion	.007	−.118	−.133	−.140	.236	.033	−.016	.139
Openness	.050	.342**	.272*	.010	−.019	−.152	−.240*	−.192
Agreeableness	.188	.141	.095	.260*	−.167	−.230	−.256*	−.215
Conscientiousness	.089	.020	−.018	.069	−.136	−.013	−.038	−.033

Table 6.7: Correlation between POS frequency and personality trait

Note: two-tailed, *$p<0.05$, **$p<0.01$

6.3.2.2 Results

The average F-score for the sub-groups, by dimension, can be seen in table 6.8. In order to aid interpretation, these results are plotted in figure 6.6.

Trait	Low	Mid	High
Neuroticism	54.1	53.0	53.1
Extraversion	54.8	52.8	53.2
Openness		52.3	53.8
Agreeableness	52.3	53.2	53.9
Conscientiousness	53.0	53.0	53.8

Table 6.8: Average F-score of corpus stratified by trait

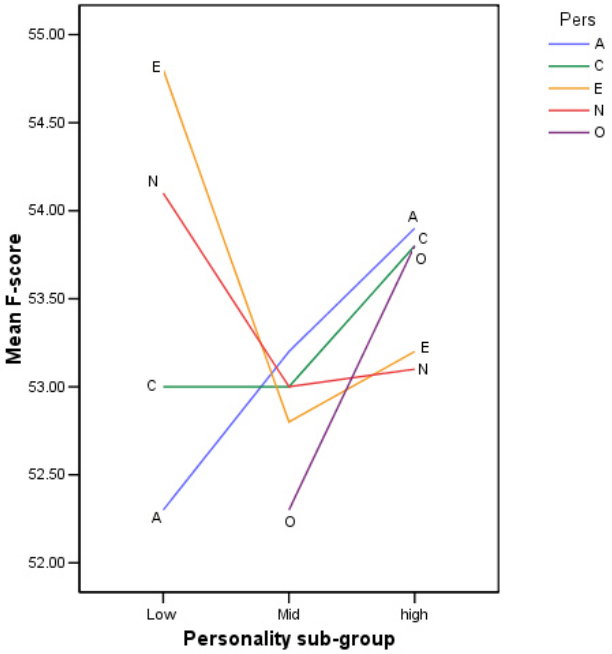

Figure 6.6: Average F-score of personality trait sub-groups

The most predictable results are those of the strongest correlations. Low scorers of Agreeableness clearly have a lower average F-score than the mid group, who in turn

score lower than the high group. Similarly with Openness, though there are only two groups. The more interesting results are for the traits which scored the lower correlations previously. In all three of Neuroticism, Extraversion and Conscientiousness, there are two groups with similar scores, and one with much higher.

Low and mid scorers of Conscientiousness have an average score very similar to that of the corpus as a whole, while the high group are less contextual. It is low Neurotics, as expected by the negative correlation who use less contextual language, while the mid and high groups again write at the average blog level. The biggest difference lies within Extraversion. It is again the low group with the least contextual writing, of *all* sub-groups, with the mid group much more contextual, and the high group just a little less so than them. This follows Heylighen and Dewaele's finding that it is only the most Introverted subjects who show a difference in the level of their writing. This is unexpected however, since this effect was previously only found in stressful oral situations.

It is worth noting however, that under statistical tests of difference none of the extreme groups actually differ significantly.

6.3.3 Discussion

Neuroticism and Extraversion did not correlate with contextuality with enough significance to confirm a definite effect. However, both factors correlate with the F-measure in the expected directions to a reasonable degree, which seems to tie with the degree of support found by Oberlander and Gill (2004, 2005). Examining the average F-scores of the stratified subgroup reveals the extent of the effect: it is most certainly not a wholly linear effect, with only the low groups differing from the average. Likewise for Conscientiousness: correlation suggested no general effect, but stratified analysis revealed a localised one. Within this corpus however, these effects have not proven significant.

The only other trait for which a hypothesis had previously been suggested was Openness. Heylighen and Dewaele hypothesised that since Openness is also considered the factor of intellect, it should correlate negatively with contextuality. The results here suggest that while it is not significant in this study, the trend is in the expected

direction.

The relationship between Agreeableness and language use, however, has not been extensively discussed previously. One aspect of Agreeableness is cooperativity: highly Agreeable individuals are most willing to cooperate and accommodate. In communication, this could be realised via a better ability—or at least willingness—to adapt to the interlocutor's communication situation or style. Interpreting this in the setting of blogs suggests that bloggers of an Agreeable nature are more likely to be aware of the lack of shared context between themselves and the reader, thus adjusting their writing away from contextuality toward a higher F-score. The results reported here show this, with both a significant negative correlation between Agreeableness and contextuality, and the steady increase in F-score (decrease in contextuality) between sub-groups.

This apparent preference of high Agreeableness scorers for a less contextual style, as it may or may not relate to ideas of formality, has been seen previously in the collocation analysis: the pattern reported for use of contractions—the low group containing several distinctive collocations containing contractions, the high group containing several that were not contracted (see figure 6.3 and table 6.3)— reflects a more 'formal' approach to writing.

This finding also helps explain why Agreeableness correlated negatively with the LIWC factor of 'Immediacy' in section 5.2, where Pennebaker and King (1999) found it to be positive. It appears the 'Immediacy' is related to contextuality: that language considered immediate assumes a level of shared context between reader and writer. Rather than making their writing more accessible by using simpler language, as essay writing subjects are free to do, highly Agreeable bloggers, with a much larger audience to consider, prefer to do so by making it less contextual.

This result proves of further interest in the context of previous work on blogs. The comments of Nardi et al. (2004; reported in section 2.5.4) that 'bloggers consider audience attention, feedback and feelings as they write [. . .] consciousness of audience is central to the blogging experience' seems most relevant. The result here suggests that not all bloggers are as conscious of their audience as some, or if they are, they are simply less willing to adapt their style to suit their audience.

6.3.4 Deictic correlates of the F-measure

The F-measure is derived from the principles of deixis as it reflects the contextual nature of language. This section reports analysis which is intended to explore the relationship that the F-measure has with notions of the principal on which it is based. Heylighen and Dewaele (2002) considered four situational factors related to deixis: the *persons* involved, the *space* of the communication, the *time*, and the prior *discourse*. From the questions asked in this study regarding blogging habit, as well as figures calculated from the blog data, further measures can be derived which reflect these categories.

- **Persons** The more interlocutors know one another, the more knowledge they share and the more contextual they are able to be when communicating. Bloggers were asked to consider who they were writing for: 'themselves', 'friends and family', 'members of a known community of interest group', 'anyone' or 'everyone'. Whilst anyone can read any weblog if it is not explicitly made private, it is felt that the author's familiarity with their intended audience will have greatest effect on their writing style. These results were scored highest for people who wrote privately for themselves, and lowest for those who wrote to as wide an audience as possible. Since the difference between 'anyone' and 'everyone' was left for the subject to decide, they were scored in three ways. First 'everyone' was scored lowest, then 'anyone', then they were both scored the same. Correlations were very similar, so only the latter will be reported here.

- **Space** The closer two individuals are geographically, the more they have knowledge of each other's physical context. There is no explicit data in this study that can be used to examine this principle. It may be possible to investigate from where a blog is read, but this relies on the subject using a webtracker and also their being aware of who reads the blog.

- **Time** The time between communication acts also affects the context available. Instant communication provides more opportunity for contextual references than time delayed communications such as letter writing. Likewise in monologues: two lectures on a subject can assume more context if there is an hour break

between them than if there was a week. Subjects of this study were asked how often they felt they wrote in their blog, but this was found to have no correlation with how many posts they actually made. Therefore this latter data is used as a measure of time: the number of personal chunks a subject makes in one month; the number of personal chunks written per day.[4] The more often an author writes, the more contextual they can be.

- **Discourse** If a communication act is part of a much larger discourse, then the prior discourse is context from which references can be drawn. One possible measure of prior discourse in blogs is to consider for how long a subject has been blogging. However, this is an unreliable measure: since this was not explicitly asked of subjects, it can only be gleaned from counting the number of months in their blog's archives, but when people switch blog providers, as they often do, they can lose these, so an accurate count is not always possible. In place of this, overall word count is used. This assumes that the volume of words produced in the month from which the corpus is taken is reflective of the subjects' long term blogging habits. So a higher number of words is used to reflect a greater prior discourse.

There is one further measure which relates to context that can be derived from the blog corpus. As discussed in section 2.6.1, Nilsson (2003) identified that bloggers use frames which allow the author to assume a degree of shared knowledge between writer and readers. She later relates this to the use of hyperlinks, suggesting that they are used in order to point the reader to more material, without having to explain it themselves. However, the perhaps obvious link count is not the measure being suggested; more relevant to contextuality is the number of words per link. A short link such as 'click here' assumes that the reader *will* follow the link in order to learn the context in which the author is situating their text; a longer link such as 'click here to see the hotel we stayed at' explicitly informs the reader of the context of the link, thus reflecting lower overall contextuality.

[4] As highlighted in section 3.4.5.1, this is the number of chunks written on each day that any writing was done.

6.3.4.1 Results

The Pearson correlation scores for the F-measure with the contextual factors described in the previous section can be seen in table 6.9. Note that these results reflect the corpus without outliers.

Deictic measure	r
Audience familiarity	–.205
Number of personal chunks	–.165
Personal chunks per day	–.241*
Word per month	–.269*
Average link length	.508**

Table 6.9: Correlation of F-score and Deictic blog measures
Note: two tailed, *$p<0.05$, **$p<0.01$

It is clear that all the deictic measures correlate in the expected direction, two proving significant at the $p < .05$ level, and average link length correlating very strongly indeed ($p < .001$). The negative correlation with audience familiarity despite being non-significant suggests that the more an author knows their intended audience, the more contextual their writing style. The raw number of personal chunks proved the weakest correlation but was still negative. However, the more fine-grained measure of the relative number of posts made per day is significantly negatively correlated. This seems to suggest that the more often a blogger posts the more contextual they are. Likewise, the total word count correlates negatively, further suggesting that the more that is written each month, the more contextual the style of the author. Note that this length effect is not an artifact of the calculation of the F-measure, which is based purely on *relative* frequencies of parts-of-speech. The strongest correlation is that for the average link length. As expected, this is positive, suggesting that authors with a more contextual style use more contextual or shorter links; less contextual authors on the other hand tend to use longer links.

What is interesting about these results, is that the contextuality of a blogger's style seems to say a lot about their blogging behaviour. Beyond this conclusion however, is

that the F-measure does indeed appear to relate well to deictic factors, as it purports to.

6.4 Word Frequency

In section 4.2 average word rank was used as a measure of word frequency to again place the blog corpus in the context of sub genres of the BNC. This section adopts the same approach to investigate differences in word frequency usage between individuals within the blog corpus.

In section 4.2.2 despite the correlation between the F-measure and average rank of the genres of the study, there were noticeable differences in the ordering of the genres. Despite these differences, it is intuitive to consider less contextual (as it was previously considered more 'formal') writing making greater use of longer words, which tend to be less frequent than shorter words. Therefore, the hypothesis for average frequency rank results is that results should be similar to those found previously for the F-measure. That is, the strongest effect should be that more Agreeable subjects use lower frequency words. This also follows from the findings of both the LIWC and MRC (see chapter 5) that both high Agreeableness and Openness scorers use longer words. Of course, this in itself appeared to contrast with positive correlations with MRC measures of word frequency for both traits.

6.4.1 Correlation analysis

6.4.1.1 Method

The method used for calculating the average word frequency rank is similar to that used in the genre comparison (see section 4.2.1). Word-POS tag pair frequencies were calculated for each subject, and these were used to lookup the ranked frequency list derived from the BNC. The average rank frequency of a file was calculated as the total rank sum divided by the number of words. Once again, a higher average rank frequency score reflects greater use of low ranking, low frequency words. Also, as with the F-measure study of the previous section, outliers are excluded.

The mean average word frequency rank across the blog corpus was 3584[5] (SD 695). Outliers are again considered as those individuals scoring outside the range of the mean plus or minus 2 standard deviations. This removed two subjects, one from either side of the boundaries. The subject removed with the higher score was in fact almost 5 standard deviations above the mean.

6.4.1.2 Results

Alongside the mean average rank, the mean percentage of words for which rank information was found was 96.9% (SD .97%). The correlations of the average rank and personality scores can be seen in table 6.10. Also included are correlations with the percentage of words that were found in the rank list. This is included since the rank list only included those words with a frequency of five or more in the BNC. Therefore a lower percentage suggests a greater use of very infrequent words.

Trait	*r* Ave rank	*r* Perc found
Neuroticism	–.169	.154
Extraversion	.084	.068
Openness	.263*	.138
Agreeableness	.132	–.025
Conscientiousness	.105	.030

Table 6.10: Correlation between personality score and average work rank, and percentage of words for which rank data was found.

Note: two-tailed, *$p<0.05$

Firstly, none of the correlations with the percentage of words found reach significance; the strongest result suggests that higher neurotics use slightly more words from the rank list. Similarly, the majority of the correlations with average rank are non-significant; Neuroticism again shows the highest non-significant relationship, suggest-

[5]Note this is different to the value reported for the blog corpus in the genre analysis (3495; table 4.3). This is because in the previous analysis, the blog corpus was treated as a whole, where here each subject is a separate file and an average for the corpus is calculated.

ing high neurotics use more frequent language.[6] Openness, however, does correlate significantly; highly Open individuals use less frequent language. This fits well with Openness being considered the factor of intellect.

6.4.2 Stratified corpus analysis

Following the approach adopted for studying the F-measure, this section takes a closer look at each personality dimension using the stratified corpus (continuing to exclude the two outliers).

6.4.2.1 Method

The method here is similar to that described above in section 6.3.2.1. For each sub-group, the mean average word rank is calculated.

6.4.2.2 Results

The mean average word rank for the sub-groups, by dimension, can be seen in Table 6.11. As with the F-measure analysis, these scores are also represented graphically in order to aid interpretation (figure 6.7).

Trait	Low	Mid	High
Neuroticism	3685	3536	3827
Extraversion	3397	3659	3632
Openness		3421	3740
Agreeableness	3463	3622	3715
Conscientiousness	3571	3586	3740

Table 6.11: Mean average word rank of corpus stratified by trait

Easily the most noticeable aspect of the stratified scores (figure 6.7) is their striking similarity to the pattern exhibited by the group F-scores (figure 6.6): Openness reflects

[6] A negative correlation with the average rank number, implies a positive relationship with rank, which in turns reflects greater use of more frequent language.

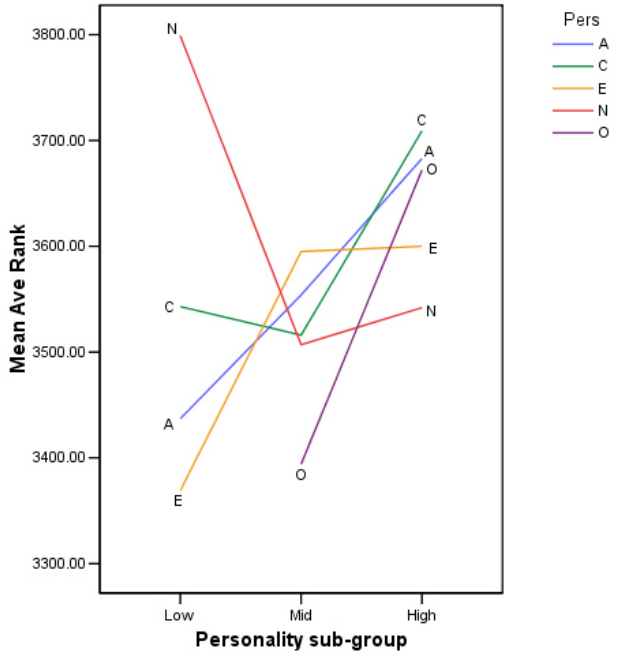

Figure 6.7: Average word frequency rank of personality trait sub-groups

its strong correlation; Agreeableness despite the much weaker correlation increases at a constant rate across groups; the low and mid groups of Conscientiousness have a similar score, while the high group scores much higher, and the low groups of both Neuroticism and Extraversion score differently to the very similar high and medium groups.

The main difference, which follows from the different direction of correlation, is that Introverts use *less* frequent words than the remaining Extraversion groups. In fact, the difference between the low and mid Extraversion group is almost as much of that between the two Openness groups. This directional result is perhaps to be expected given the negative relationship between Extraversion and Neuroticism. It does not however, follow from the F-measure result, as had been predicted. Of course, the overall correlations of Extraversion with both measures were among the lowest recorded for each.

6.4.3 Discussion

With the exception of Extraversion, the pattern of the stratified corpus results seem to follow those of the F-measure. The strengths of the correlation results also appear reasonably well related; the only real difference is that Agreeableness, which correlated significantly with the F-measure, correlates very little with rank.

Indeed, the average word rank and the F-measure correlate strongly ($r = .554$[7], $p < .001$). This suggest that use of infrequent words is a good indicator of a less contextual style of writing; greater use of more common words indicates a more contextual style. The exception is of course Introverts, who appear to be less contextual than mid or high scorers of Extraversion, while using more frequent language.

The results for Openness, and to a lesser extent Agreeableness, that they relate negatively with word frequency, are in opposition to the prior MRC finding for those traits (section 5.4.2). One reason for this, as argued in section 5.5, is that the frequency measures of the MRC may very well be out-of-date. Beyond this however, an argument could be made of the coverage of the categories. The three frequency measures of the MRC averaged between 80-90% coverage on the blog corpus (SD 6%). In contrast, the approach adopted in this section garnered an average coverage of 97.9% (SD .97%).

6.5 Top-down Versus Bottom-up

Both this chapter and the last have used multiple regression analysis in an attempt to explain the variance within personality traits. This section will report the results of one further set of personality-centred analyses: the purpose of these regressions is to enable a simplistic comparison between the top-down and bottom-up techniques used in this thesis.

It is fair to say that the collocation approach of section 6.1 ultimately resulted in more variance being accounted for than either the LIWC (section 5.3) or MRC (section 5.4) analyses. In order to compare the methodologies, the features from all techniques employed so far will be entered together into multiple regression analysis. By examining which features are retained in the regression equations, an indication

[7] This is calculated discounting the outliers for both measures, four subjects in total.

of which approach may be most suited to exploration of individual differences will be provided.

6.5.1 Method

As with the previous multiple regression analyses the personality traits are considered as the dependent variables, and the dependent variables are entered into a stepwise regression analysis. Due to being unable to implement the collocation analysis on Openness, there is a sparseness of data derived from the bottom-up approaches. Therefore, Openness is omitted from this study.

The dependent variables for each trait is the set of variables from all approaches which have shown a correlation of marginal significance ($p < .1$). That is to say, all LIWC categories, MRC variables, and distinctive collocations which have shown a relationship with each trait. In addition to these, any of the LIWC factors of section 5.1, along with the F-measure, its constituent parts-of-speech relative frequencies and the average rank are included if they have shown a marginally significant relationship. The exception is the Dolby categories of the MRC, which have been shown to be largely unreliable.

Note that *all* correlating LIWC categories will be included; none of the genre, topic, or sparsity restrictions introduced in section 5.3.2.1 will be employed. However, control for independence is maintained: the most specific variable is entered into the regression, and only if it is not retained will the analysis be re-run with the more general category in its place.

6.5.2 Result

The equations resulting from the regression analyses for Neuroticism, Extraversion, Agreeableness and Conscientiousness can be seen in table 6.12.

Most obviously, there is little difference between these results and those from just the distinctive collocations. In fact, importantly, the only collocations lost from those analyses, are two from Agreeableness ([*they don't*] and [<*p*> more]), which are replaced by three further collocations ([<*p*> and the], [*of me*] and [*have an*]), and [<*p*>

NP1 and] is replaced in the Extraversion equation by the length-variant [<*p*> NP1].

For Neuroticism, in addition to the collocations, Discrepancy words and Physical states are retained, increasing the variance by just 7% to 67%. In addition to the loss and gain of collocations, the Thorndike-Lorge Frequency mean (high Agreeableness scorers using more frequent words) is also retained. The net effect of these changes is an increase of 8% of the variance explained to 65%.

Nothing additional is retained for Conscientiousness, 66% of the variance remains explained. The LIWC category Death entered the equation at an early stage in the analysis but fell out before the end.

Perhaps the most interesting equation is that of Extraversion. The equation consists of: five collocations, four found previously, and the other a variant of the dropped collocation; two LIWC categories, School and TV words, neither of which were retained in any stage of the original LIWC regression equations, and the relative frequency of pronouns - as computed for calculating the F-measure. The increase of variance is larger than has been found for the other traits, 16%. The F-measure pronoun figure however, is part of the bottom-up derived data set, so this increase is not all due to top-down data.

6.5.3 Discussion

There are a number of important general observations that can be made about the regression equations. The first is that for all intents and purposes, when adding all the other features to the set of collocations, none of these are lost from when they are regressed alone. The second is that additional features contribute very little extra to the variance explained, regardless of how much they could account for alone. This is perhaps further proof that despite being drawn from a much larger variable space the context-based n-grams are still useful data.

It is worth noting that the methodology here is perhaps slightly naïve. It has ignored all previously discussed doubts over the reliability of any of the linguistic features identified throughout the thesis. With the exception of excluding the Dolby categories, concerns over MRC frequency lists or collocation over specificity have not been taken into account. The reason for this is that this analysis has purely been an exploratory

illustration of the differences between the two methodologies. If these features were applied in an automatic classification scenario, there may be a degree of over-fitting, or some of the features may not measure that which they claim to as accurately as they should.

However, the results seem to show with convincing argument that the variance explained by data-driven approaches, and particular distinctive collocations, is considerably more than explained by dictionary-based methods.

6.6 Summary

Following the top-down methodology of chapter 5, this chapter adopted a number of bottom-up or data-driven approaches. The techniques used to investigate personality differences in language included word n-gram analysis and two unitary measures previously used to compare corpora.

The n-gram study began by identifying a number of distinctive collocations for the extreme groups of each personality trend. After identifying a number of patterns, attention returned to individual personality scores to investigate which of these collocations related to the overall trait. Those collocations were entered into multiple regression analyses, and the resulting levels of variance were higher than had previously been seen from the top-down feature sets.

The first unitary measure was Heylighen and Dewaele's F-measure (2002), a measure of contextuality. The strongest direct relationship found was with Agreeableness, showing that high scorers prefer a less contextual style; high scorers perhaps showing more willingness to acknowledge the lack of shared extra-linguistic context between themselves and their readers. There also appear to be some small effects for Introverts, low scorers of Neuroticism and high scorers of Conscientiousness. In addition to personality, a number of measures which could be considered to reflect aspects of contextuality showed significant relationships with the F-measure. This suggests that the F-measure does indeed appear to be related to the notion of deixis upon which it is based.

Despite earlier differences, the average word frequency rank appeared for the most

part to show similar relationships with personality traits. The strongest relationship was with Openness; more Open individuals are more likely to use less frequent words. The exception to the F-measure pattern was Extraversion. Despite showing a similar pattern of effect—just Introverts differ from the remainder—the effect was in the opposite direction, suggesting that whilst Introverts are slightly less contextual, they use slightly more frequent words.

This chapter concluded with a simple comparison between the methodologies of the previous chapter and those employed here. By using multiple regression analysis, it is clear that bottom-up approaches, specifically collocation, account for more variance within personality traits than features derived from dictionary.

This section concludes the study of language differences due to personality in blogs. In the next chapter attention turns to gender differences.

Dependent Variable	Independent variable	β	*p*	R^2	*p*
N score	[*was that*]	–.30	.000		
	[*this year*]	.18	.035		
	[*if i*]	.21	.028		
	[*the best*]	–.38	.000		
	[*<eop> <sop> i*]	–.22	.008		
	[*<p> i had*]	–.19	.021		
	[*and i*]	–.28	.001		
	[*is that*]	–.22	.007		
	Discrepancies	.28	.003		
	Physical states	.20	.017	**.67**	**.000**
E score	[*and he*]	.23	.024		
	[*1 <p>*]	.22	.020		
	School	–.30	.001		
	TV	–.25	.008		
	[*<p> NP1*]	.36	.000		
	[*last night <p>*]	–.30	.001		
	[*<p> as*]	.28	.003		
	F-Pronoun	.23	.021	**.55**	**.000**
A score	[*is not*]	.37	.000		
	[*have to*]	–.17	.045		
	[*bank holiday*]	–.34	.000		
	[*have any*]	–.23	.004		
	Thorndike-Lorge Freq. mean	.32	.000		
	[*<p> and the*]	.28	.002		
	[*of me*]	–.24	.003		
	[*have an*]	.20	.019	**.65**	**.000**
C score	[*case <p>*]	–.39	.000		
	[*a few weeks*]	–.24	.005		
	[*<p> i hope*]	.18	.019		
	[*i was*]	–.29	.055		
	[*that my*]	–.36	.000		
	[*how i*]	.30	.000		
	[*kind of*]	.26	.001		
	[*do is*]	–.17	.034	**.66**	**.000**

Table 6.12: N-Gram relative frequency multiple regression analysis with personality scores

Chapter 7

Linguistic Differences of Gender

So far this thesis has concentrated on personality traits, yet perhaps the most studied individual difference is that of gender. Section 2.3.1 discussed different approaches to studying gender differences in language. The simplest approach is to treat men and women as homogeneous groups, but it has been argued this loses a great deal of intra-gender variation. Methodologically speaking, this work adopts this simple approach, not least due to the small number of subjects in the study. Therefore, differences between men and women are explored wholesale. Relevant previous findings for general language differences due to gender were reported in section 2.3: typically, male language consists of more articles and prepositions suggesting greater concreteness, more swearing, and is more likely to contain opinions and insults and show a preference for discussing more impersonal topics; female language is more personal, with a higher use of pronouns and references to other people, and females also use more emotional language and more questions. This chapter will investigate gender differences as they relate to the measures used so far. Analysis follows the order of presentation in the thesis: first the top-down LIWC and MRC dictionaries are used; subsequently the data-driven unitary measures are employed. Note that as with Openness, the three way-comparison technique for identifying distinctive collocations is not applicable to gender.

This chapter also has a further methodological aim which is to assess the suitability of the techniques employed in this thesis for detecting language variation due to individual difference. If the methods used here can confirm previous findings for gender

and language, the findings for personality will be still more convincing.

Note on methodology

In order to examine differences between genders it is perhaps more commonplace to use t-tests which highlight significant differences in group means. However, in order to remain consistent with this thesis' earlier work on personality traits, correlation analysis is used. This is done by assigning a numeric value to each group and results in identical significance values as produced by t-tests. Here, females were assigned the value 0 and males 1. Therefore, a positive correlation suggests a property of greater male use; a negative correlation indicates a feature more likely to be used by females.

7.1 Top-down Approaches to Gender Differences

Analysis begins, as it did for the personality traits, with the top-down dictionary-based approaches. This section will discuss correlations with the variables of both the LIWC and MRC.

7.1.1 Correlation of LIWC factors with gender

In chapter 5 (section 5.2) personality traits were first correlated with the variants of Pennebaker and King's factor analysis (1999), as derived from the blog data. The factor analysis resulted in three factors equivalent to their 'Making distinctions', 'Immediacy' and 'the Social past' factors. This section explores the correlations, if any, that gender shows with the three factors and the variables from they were derived.

Since women are considered the more social gender, a negative correlation of 'the Social past' with gender is expected. Pennebaker and King previously found a significant correlation for gender with 'Immediacy': women preferring a more immediate writing style.

LIWC	*r*
Factor 1	**–.055**
Exclusive	–.058
Discrepancies	–.169
Tentative	.000
Negations	–.146
Present tense	–.186
Factor 2	**–.270***
– Articles	.333**
First-person singular	–.329**
Present tense	*–.186*
– Words $>$ 6 letters	–.009
Insight	–.163
Factor 3	**–.168**
Positive emotions	.060
Social	–.390**
Inclusive	–.072
Past tense	–.124

Table 7.1: Correlation of 13 LIWC categories with gender

Note: two tailed, *$p<0.05$, **$p<0.01$, $n = 71$. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading.

7.1.1.1 Results

Table 7.1 shows the Pearson correlation of the three factors with gender. The most striking observation is that with only two exceptions (Words of greater than six letters and Positive emotions, among the smallest of the correlations) all the variables correlate in the expected direction following that of the factor onto which they load. The first factor, 'Making distinctions' correlates only minimally with gender. Some of the variables associated with it correlate with a small degree of strength, though not significantly. Likewise the third factor, 'The social past' correlates negatively, as expected, but also not significantly. Words reflecting Social processes, however, show a highly significant correlation, suggesting that females do indeed talk more about social matters than males. Factor 2, 'Immediacy' shows the strongest, and only significant correlation of the three factors. This follows Pennebaker and King's results (1999) suggesting that to a certain degree, women tend to have a more immediate style in their blogs. The factor variables again show some strong correlations. As predicted by the literature, men use significantly more Articles, but fewer First-person singular pronouns.

7.1.2 LIWC and content differences

Once Pennebaker and King's factors (1999) had been reproduced and studied, it made sense to widen the scope of examination to include all LIWC variables. Section 5.3.1 describes the correlation study of the LIWC variables with personality, while section 5.3.2 describes the multiple regression of the correlating variables. This section replicates these studies while focusing on gender. First it reports the correlation analysis of all 71 LIWC variables with gender. Then it reports the results of the four stages of multiple regression analysis (as described in 5.3.2.1): first with all variables, and then controlled for topic, genre, and language sparsity.

7.1.2.1 Correlation of the LIWC with gender

The categories which show a significant relationship at the $p < .1$ level are reported, but note that 15 of the 18 variables reported show significance at the $p < .05$ level,

LIWC Variable	Example words	*r*	*p*
Pronouns	*I, our, they*	–.407	.000
Total third-person	*she, them, their*	–.405	.000
Social processes	*talk, us, friend*	–.390	.001
Communication	*talk, share, conversation*	–.365	.002
Hearing	*heard, listen, sound*	–.344	.003
Total first-person	*I, we, me, us*	–.342	.004
Article	*a, an, the*	.333	.005
First-person singular	*I, me, my*	–.329	.005
Other ref to people	non 1st pers pron	–.279	.018
Inhibitions	*block, constrain*	.276	.020
Anger	*hate, kill, pissed*	–.269	.023
Family	*dad, brother, cousin*	–.256	.031
Positive feelings	*happy, joy, love*	–.255	.032
Physical states & functions	*ache, breast, sleep*	–.255	.032
Optimism & energy	*certainty, pride, win*	.248	.037
Humans	*boy, woman, group*	–.226	.058
Negative emotions	*hate, worthless, enemy*	–.225	.059
Money	*cash, taxes, income*	.197	.099

Table 7.2: Correlation of gender with LIWC variables

with 8 of these at the $p < .01$ level (table 7.2).

Perhaps the clearest result is that as well as using more First-person pronouns, women also talk about other people a great deal more than men. This is reflected in the negative correlation with Third-person and Total pronouns, Social words, References to other people, and words reflecting Humans and Family. This is also reflected by the use of words relating to Communication and Hearing. This does not necessarily mean that women are more social than men, merely that they write more about other people and their relationships with them.

As predicted by the literature, women also use more terms relating to emotions, both positive and negative. The exception is the category reflecting Optimism and energy, which relates to greater male use. The reason for this could be that while the category may reflect an 'emotional' state such as optimism, many of the words within do not directly concern emotions. Many of the words in the category reflect aggressive confidence and competitiveness (eg. 'bold', 'determined', 'glorious', 'triumph') which tend to be more associated with masculinity (Schaffer, 1981; Lynn, 1993).

Women also use more terms relating to Physical states and functions, while men talk more about Money. This seems to reflect previous findings that men discuss more impersonal topics, while women prefer those of a more personal nature. Also predicted by the literature was that men use more Articles. Men also use terms reflecting Inhibitions.

7.1.2.2 Regression of the LIWC

Linear regression makes a number of assumptions about dependent variables, not least of which is that they be numerical. Gender is a binary categorial variable — it has two distinct classes — and so linear regression is unsuitable for analysis. The alternative employed here is logistic regression. The outcome is similar to linear regression, in that independent variables are used to best model the dependent. However results are reported differently, and in fact there is little by way of an agreed standard (Peng et al., 2002). Here a very simple approach is adopted, in order to make the format similar to the linear regressions previously reported in chapters 5 and 6.

Rather than listing the correlation of each contributing variable, the Wald statistic

Level of Control	Independent variable	Wald	p	Accuracy
None	Third-person	9.509	.002	
	First-person sing.	6.659	.010	**73.2%**
Topic	Third-person	9.509	.002	
	First-person sing.	6.659	.010	**73.2%**
Genre	Third-person	8.090	.004	
	Articles	4.117	.042	**76.1%**
Sparsity	Third-person	8.090	.004	
	Articles	4.117	.042	**76.1%**

Table 7.3: LIWC logistic regression analyses with gender

is reported. This is a measure each regressors relevance and is much like a t-value. The significance of each is also reported. As a summary statistic of the model produced, logistic regression does not yield an R^2 statistic. Instead, the model is used to classify instances accordingly and the accuracy of this is reported.

As explained in section 5.3.2.1 there were three levels of control applied to variables when performing linear regression analysis. These are once more applied here. The entered dependent variables are those which showed a relationship at the $p < .1$ level with gender. A summary of the four regression analyses can be seen in table 7.3.

Both the analysis with no control, and topic controlled, produced the same regression equation: use of third and first-person singular pronouns, both of which are greater in females. This model produced a classification accuracy of 73.2%. Controlling for genre removes first-person singular pronouns from the analysis so the equation changes, though no further changes occur: use of articles, greater in men, is now introduced into the equation, which produces 76.1% classification accuracy. Clearly the use of pronouns and articles are important in distinguishing between genders within personal blogs.

MRC Variable	*r*	*p*
Kucera & Francis written freq. Mean	.257	.030
Thorndike-Lorge freq. Mean	.234	.049
Concreteness StDev	–.233	.051
Concreteness Mean	–.230	.053
Kucera & Francis written freq. StDev	.229	.054
Thorndike-Lorge freq. StDev	.211	.077
Imagability Mean	–.202	.091

Table 7.4: Correlation of gender with MRC variables

7.1.3 MRC and psycholinguistic differences

Where the LIWC was a dictionary that put words into classes, the MRC database holds psycholinguistic data about the words it contains. As with the LIWC, the MRC database can be used to find variables that correlate with gender and ultimately go some way to explaining the differences between the males and females.

7.1.3.1 Correlation of the MRC with gender

Again, variables which showed a relationship significant at the $p < .1$ level are reported in table 7.4. Unlike the previous LIWC analysis however, only two of the seven show significance at the $p < .05$ level, and none at a higher level. It has already been established that the results of the Dolby categories are unreliable (see section 5.4.2.3), so these are ignored here.

Both the significant Kucera and Francis positive correlation, and that of similar strength with the Thorndike-Lorge score suggests that women use less frequent language than men. This does not tie with the previous finding for women scoring higher on the LIWC factor ‘Immediacy’, since use of lower frequency words would suggest less immediate writing. Correlation with not just the mean but standard deviation of these measures suggests that men also use language with a greater range of frequency.

Just short of being significant are the results for Concreteness. Not only do women use more concrete language, but they use a greater range of language. This does not

Dependent Variable	Independent variable	Wald	*p*	Accuracy
Gender	Concreteness StDev	15.115	.000	
	Thorndike-Lorge freq. StDev	15.246	.000	**77.5%**

Table 7.5: MRC logistic regression analysis of gender

confirm previous hypotheses that men use more concrete language since they use more articles. However, conceptually at least, concreteness seems to follow from the more immediate style of female writing.

7.1.3.2 Multiple regression of the MRC

As before, correlating variables were entered into a logistic regression with gender as the dependent variable. The summarised results can be seen in table 7.5).

Interestingly, none of the mean variables are retained in the model. Females' greater range of concreteness levels, and males' greater range of frequency of language produces a model capable of 77.5% classification accuracy.

7.1.4 Discussion

The LIWC analysis serves to confirm a number of prior findings in the field of language and gender differences within the blog corpus: women use more social language, and have a more immediate style of language; females also use more pronouns, while males use more articles.

Results from the MRC paint a less clear picture. That women use more concrete language seems to go against the commonly accepted association of frequency of articles with concreteness. However, it does appear to follow from the more immediate style of writing. The finding that it is men who use more common language however seems to contradict this.

The limitations of the frequency measures of the MRC, and particularly the Thorndike-Lorge frequency scale have been highlighted previously: they achieve only between 80-90% coverage and are arguably out-of-date. The concreteness data also has issues of coverage: data is only available for 8000 of the 150,000 words of the MRC and

captures only 75% of the words in the blog corpus. These concerns throw doubt on validity of the regression result.

7.2 Bottom-up Approaches to Gender Differences

Despite concerns over reliability of the MRC as it has been applied here, the LIWC appears to confirm commonly detected differences in the language of gender. In this section, the two unitary data-driven measures are used to explore further language differences.

7.2.1 Contextuality

Heylighen and Dewaele (2002) applied their measure of contextuality, the F-measure, to texts of known gender and found a distinct difference between the sexes. Females score lower, preferring a more contextual style, while men prefer a less contextual style. This result was taken to be consistent with previous findings from sociolinguistic and psychological studies.

In section 4.1 the F-measure was calculated for a number of sub genres of the BNC, in order to investigate the contextuality pattern of those genres. A number of these are marked for author gender, as are Gill's e-mail corpus and the blog corpus of this thesis. This data was used to calculate the average contextuality for texts of each genre dependent on author gender.

7.2.1.1 Method

The F-measure of texts was calculated as it was in sections (4.1.1 and 6.3.1.1). A number of the genres in that study were marked for gender, and so calculation of average male and female scores was possible. Note that in being consistent with the application of the F-measure to personality differences, outliers in the blog corpus are excluded. This leaves 45 females and 23 males.

Genre	Male	Female
Fiction prose Adult	47.8	45.0
E-Mail Corpus	*53.1*	*49.5*
Blog Corpus	*54.8*	*52.4*
Non academic Social Science	59.5	52.1
Academic Social Science	60.5	60.8

Table 7.6: Average F-score for male and female authors in selected genres

7.2.1.2 Results

Table 7.6 shows the average F-score for males and females in the genres for which data was available. For both genders, the ordering of genres remains as shown in table 4.1. Females score lower in four out of the five genres. Within the blog corpus this difference is significant (t=–2.75, DF=66, $p<.01$). The exception is when the writing is for academic purposes. Here there is little difference between male and female F-scores; both are relatively high. It appears that while females prefer a more contextual style, when required they can adopt a less contextual style similar to that projected by males.

As in the personality analysis (section 6.3), a closer inspection is made of the component parts-of-speech. Note that since women are more contextual in style, a negative correlation between gender and those related parts of speech (pronouns, verbs, adverbs and interjections) is expected; a positive correlation is expected for those POS considered least contextual (nouns, adjectives, prepositions and articles). The results of the Pearson correlation between gender and POS relative frequency can be seen in table 7.7.

With the exception of adverbs, all the POSs correlated in the expected directions. Most significantly, men use more articles, while women use more pronouns (cf. the LIWC findings of section 7.1.2). Women also use considerably more verbs, and also following previous findings, men use more prepositions.

Trait	*r*
Noun	.224
Adjective	.109
Preposition	.202
Article	.317**
Pronoun	–.444**
Verb	–.344**
Adverb	.144
Interjection	–.073

Table 7.7: Correlation between POS frequency and gender

Note: two-tailed, *$p<0.05$, **$p<0.01$

7.2.2 Word Frequency

During the investigation of personality differences (section 6.4) average word frequency rank appeared to show a similar relationship with the F-measure. With the exception of Extraversion, the pattern of score distributions was strikingly similar, even if correlation scores differed somewhat. Following this relationship, it is expected that women, with their more contextual style, use more frequent language. Similarly, this would also follow from their more immediate style (section 7.1.1).

7.2.2.1 Results

Summary statistics for both genders on average rank, along with the percentage of words for which rank was available, can be found in table 7.8. Again, outliers were excluded, leaving 45 females and 24 males. It is immediately clear that there is very little difference between the genders where average rank is concerned. The correlation between rank and gender is practically zero, and the coverage provided by the rank list is practically the same. This result brings into further question the frequency findings from the MRC.

	Average rank	Perc words ranked
Female mean (SD)	3557 (555)	96.9 (.99)
Male mean (SD)	3564 (563)	96.8 (.96)
Pearson's *r*	.005	–.056

Table 7.8: Summary statistics for average rank and gender

Note: two-tailed

7.2.3 Discussion

As has been found previously (Heylighen and Dewaele, 2002) women have a more contextual style of writing than men. As was discussed in section 2.3 this ties well with the finding that fiction is more contextual than non-fiction. This was further linked to females having a more involved narrative style, compared to males more informational style. This result also seemed to follow from the more immediate style of female language.

The exception of course was when in the least contextual genre, in which women are able to adopt a style equally as un-contextual as men. The F-measure has also presented further evidence that men use more articles and prepositions, while women use more pronouns. Women also use more verbs.

There was no finding for frequency between genders. From calculations here, it appears the women use language of a similar frequency to men, not less frequent as had been predicted by the MRC.

7.3 Summary

This chapter has adopted both the top-down and bottom-up analytic techniques that had previously been applied to personality. Using the LIWC confirmed a number of previous findings for gender differences in language. In fact, proving reasonably successful upon logistic regression, women's greater use of pronouns and men's higher use of articles seem to be among the most important differences.

The MRC also proved successful under regression, despite apparent internal con-

flicts between findings. Ignoring criticisms of these results for a moment, it appears that the dictionary-based approaches do appear to perform reasonably well with gender.

The bottom-up approaches proved perhaps less fruitful, in part due to the exclusion of the n-gram analysis. The F-measure proved significantly capable of detecting male and female writing, and further confirmed previous hypotheses regarding gender use of parts-of-speech. There appears to be no effect for frequency of language between genders.

On the methodological side, this chapter has confirmed a number of previous findings for gender differences of language. This suggests that the techniques applied are robust at detecting such differences. Following this is that the differences identified for personality traits and language are genuine differences and not merely artifacts of the technique.

Chapter 8

Conclusion

8.1 Summary of thesis

This thesis aimed to investigate how individual differences affect the language produced in personal diary weblogs, or blogs. A secondary objective was to investigate the distinctiveness, yet representativeness of blogs as a genre. More formally, the thesis has addressed the following general hypotheses concerning whether and to what extent:

Hypothesis 1: Blogs are distinct yet representative of more general language.

Hypothesis 2: Personality is projected linguistically in blogs.

Hypothesis 3: Gender is projected linguistically in blogs.

This thesis has found support for all three hypotheses. This section traces the path through the thesis, and summarises the findings which led to this conclusion.

The first chapter began with quotes from two different blogs. These were used to illustrate that differences between individuals can be detected through what they write, through their language. The section discussed the notion of studying gender and personality differences in language, and further, how these can be investigated in computer-mediated communication (CMC), specifically blogs. The aim of the chapter was to introduce the main focus of the thesis. In addition to this the objectives, boundaries and structure of the thesis were all outlined.

The second chapter constituted a review of the literature which informed the work

of the thesis. It was thematically divided into two sections: the first concerned the data to be studied, the second the tools to be used. The first section began by providing background on personality trait theory, highlighting the specific model used here. Alternative theories were briefly discussed before previous findings relating personality to differences in language use were reviewed. Many of these studies have focused solely on two factors of personality, Neuroticism and Extraversion, while most have studied language through speech. There are, however, some studies of particular relevance to the work of this thesis (Pennebaker & King, 1999; Gill, 2004). Following personality, gender differences of language were discussed, first with a general introduction to the field, followed by a summary of previous findings. Due to the limited size of the corpus, the approach of treating men and women as whole groups — ignoring intra-gender differences — was chosen and the studies that were reported reflected this. Though there are some inconsistencies in findings, there are a number that have been replicated across many studies and various genres of text.

Since blogs are treated here as a genre, it was appropriate to attempt to indicate what is meant by genre, though there is no clear definition. With this in mind, genre studies of CMC were reported. Having paid specific attention to weblogs, it is clear that they are considered a distinct genre (Herring et al., 2004a). Following this initial discussion of work looking at weblogs, they were discussed in greater depth. Background was given as to why blogs are an emerging topic of great interest, including a summary of the small yet rapidly increasing volume of work in the field. The final background section, one of the most important, looked at studies of language in CMC, and specifically weblogs. The most interesting point is that despite being a distinct genre, weblogs share a general property with many forms of CMC: the language of weblogs has properties of both written and spoken language.

The second section introduced the tools and analytic techniques to be used in the thesis. These were broadly divided into two categories: top-down, or dictionary-based approaches; and bottom-up techniques, derived more directly from data. The chapter concluded by noting the methodological issues arising from some of these approaches.

The third chapter introduced the corpus which would form the basis of the work. Not only did it describe how the corpus was created but also provided some demo-

graphic statistics on the corpus. It also described the sub-division of the corpus that would be used in certain stages of analysis. The corpus was found to consist of more women than men, as previous findings had suggested. Women also write more than men, reflected in longer texts and more frequent posts. No such effects were found for personality. The distribution of Openness scores across individuals was unusual, which would affect the subsequent analysis.

The fourth chapter explored the linguistic properties of blogs, in an attempt to explore the situation of the genre. Analysis using two unitary linguistic measures situated blogs in the context of other genres. These were both written and spoken, and included genres that were also computer-mediated in nature. The first of these was a measure of contextuality (the F-measure, due to Heylighen & Dewaele, 2002) which ordered genres similarly to previous results based on factor analysis. This showed that while blogs are less contextual than speeches, they are more contextual than biographies. The two most interesting findings were that similarly to e-mails and Mailing List texts, blogs were situated between written and spoken genres. However, despite the similar scores for the other two CMC-based genres, blogs were significantly less contextual than e-mails. The second measure was an approximation of word frequency, calculated as a rank sum of words, word rank derived from use in the British National Corpus. Ordering of genres appeared less systematic with this approach; however, a similar positioning for blogs with respect to spoken and written genres, and CMC genres was evident. Though the measures used in this chapter were simple unitary measures, they were chosen since they could be used to explore individual difference as easily as they could compare corpora. They both provided results to show that though blogs may be distinct from other forms of CMC, they also share similar properties.

The fifth chapter was the first of two which concentrated on personality differences within the blog corpus. This chapter employed top-down or dictionary-based approaches to linguistic content analysis. The first step was the replication of a factor analysis using selected variables from the LIWC (Pennebaker & King, 1999; cf. Gill, 2004). With only minor differences, the three strongest factors were successfully replicated with blogs (compared with personal student essays and e-mails in the previous studies). However, correlating these factors with personality revealed differences in the

nature of the language in the texts. In blogs, there were strong positive effects for Neuroticism with 'Making distinctions', Extraversion with 'Immediacy', and Extraversion and Openness with the 'Social past'; there were strong negative effects for Extraversion and Agreeableness with 'Making distinctions', Openness and Agreeableness with 'Immediacy', and Conscientiousness with the 'Social past.'

Following this, correlation and multiple regression were carried out with the full variable set of the LIWC. While a number of the variables correlated with the personality traits (with the exception of Conscientiousness), very few remained in the regression equations account for very little variance: between 0% and 28%. A similar analysis was then carried out using the MRC Psycholinguistic Database: effects of correlation were unclear, and once more regression explained very little: between 0% and 26%. These results prompted a review of the limitations of dictionary-based approaches noted at the end of chapter 2.

Chapter 6 continued to focus on variation according to personality, but adopted a set of bottom-up, more data-driven, techniques. The first technique employed was a word n-gram comparison of the stratified corpora. High and Low extreme groups of each personality were submitted to a three way comparison with a group of neutral personality. This analysis was carried out for only the four normally distributed traits since there was no low Openness group for comparison. The product of this was a set of representative n-grams for each extreme personality type. These were robust of the over-influence of individuals since at least half the group were required to use the collocations reported. These results were placed back within the scope of the individual, by studying their relative frequencies in each blog with relation to the specific personality of the author. This was not to improve upon the group results, or in anyway determine their correctness, but to see which of them could be used to explain more fine grained variation within personality dimensions. The majority correlated as expected (high group n-grams positively, and vice versa) and a number of them did so significantly. Upon regression fewer remained, but still enough to account for between 40% and 66% of the trait variance.

The remainder of the chapter returned to the measures of contextuality and word frequency used for the genre study of chapter 4. The only significant correlation sug-

gested that highly Agreeable individuals are less contextual, reflecting their natural consideration for other people, more specifically the lack of shared extra-linguistic context between themselves and their readers. Approaching significance was a similar effect for Openness, suggesting that as previously hypothesised (Heylighen & Dewaele, 2002) the factor of intellect reflects a preference for less contextual language. There were also some effects observed by comparing the average scores of the stratified corpora, though none of these were significant. In an extra analysis, the F-measure, which was constructed upon the principles of deixis, was correlated against a number of deictic measures derived directly from the blog data. The significant correlations found confirmed that the F-measure indeed relates to contextuality of situation.

The strongest correlation for average word frequency rank was with Openness, showing that highly Open individuals are more likely to use infrequent words. When stratified averages were computed, patterns were strikingly similar to those for the F-measure, which was unsurprising given the highly significant correlation between the two measures within the blog corpus. Despite this, none of the group differences were significant.

In the seventh chapter, attention turned from personality to gender. The set of analyses which had thus far been used to investigate differences in personality were employed to examine differences between the language of men and women. Both the methods of the previous dictionary-based and data-driven chapters were used. Some results of the former were consistent with previous findings, such as strong relationships for women with immediate writing and social references; others stood merely to re-emphasise the limitations of the technique. Using the F-measure across a number of genres, including the blog corpus showed that men are significantly less contextual than women in their writing style. The only genre for which this was not found, was the least contextual genre from chapter 4 — academic writing — which shows that women can adapt their writing style as required. There was no difference in the frequency of words used, according to the measure used here. From the results which echoed previous work in gender and language, techniques were shown to be capable of identifying robust differences in language. This suggests that the personalty differences found in this thesis are indeed genuine differences, and not just by-products of

the analyses.

8.2 Contributions

The main contributions of this thesis, listed by the field to which they contribute, are that it has:

Personality

- Demonstrated that personality is projected through language in a popular, emerging CMC environment.
- Explored the specific linguistic features associated with different personality dimensions.
- Extended the study of language and personality giving as much attention to Agreeableness and Conscientiousness as to Extraversion and Neuroticism.

Gender

- Demonstrated that gender is projected through language in a popular, emerging CMC environment.

Empirical Linguistics

- Demonstrated that blogs are a distinctive yet representative genre.
- Confirmed the utility of n-gram context in linguistic studies.
- Implemented and extended a variety of corpus comparison techniques.
- Gathered, annotated and analysed a personality informed corpus of blogs.

8.3 Limitations of thesis

Chapter 1 of this thesis introduced boundaries which restricted the scope of the research. This section discusses further limitations that have become apparent in the course of its development.

The first limitation draws on the idea of generalisabilty of findings. Koppel et al. (2002) identified different features in the language of fiction and non-fiction in the BNC which could be used to distinguish between genders; in fact gender differences were similar to genre differences. This work has found some different relationships with personality traits than have been found previously (cf. Gill, 2004; Pennebaker & King, 1999). However, despite these different relationships, factor analysis appears to find a similar structure to language in blogs as it does in other genres (see section 5.1). Despite the distinctiveness of the genre, blogs do share many traits with other forms of computer-mediated communication. It seems that many of the findings of this thesis may only reflect language from personal written monologue genres, but a number will generalise beyond.

Another limitation of the thesis, which would in itself further limit generalisabilty, is the limited corpus size. Despite a significantly larger contribution of text per subject than previous work (cf. Gill, 2004), the corpus only consists of 71 subjects. This limited the strictness which could be applied when stratifying the corpus into extreme personality groups (see section 3.4.5). 71 subjects, when dealing with such a complex level of individual differences (five personality traits plus gender) is too few on which to form a solid base for work in automatic text classification.

A related criticism that could be levelled at the corpus is the lack of balance and control. The corpus was not balanced for gender, for example, and despite an upper limit being placed on text size for group analysis, there was no lower limit below which subjects were excluded. More subjects would have permitted more rigorous limits to have been enforced.

A further limit imposed by the corpus size relates to the methodology of studying gender differences, as discussed in section 2.3.1. There are many who feel that treating men and women as two homogeneous groups ignores possible language differences within each gender. With so few subjects, the best recourse was to adopt the general

methodology, treating each individual as the same as the rest of their gender group.

8.4 Future work

This section outlines future work that would allow both further investigation of many of the results found thus far, and to answer some of the limitations discussed above.

The most obvious direction for future work given the limitations imposed by the small number of subjects is a large scale reproduction. As highlighted in section 2.5 the number of blogs has increased at an incredible rate since the data gathering period of this thesis. Not only does this create a much larger potential data pool, but it would make it easier to recruit subjects. A larger corpus would allow for better control of the gender split, and the amount of data drawn from each subject. It would also allow for better control of factors when the corpus is stratified by personality dimension. With a larger base of subjects, it would be possible to use only those subjects who were extreme on just one dimension and still retain a group with a size worth studying. With respect to gender, a larger corpus would also allow the exploration of intra-group differences.

Another potential future investigation draws on work carried out by Gill (2004), and would potentially better focus work in this area. Gill investigated human perception of personality traits through the texts he had gathered for his e-mail corpus. He found good level of agreement between raters on target personality, but only rater-target agreement for Extraversion and Psychoticism. He found that subjects were particularly poor at determining the target level of Neuroticism of a text. This is a similar result to that found by Markey and Wells (2002), who studied perception of personality in chat rooms, finding only Extraversion and Openness to achieve significant judge-target agreement.

One of the areas which this work is intended to inform is that of personality-rich text generation. It seems that in light of the Gill's perception result for example, enabling agents to generate e-mail like texts with different levels of Neuroticism would make little difference to how they are read. That is not to say Neuroticism should be ignored, merely that work could be focused on the other aspects of personality first. This

method would need to be employed from scratch since previous findings have studied areas of markedly different properties to blogs. The approach would be employed with blog texts gathered with five factor model personality data in order to best direct focus in evaluating language differences.

This approach would also identify those individuals whose personalities are most easily observed. This would allow further focusing on the linguistic features which are exhibited by these individuals. This would identify not only those features which, according to the data, relate to the projection of personality, but those which must relate to perception of personality.

An analogous direction for future work is automatic classification of blogs by author personality, or in fact by gender. This is not only of interest commercially, with the ability to profile consumers highly sought after, but has implications for other areas of language-based research. Sentiment analysis for example: different types of people may express sentiment in different ways, and so automatic detection and classification of sentiment may be aided by applying a different model based on the author in question. Of further interest is to compare the performance of human judges against automated systems of classification. Human judgement is the best possible baseline to compare an automated system to in this situation.

A further interesting study would be to investigate the anomaly of the Openness scores distribution (see section 3.4.3). To do this, a suitable comparison set would need to be collected, preferably in the same way. One approach to this would be to advertise the large scale reproduction so as to attract non-bloggers alongside those who will provide textual data. This data would serve no other purpose other than to compare factors such as personality score distribution. This would either confirm or falsify the hypothesis that people who keep blogs are generally more Open.

8.5 Final words

In concluding this thesis, it seems most appropriate to reiterate the central findings: blogs are both a distinct genre yet representative of more general language; personality and gender are both projected through language in a computer-mediated environment.

These are both important findings. The former, while only a secondary objective of this work provides justification for exploration of a new and rapidly expanding potential corpus for researchers. The second finding is important because it has often been considered that individuals could hide behind words on the anonymous medium of the internet. This finding has significant implications for knowing how individuals portray themselves by their choice of language. This is particularly useful if trying to deliberately portray certain character traits, as in personality rich natural language generation.

Finally, the thesis returns to where it began, with two samples from blogs:

- I don’t know how many of you ever experience a similar thing, but well, I just see possibilities around me everyday to be evil, and I have to make an active decision NOT to do it.
- I am writing this hesitantly, because I am conscious that you reading this may be thinking ”What category do I fit into?”

It turns out that the first blogger scores particularly low on Agreeableness, as the reader no doubt anticipated from the short excerpt alone. The second author should indeed be worried.

Appendix A

Collection and Construction of Corpus

Item Number	Item Text	Factor	Direction
1	Tend to vote for conservative political candidates.	O	–ve
2	Have frequent mood swings.	N	+ve
3	Am not easily bothered by things.	N	–ve
4	Believe in the importance of art.	O	+ve
5	Am the life of the party.	E	+ve
6	Am skilled in handling social situations.	E	+ve
7	Am always prepared.	C	+ve
8	Make plans and stick to them.	C	+ve
9	Dislike myself.	N	+ve
10	Respect others.	A	+ve
11	Insult people.	A	–ve
12	Seldom feel blue.	N	–ve
13	Don’t like to draw attention to myself.	E	–ve
14	Carry out my plans.	C	+ve
15	Am not interested in abstract ideas.	O	–ve
16	Make friends easily.	E	+ve
17	Tend to vote for liberal political candidates.	O	+ve
18	Know how to captivate people.	E	+ve
19	Believe that others have good intentions.	A	+ve
20	Do just enough work to get by.	C	–ve
21	Find it difficult to get down to work.	C	–ve
22	Panic easily.	N	+ve
23	Avoid philosophical discussions.	O	–ve
24	Accept people as they are.	A	+ve
25	Do not enjoy going to art museums.	O	–ve
26	Pay attention to details.	C	+ve
27	Keep in the background.	E	–ve
28	Feel comfortable with myself.	N	–ve
29	Waste my time.	C	–ve
30	Get back at others.	A	–ve
31	Get chores done right away.	C	+ve
32	Don’t talk a lot.	E	–ve
33	Am often down in the dumps.	N	+ve
34	Shirk my duties.	C	–ve
35	Do not like art.	O	-ve
36	Often feel blue.	N	+ve
37	Cut others to pieces.	A	–ve
38	Have a good word for everyone.	A	+ve
39	Don’t see things through.	C	–ve
40	Feel comfortable around people.	E	+ve
41	Have little to say.	E	–ve

Table A.1: 41 items of the IPIP online implementation inventory (Buchanan, 2001)

```
<?xml version="1.0" encoding="UTF-8"?>

<!ENTITY text "#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK">
<!ENTITY content "COMMENTARY|DIARY|JOURNAL|LINKAGE|FRIDAYFIVE|QUIZ|PERSONALDATA">

<!ELEMENT BLOG (BLOGHEADER*, (BLOGEXTRA*, BLOGBODY)+,BLOGEXTRA*,BLOGFOOTER*)>
   <!ELEMENT BLOGHEADER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
   <!ELEMENT BLOGEXTRA (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
   <!ELEMENT BLOGBODY (DAY|HOLIDAY)+>
      <!ELEMENT DAY (DAYHEADER*,POST+,DAYFOOTER*)>
         <!ELEMENT DAYHEADER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
            <!ELEMENT POST (POSTTITLE*, POSTBODY, POSTFOOTER*)>
            <!ELEMENT POSTTITLE (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
            <!ELEMENT POSTBODY (PERSONAL|COMMENTARY|LINKAGE|FRIDAYFIVE|QUIZ|PERSONALDATA)+>
               <!ELEMENT PERSONAL (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK|DIARY|JOURNAL)*>
               <!ELEMENT JOURNAL (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT DIARY (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT COMMENTARY (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT LINKAGE (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT FRIDAYFIVE (FFINTRO*, (FFQUESTION, FFANSWER)+, FFOUTRO*)>
                  <!ELEMENT FFINTRO (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT FFQUESTION (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT FFANSWER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT FFOUTRO (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT QUIZ (QUIZINTRO*, QUIZRESULT, QUIZLINK, QUIZDISCUSSION*)>
                  <!ELEMENT QUIZINTRO (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT QUIZRESULT (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT QUIZLINK (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                  <!ELEMENT QUIZDISCUSSION (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
               <!ELEMENT PERSONALDATA (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>

                     <!ELEMENT BOLD (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT IMAGE (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT ITALIC (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT LINK (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT LIST (LISTITEM, LISTCONTENT)+>
                        <!ATTLIST LIST blogorlist (blog|justlist) "justlist">
                        <!ELEMENT LISTITEM EMPTY>
                        <!ELEMENT LISTCONTENT (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT QUOTE
(LYRICS|CONVERSATION|OTHERBLOG|WEBSITE|FACTSNFIGURES|EMAIL|OTHERMEDIA|UNKNOWN|NEWS)>
                        <!ELEMENT LYRICS (SONG|POEM)>
                           <!ELEMENT SONG (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                           <!ELEMENT POEM (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT CONVERSATION (INSTANTMESSENGER|REALLIFE)>
                           <!ELEMENT INSTANTMESSENGER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                           <!ELEMENT REALLIFE (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT OTHERBLOG (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT WEBSITE (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT FACTSNFIGURES (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT EMAIL (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT OTHERMEDIA (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT UNKNOWN (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT NEWS (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                     <!ELEMENT OWNWORK (OWNSONG|OWNPOEM|FICTION|NONFICTION|PREVIOUSPOST|DREAM)>
                        <!ELEMENT OWNSONG (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT OWNPOEM (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT FICTION (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT NONFICTION (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT PREVIOUSPOST (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
                        <!ELEMENT DREAM (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>

            <!ELEMENT POSTFOOTER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
         <!ELEMENT DAYFOOTER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
      <!ELEMENT HOLIDAY EMPTY>
   <!ELEMENT BLOGFOOTER (#PCDATA|BOLD|IMAGE|ITALIC|LINK|LIST|QUOTE|OWNWORK)*>
```

Figure A.1: XML encoding of tagset for blog encoding

Appendix B

Previous Results

	Factor 1: Immediacy (22.4% variance)	Factor 2: Making Distinctions (10.3% variance)	Factor 3 The Social Past (9.8% variance)	Factor 4: Rationalization (8.6% variance)
Exclusive		.674		
Discrepancies	.485	.427		
Tentative		.644		
Causation				.598
Present Tense	.593		.596	
Negations		.579		
Insight				.627
Words >6 letters	–.683			
Articles	–.765			
First-person sing.	.823			
Positive emotion			–.469	
Social			.425	
Past Tense			.856	
Inclusive		–.463		
Negative emotion				–.443

Table B.1: Pennebaker and King's rotated factor loadings for exploratory analysis of 15 LIWC variables.
Note: variables ordered as in this study.

	Factor 1: Immediacy (19.4% variance)	Factor 2: Making Distinctions (12.8% variance)	Factor 3 The Social Past (11.5% variance)	Factor 4: Rationalization (9.6% variance)
Exclusive		.697		
Discrepancies	.553			
Tentative		.604		
Causation				.707
Present Tense	.788			
Negations		.651		
Insight	.561			
Words >6 letters	–.669			
Articles	–.522			–.506
First-person sing.	.587			
Positive emotion			.569	
Social			.676	
Past Tense			.732	
Inclusive		–.546	–.465	
Negative emotion				.737

Table B.2: Gill's rotated factor loadings for exploratory analysis of 15 LIWC variables.
Note: variables ordered as in this study.

original	Factor 1: Making Distinctions (21.9% variance)	Factor 2: Immediacy (14.5% variance)	Factor 3: The Social Past (11.8% variance)
Exclusive	.697		
Discrepancies	.593		
Tentative	.581		
Negations	.598		
Articles		–.710	
First-person singular		.622	
Present Tense		.812	
Words >6 letters		–.556	
Insight			
Positive emotions			.755
Social			.658
Inclusive	–.561		–.457
Past Tense			.755

Table B.3: Gill's rotated factor loadings for exploratory analysis of 13 LIWC variables.
Note: variables ordered as in this study.

Dimension	N	E	O	A	C
Making Distinctions	**.05**	**–.14****	**.06**	**–.05**	**–.13****
Exclusive	.00	–.08*	.10	–.06	–.08*
Tentative	.06	–.14**	.11*	–.02	–.06
Negations	.05	–.12**	.00	–.04	–.15**
Inclusive	–.01	.07*	.01	.03	.06
Immediacy	**.10***	**.04**	**–.16****	**.07****	**–.02**
First-person singular	.13**	.04	–.13**	.07*	.01
Articles	–.09*	–.09*	.13**	–.15**	–.04
Words $>$ 6 letters	–.03	–.04	–.16**	–.03	.06
Present tense	.06	.01	–.15**	.04	.00
Discrepancies	.05	–.03	–.01	–.02	–.07*
The Social Past	**.04**	**.00**	**.08***	**–.02**	**–.04**
Past tense	.03	.04	–.03	.06	–.06
Social	–.01	.12**	.02	.00	.02
Positive emotions	–.13**	.15**	–.06	.07*	.07*
Rationalization	**–.06**	**.02**	**–.03**	**.07**	**.04**
Insight	.03	–.02	.07*	.05	–.01
Causation	.03	–.08*	–.08*	.00	–.07*
Negative emotions	.16**	–.08*	.05	–.07*	–.15**

Note. $N = 841$. Two variables are coded onto two factors: Present tense is also part of The Social Past; Discrepancy is a part of Making Distinctions. The following variables are negatively loaded on their respective factors: Articles, Words of more than 6 letters, Inclusive, Present tense (for The Social Past only), and negative emotion. The ordering of the factors has been altered to match that of the present study.
$*p < .05$. $**p < .001$, two tailed.

Table B.4: Pennebaker and King's correlation of 15 LIWC categories with personality scores

LIWC factor	EPQ-R Dimension Psychoticism	Extraversion	Neuroticism
Making Dist.	**.11**	**−.02**	**−.13**
Exclusive	−.01	−.10	−.02
Negations	−.02	−.08	−.03
Tentative	.13	.00	−.14
Discrepancies	.13	.09	.04
Inclusive	−.11	−.02	.26**
Immediacy	**−.10**	**−.07**	**.14**
Present tense	−.06	−.10	.14
Words $>$ 6 letters	−.01	−.05	.04
First-person Sing.	−.23*	−.12	.16
Insight	.07	.00	.01
Articles	.12	.11	−.02
The Social Past	**.01**	**.09**	**−.21***
Past tense	−.09	.06	−.19
Social	.02	.01	−.05
Positive emotion	.07	.15	−.13
Rationalisation	**.04**	**−.01**	**.01**
Negative emotion	.20*	.13	−.07
Causation	.04	−.05	.08

Note. $N = 105$. Two variables are coded onto two factors: Articles is also part of Rationalization; and Inclusive is a part of The Social Past. The following variables are negatively loaded on their respective factors: Words of more than 6 letters, Articles, and Inclusive words. LIWC categories are ordered as they load onto their Factor. The ordering of the factors has been altered to match that of the present study.
$*p < .05$. $**p < .001$, two tailed.

Table B.5: LIWC Factors and Simple Correlations with EPQ-R Scores using E-mail data and 4 LIWC factor model

	EPQ-R Dimension		
	Psychoticism	Extraversion	Neuroticism
Making Dist.	**.11**	**−.03**	**−.11**
Exclusive	−.01	−.10	−.02
Negations	−.02	−.08	−.03
Discrepancies	.13	.09	.04
Tentative	.13	.00	−.14
Inclusive	−.11	−.02	.26**
Immediacy	**−.11**	**−.08**	**.12**
Present tense	−.06	−.10	.14
Articles	.12	.11	−.02
First-person Sing.	−.23*	−.12	.16
Words $>$ 6 letters	−.01	−.05	.04
The Social Past	**.04**	**.11**	**−.24***
Past tense	−.09	.06	−.19
Social	.02	.01	−.05
Positive emotion	.07	.15	−.13

Note. $N = 105$. One variable is coded onto two factors: Inclusive is a part of The Social Past. The following variables are negatively loaded on their respective factors: Articles, Words of more than 6 letters, and Inclusive words. LIWC categories are ordered as they load onto their Factor. Immediacy and Making Distinction factors have been switched to aid comparison.
$*p < .05$. $**p < .001$, two tailed.

Table B.6: LIWC Factors and Simple Correlations with EPQ-R Scores and E-mail data using 3 LIWC factor model.

Appendix C

Extra Results From This Work

Dimension	N	E	O	A	C
Factor 1	**.212**	**–.135**	**.027**	**–.167**	**.059**
Exclusive	.133	–.079	.143	–.189	–.061
Discrepancies	.339**	–.251*	–.118	–.290*	–.035
Tentative	.140	–.144	–.107	–.198	.060
Causation	.056	.025	.115	–.005	.069
Present tense	.159	.200	.009	–.092	.102
Negations	.163	.020	–.222	–.245*	.098
Insight	–.111	.063	.106	.094	.075
Factor 2	**.072**	**.040**	**–.345****	**–.310****	**.012**
Discrepancies	*.339***	*–.251**	*–.118*	*–.290**	*–.035*
Present tense	*.159*	*.200*	*.009*	*–.092*	*.102*
Negations	*.163*	*.020*	*–.222*	*–.245**	*.098*
– Words $>$ 6 letters	.020	–.055	.290*	.262*	.034
– Articles	–.072	.031	.136	.255*	–.054
First-person singular	–.017	.175	–.098	–.081	–.060
Factor 3	**–.099**	**.268***	**.282***	**.172**	**–.086**
Insight	*–.111*	*.063*	*.106*	*.094*	*.075*
Positive emotions	–.043	.162	.127	.069	–.060
Social	–.035	.238*	.195	.037	–.109
Factor 4	**.005**	**–.065**	**.090**	**–.033**	**–.135**
Past tense	.011	–.116	–.028	–.125	–.157
Inclusive	–.012	.015	.249*	.094	–.091
– Negative emotions	.158	–.038	–.097	–.202	–.046

Table C.1: Correlation of LIWC factors (15 variables) with personality scores

Note: $n = 71$, two tailed, *$p<0.05$, **$p<0.01$. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading.

Dimension	N	E	O	A	C
Factor 1	**.253***	**–.183**	**–.057**	**–.253***	**.025**
Exclusive	.133	–.079	.143	–.189	–.061
Tentative	.140	–.144	–.107	–.198	.060
Discrepancies	.339**	–.251*	–.118	–.290*	–.035
Negations	.163	.020	–.222	–.245*	.098
Present tense	.159	.200	.009	–.092	.102
Factor 2	**–.001**	**.115**	**–.212**	**–.171**	**.052**
Discrepancies	*.339***	*–.251****	*–.118*	*–.290****	*–.035*
– Articles	–.072	.031	.136	.255*	–.054
– Words > 6 letters	.020	–.055	.290*	.262*	.034
First-person singular	–.017	.175	–.098	–.081	–.060
Present tense	*.159*	*.200*	*.009*	*–.092*	*.102*
Insight	–.111	.063	.106	.094	.075
Factor 3	**–.036**	**.220**	**.219**	**.103**	**–.091**
Positive emotions	–.043	.162	.127	.069	–.060
Social	–.035	.238*	.195	.037	–.109
Negative emotions	.158	–.038	–.097	–.202	–.046
Factor 4	**–.065**	**.026**	**.204**	**.082**	**–.125**
– *Negative emotions*	*.158*	*–.038*	*–.097*	*–.202*	*–.046*
Inclusive	–.012	.015	.249*	.094	–.091
Past tense	.011	–.116	–.028	–.125	–.157

Table C.2: Correlation of LIWC factors (14 variables) with personality scores
Note: $n = 71$, two tailed, *$p<0.05$, **$p<0.01$. Italics are used to indicate variables loading on a second factor. '–' is used to indicate a negative factor loading.

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> <P>	1	155	0.0029	61	0.0008	101	0.0011	3.74	1.43	2.61	86.81****	5.08*	57.99****	+		
<P> <P> <P>	2	63	0.0012	16	0.0002	10	0.0001	5.80	0.54	10.71	51.07****	2.39	76.11****	+		
IN NP1	4	30	0.0006	20	0.0003	91	0.0010	2.21	3.94	0.56	7.76**	39.56****	8.25**			+
NP1 <P>	6	156	0.0029	223	0.0028	415	0.0045	1.03	1.61	0.64	0.08	34.43****	24.14****			+
OK <P>	13	8	0.0001	39	0.0005	8	0.0001	0.30	0.18	1.70	12.01***	26.96****	1.11		+	
<EOP> <SOP> SO	14	8	0.0001	21	0.0003	1	0.0000	0.56	0.04	13.61	2.08	25.34****	10.54**			−
<P> OK	15	2	0.0000	24	0.0003	2	0.0000	0.12	0.07	1.70	14.38***	25.23****	0.28		+	
THAT I	16	63	0.0012	186	0.0023	178	0.0019	0.50	0.83	0.60	25.13****	3.19	12.90***	−		
AND I	18	74	0.0014	208	0.0026	194	0.0021	0.52	0.81	0.65	24.90****	4.56*	10.64**	−		
A COUPLE	19	8	0.0001	6	0.0001	41	0.0004	1.96	5.92	0.33	1.58	24.48****	10.21**			+
A COUPLE OF	20	7	0.0001	4	0.0001	35	0.0004	2.58	7.58	0.34	2.40	24.03****	8.44**			+
IN NP1 <P>	21	16	0.0003	9	0.0001	48	0.0005	2.62	4.62	0.57	5.63*	24.00****	4.21*			+
<P> AS	24	38	0.0007	68	0.0009	145	0.0016	0.82	1.85	0.45	0.94	18.53****	22.66****			+
<P> OK <P>	25	2	0.0000	22	0.0003	2	0.0000	0.13	0.08	1.70	12.65***	22.49****	0.28		+	
<P> NP1	26	61	0.0011	137	0.0017	200	0.0022	0.66	1.26	0.52	7.87**	4.54*	22.37****	−		
I THINK	29	41	0.0008	105	0.0013	58	0.0006	0.58	0.48	1.20	9.67**	21.32****	0.80		+	
<SOP> SO	31	8	0.0001	21	0.0003	2	0.0000	0.56	0.08	6.80	2.08	21.14****	7.74**			−
<P> I	33	745	0.0138	1237	0.0155	1186	0.0129	0.89	0.83	1.07	6.70**	20.83****	1.98		+	
THAT HE	40	2	0.0000	19	0.0002	33	0.0004	0.16	1.50	0.10	10.10**	2.08	19.17****	−		
COUPLE OF	41	7	0.0001	8	0.0001	40	0.0004	1.29	4.33	0.30	0.24	18.95****	11.35***			+
<P> AND	43	305	0.0056	605	0.0076	590	0.0064	0.74	0.84	0.88	18.46****	8.48**	3.38		+	
NP1 <P> NP1	44	4	0.0001	22	0.0003	40	0.0004	0.27	1.58	0.17	7.71**	3.04	18.14****	−		
TO NP1	46	24	0.0004	29	0.0004	80	0.0009	1.22	2.39	0.51	0.51	18.07****	9.32**			+
MANAGED TO	48	14	0.0003	2	0.0000	23	0.0003	10.31	9.96	1.04	15.37****	17.84****	0.01		−	
AND THEN	49	32	0.0006	75	0.0009	39	0.0004	0.63	0.45	1.40	5.11*	17.31****	1.92		+	
<P> I WANT	50	8	0.0001	28	0.0004	7	0.0001	0.42	0.22	1.94	5.36*	16.68****	1.64		+	
<P> AND I	51	42	0.0008	109	0.0014	68	0.0007	0.57	0.54	1.05	10.46**	16.37****	0.06		+	
<P> WE	55	47	0.0009	99	0.0012	151	0.0016	0.70	1.32	0.53	4.23*	4.72*	16.04****			+
<EOP> <SOP> I	56	21	0.0004	77	0.0010	77	0.0008	0.40	0.87	0.46	15.97****	0.79	11.11***	−		
<P> HOWEVER	57	14	0.0003	14	0.0002	49	0.0005	1.47	3.03	0.49	1.05	15.90****	6.40*			+
<P> YOU	58	30	0.0006	87	0.0011	110	0.0012	0.51	1.10	0.46	11.26***	0.41	15.87****	−		
IF I	60	49	0.0009	63	0.0008	35	0.0004	1.15	0.48	2.38	0.51	12.63***	15.63****			−
LIKE A	61	41	0.0008	22	0.0003	31	0.0003	2.75	1.22	2.25	15.54****	0.52	11.72***	+		
ALL OF	63	11	0.0002	52	0.0007	30	0.0003	0.31	0.50	0.62	15.44****	9.55**	1.92		+	
FUN <P>	64	5	0.0001	25	0.0003	6	0.0001	0.29	0.21	1.42	7.92**	15.40****	0.33		+	
I NEED TO	68	13	0.0002	28	0.0004	8	0.0001	0.68	0.25	2.76	1.33	14.82***	5.32*			−
<P> NP1 <P>	69	14	0.0003	41	0.0005	66	0.0007	0.50	1.39	0.36	5.43*	2.86	14.67***	−		
<P> PERHAPS	70	7	0.0001	26	0.0003	7	0.0001	0.40	0.23	1.70	5.51*	14.54***	0.98		+	
HOWEVER <P>	72	10	0.0002	13	0.0002	45	0.0005	1.13	3.00	0.38	0.09	14.39***	9.34**			+
<P> HOWEVER <P>	72	10	0.0002	13	0.0002	45	0.0005	1.13	3.00	0.38	0.09	14.39***	9.34**			+

Table C.3: Collocations Significant to $p < 0.001$ for Neuroticism

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
MEAN <P>	73	2	0.0000	24	0.0003	25	0.0003	0.12	0.90	0.14	14.38***	0.13	12.84***	–		
AT WORK	74	10	0.0002	24	0.0003	6	0.0001	0.61	0.22	2.83	1.78	14.30***	4.25*			–
INSTEAD <P>	75	12	0.0002	7	0.0001	2	0.0000	2.53	0.25	10.20	3.98*	3.71	14.21***	+		
FEEL LIKE	76	7	0.0001	22	0.0003	5	0.0001	0.47	0.20	2.38	3.42	14.13***	2.23		+	
WAS THAT	77	1	0.0000	4	0.0001	22	0.0002	0.37	4.77	0.08	0.95	11.27***	14.11***			+
YOU SEE	80	4	0.0001	4	0.0001	25	0.0003	1.47	5.42	0.27	0.30	14.07***	7.80**			+
WHAT A	81	15	0.0003	3	0.0000	9	0.0001	7.37	2.60	2.83	14.06***	2.34	6.37*	+		
I NEED	82	23	0.0004	41	0.0005	17	0.0002	0.83	0.36	2.30	0.54	13.97***	6.88**			–
<P> STILL	84	19	0.0004	7	0.0001	7	0.0001	4.00	0.87	4.62	11.38***	0.07	13.94***	+		
ME <P> <EOP>	86	0	0	11	0.0001	15	0.0002	-	1.18	-	-	0.18	13.87***	–		
<P> I HAD	87	7	0.0001	39	0.0005	37	0.0004	0.26	0.82	0.32	13.85***	0.73	9.57**	–		
THIS YEAR	89	20	0.0004	7	0.0001	8	0.0001	4.21	0.99	4.25	12.57***	0.00	13.64***	+		
GET A	90	31	0.0006	15	0.0002	21	0.0002	3.05	1.21	2.51	13.60***	0.33	10.87***	+		
THE BEST	91	5	0.0001	15	0.0002	37	0.0004	0.49	2.14	0.23	2.10	6.72**	13.50***			+
<P> AND THEN	92	11	0.0002	33	0.0004	12	0.0001	0.49	0.32	1.56	4.63*	13.43***	1.12		+	
IS THAT	93	6	0.0001	24	0.0003	40	0.0004	0.37	1.44	0.26	5.71*	2.08	13.30***	–		
<P> NPD1	94	9	0.0002	23	0.0003	6	0.0001	0.58	0.23	2.55	2.11	13.22***	3.24		+	
<P> I'LL	95	15	0.0003	47	0.0006	22	0.0002	0.47	0.41	1.16	7.25**	13.21***	0.19		+	
GOING TO	96	68	0.0013	152	0.0019	112	0.0012	0.66	0.64	1.03	8.55**	13.17***	0.04		+	
<P> LAST	97	16	0.0003	4	0.0001	20	0.0002	5.89	4.33	1.36	13.11***	9.47**	0.83		–	
<SOP> I	98	24	0.0004	78	0.0010	78	0.0008	0.45	0.87	0.52	12.98***	0.80	8.53**	–		
YESTERDAY <P>	99	23	0.0004	9	0.0001	16	0.0002	3.77	1.54	2.44	12.96***	1.11	7.70**	+		
NEED TO	100	33	0.0006	48	0.0006	23	0.0003	1.01	0.42	2.44	0.00	12.94***	11.01***			–
<P> WHICH	100	83	0.0015	68	0.0009	111	0.0012	1.80	1.41	1.27	12.94***	5.19*	2.70		–	
<P> I THINK	103	23	0.0004	62	0.0008	34	0.0004	0.55	0.48	1.15	6.64**	12.79***	0.27		+	
I MEAN <P>	104	2	0.0000	22	0.0003	23	0.0003	0.13	0.91	0.15	12.65***	0.11	11.31***	–		
AND ON	106	3	0.0001	0	0	10	0.0001	-	-	0.51	5.43*	12.48***	1.17		–	
FIND IT	106	3	0.0001	0	0	10	0.0001	-	-	0.51	5.43*	12.48***	1.17		–	
WHEN I WAS	107	6	0.0001	24	0.0003	7	0.0001	0.37	0.25	1.46	5.71*	12.45***	0.45		+	
TO EAT	107	14	0.0003	9	0.0001	4	0.0000	2.29	0.39	5.95	3.89*	2.76	12.45***	+		
SO <P>	108	16	0.0003	59	0.0007	53	0.0006	0.40	0.78	0.51	12.35***	1.76	6.08*	–		
I GET	110	9	0.0002	42	0.0005	26	0.0003	0.32	0.54	0.59	12.28***	6.44*	2.03		+	
<P> NOW <P>	112	8	0.0001	9	0.0001	34	0.0004	1.31	3.27	0.40	0.31	12.13***	6.44*			+
I HAD	113	41	0.0008	111	0.0014	100	0.0011	0.54	0.78	0.70	12.03***	3.23	3.96*	–		
TO NP1 <P>	116	4	0.0001	8	0.0001	31	0.0003	0.74	3.36	0.22	0.26	11.39***	11.75***			+
I REMEMBER	117	7	0.0001	16	0.0002	3	0.0000	0.64	0.16	3.97	0.99	11.73***	4.47*			–
THIS MORNING	118	23	0.0004	15	0.0002	45	0.0005	2.26	2.60	0.87	6.22*	11.70***	0.30		–	
SEE IT	119	4	0.0001	1	0.0000	14	0.0002	5.89	12.13	0.49	3.28	11.66***	1.83			+
ABOUT THE	121	15	0.0003	55	0.0007	32	0.0003	0.40	0.50	0.80	11.41***	9.90**	0.54		+	
WRITE ABOUT	122	6	0.0001	11	0.0001	1	0.0000	0.80	0.08	10.20	0.19	11.25***	7.11**			–

Table C.4: Collocations Significant to $p < 0.001$ for Neuroticism (cont.)

Table C.5: Collocations Significant to $p < 0.001$ for Neuroticism (cont.)

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
GO HOME	122	2	0.0000	11	0.0001	1	0.0000	0.27	0.08	3.40	3.86*	11.25***	1.08		+	
<P> AFTER	126	19	0.0004	23	0.0003	58	0.0006	1.22	2.18	0.56	0.40	11.03***	5.36*			+
A BIT OF	127	7	0.0001	5	0.0001	24	0.0003	2.06	4.16	0.50	1.56	10.97***	2.99			+
AND BUY	128	6	0.0001	0	0	1	0.0000	-	-	10.20	10.87***	1.25	7.11**	+		
SLOWLY <P>	128	6	0.0001	0	0	1	0.0000	-	-	10.20	10.87***	1.25	7.11**	+		
<P> TO	129	30	0.0006	28	0.0004	66	0.0007	1.58	2.04	0.77	3.01	10.84***	1.41			+

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> AND	1	542	0.0067	605	0.0076	612	0.0132	0.89	1.74	0.51	3.91*	92.93****	129.42****			+
<P> AND I	2	61	0.0008	109	0.0014	101	0.0022	0.56	1.60	0.35	14.07***	11.32***	44.19****			+
<P> BUT	4	361	0.0045	429	0.0054	333	0.0072	0.84	1.34	0.62	6.34*	15.64****	37.77****			+
TO NP1	5	92	0.0011	29	0.0004	41	0.0009	3.15	2.44	1.29	34.04****	13.72***	1.93		–	
OK <P>	10	5	0.0001	39	0.0005	0	0	0.13	-	-	30.07****	-	4.55*			–
NP1 <P>	11	355	0.0044	223	0.0028	146	0.0032	1.58	1.13	1.40	29.51****	1.28	12.29***	+		
<P> NP1	12	224	0.0028	137	0.0017	64	0.0014	1.62	0.81	2.02	20.58****	2.10	27.55****	+		
IN NP1	13	67	0.0008	20	0.0003	24	0.0005	3.33	2.07	1.61	26.48****	5.76*	4.28*	+		
NP1 AND	18	144	0.0018	79	0.0010	35	0.0008	1.81	0.76	2.37	18.78****	1.82	24.59****	+		
<P> BUT I	19	57	0.0007	97	0.0012	77	0.0017	0.58	1.37	0.43	10.79**	4.15*	24.08****	–		
<P> AS	21	131	0.0016	68	0.0009	31	0.0007	1.91	0.79	2.43	19.86****	1.27	23.44****	+		
THAT I	23	126	0.0016	186	0.0023	131	0.0028	0.67	1.21	0.55	12.03***	2.85	22.14****	–		
<P> <P>	25	125	0.0016	61	0.0008	48	0.0010	2.04	1.36	1.50	22.04****	2.46	6.04*	+		
LAST NIGHT	26	58	0.0007	28	0.0004	48	0.0010	2.06	2.95	0.70	10.48**	21.78****	3.39		–	
<P> HOWEVER	27	50	0.0006	14	0.0002	16	0.0003	3.55	1.97	1.80	21.24****	3.42	4.60*	+		
HOWEVER <P>	29	48	0.0006	13	0.0002	14	0.0003	3.67	1.86	1.98	21.12****	2.56	5.62*	+		
<P> HOWEVER <P>	29	48	0.0006	13	0.0002	14	0.0003	3.67	1.86	1.98	21.12****	2.56	5.62*	+		
I THINK	31	50	0.0006	105	0.0013	50	0.0011	0.47	0.82	0.58	20.33****	1.35	7.50**	–		
<P> OK <P>	36	2	0.0000	22	0.0003	0	0	0.09	-	-	19.64****	-	1.82	–		
NP1 AND I	38	51	0.0006	16	0.0002	8	0.0002	3.17	0.86	3.67	18.98****	0.12	15.68****	+		
I WANT	39	18	0.0002	46	0.0006	35	0.0008	0.39	1.31	0.30	12.87***	1.44	18.90****	–		
<P> OK	40	3	0.0000	24	0.0003	0	0	0.12	-	-	18.74****	-	2.73	–		
<P> PERHAPS	42	4	0.0001	26	0.0003	12	0.0003	0.15	0.80	0.19	18.18****	0.44	9.80**	–		
LAST NIGHT <P>	44	22	0.0003	13	0.0002	29	0.0006	1.68	3.84	0.44	2.28	18.04****	8.65**			+
MY FRIENDS	47	3	0.0000	23	0.0003	3	0.0001	0.13	0.22	0.58	17.58****	8.46**	0.45		+	
<P> ON	48	39	0.0005	18	0.0002	34	0.0007	2.15	3.26	0.66	7.78**	17.52****	3.06		–	
MONEY <P>	49	23	0.0003	3	0.0000	3	0.0001	7.61	1.72	4.42	17.31****	0.44	8.37**	+		
A COUPLE OF	50	15	0.0002	4	0.0001	17	0.0004	3.72	7.33	0.51	6.71**	17.28****	3.63		–	
IN NP1 <P>	52	36	0.0004	9	0.0001	12	0.0003	3.97	2.30	1.73	17.16****	3.60	2.93	+		
I AM	53	97	0.0012	146	0.0018	43	0.0009	0.66	0.51	1.30	10.29**	17.07****	2.11		+	
KIND OF	54	33	0.0004	58	0.0007	10	0.0002	0.57	0.30	1.90	7.13**	16.32****	3.52		+	
<P> OR	55	62	0.0008	63	0.0008	72	0.0016	0.98	1.97	0.50	0.02	15.38****	16.31****			+
IS TO	58	4	0.0001	15	0.0002	16	0.0003	0.26	1.84	0.14	6.86**	2.85	15.83****	–		
WENT TO THE	59	19	0.0002	2	0.0000	8	0.0002	9.43	6.89	1.37	15.79****	7.85**	0.58		–	
WAS THAT	60	24	0.0003	4	0.0001	5	0.0001	5.96	2.15	2.77	15.71****	1.31	5.24*	+		
GOING TO	62	92	0.0011	152	0.0019	55	0.0012	0.60	0.62	0.96	15.32****	9.63**	0.05		+	
<P> I AM	63	41	0.0005	53	0.0007	9	0.0002	0.77	0.29	2.62	1.62	15.17****	8.29**			–
AND HE	64	18	0.0002	28	0.0004	2	0.0000	0.64	0.12	5.19	2.26	14.93***	7.40**			–
MYSELF <P>	65	10	0.0001	26	0.0003	23	0.0005	0.38	1.52	0.25	7.48**	2.14	14.91***	–		
COOL <P>	66	10	0.0001	23	0.0003	1	0.0000	0.43	0.07	5.76	5.35*	14.74***	4.41*			–

Table C.6: Collocations Significant to $p < 0.001$ for Extraversion

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> NP1 AND	68	46	0.0006	17	0.0002	7	0.0002	2.69	0.71	3.79	13.67***	0.61	14.58***	+		
YET <P>	69	17	0.0002	6	0.0001	18	0.0004	2.81	5.17	0.54	5.41*	14.57***	3.21		−	
IT THE	72	4	0.0001	1	0.0000	10	0.0002	3.97	17.24	0.23	1.91	14.25***	7.01**			+
OF IT	73	27	0.0003	21	0.0003	34	0.0007	1.28	2.79	0.46	0.71	14.20***	9.24**			+
THIS MORNING <P>	74	27	0.0003	7	0.0001	19	0.0004	3.83	4.68	0.82	12.42***	14.19***	0.44		−	
OF IT <P>	77	13	0.0002	8	0.0001	20	0.0004	1.61	4.31	0.37	1.17	13.90***	7.83**			+
<P> I ALSO	78	22	0.0003	4	0.0001	1	0.0000	5.46	0.43	12.68	13.60***	0.66	13.81***	+		
AND NP1	80	79	0.0010	61	0.0008	19	0.0004	1.29	0.54	2.40	2.20	6.19*	13.74***			−
WENT TO	82	57	0.0007	24	0.0003	20	0.0004	2.36	1.44	1.64	13.62***	1.41	3.92*	+		
<P> IN FACT	83	1	0.0000	14	0.0002	3	0.0001	0.07	0.37	0.19	13.54***	2.98	2.45	−		
WELL <P>	85	106	0.0013	68	0.0009	30	0.0006	1.55	0.76	2.04	8.11**	1.61	13.32***	+		
A COUPLE	86	19	0.0002	6	0.0001	17	0.0004	3.14	4.88	0.64	7.01**	13.15***	1.71		−	
THIS MORNING	87	42	0.0005	15	0.0002	24	0.0005	2.78	2.76	1.01	13.13***	9.85**	0.00		−	
OH WELL	88	21	0.0003	7	0.0001	1	0.0000	2.98	0.25	12.10	7.23**	2.38	12.99***	+		
1 <P>	90	14	0.0002	4	0.0001	0	0	3.48	-	-	5.82*	-	12.74***	+		
I HAVE BEEN	90	14	0.0002	11	0.0001	0	0	1.26	-	-	0.34	-	12.74***			−
ALL OF	91	22	0.0003	52	0.0007	23	0.0005	0.42	0.76	0.55	12.73***	1.21	3.95*	−		
IN FACT	94	2	0.0000	16	0.0002	5	0.0001	0.12	0.54	0.23	12.49***	1.61	3.51	−		
LISTEN TO	94	2	0.0000	16	0.0002	0	0	0.12	-	-	12.49***	-	1.82	−		
<P> I THINK	95	29	0.0004	62	0.0008	27	0.0006	0.46	0.75	0.62	12.47***	1.60	3.17		+	
FOR NP1 <P>	97	9	0.0001	0	0	2	0.0000	-	-	2.59	12.41***	4.01*	1.78		−	
AND I WENT	97	9	0.0001	0	0	1	0.0000	-	-	5.19	12.41***	2.00	3.70	+		
OH WELL <P>	98	20	0.0002	7	0.0001	1	0.0000	2.84	0.25	11.52	6.44*	2.38	12.17***	+		
POINT IN	99	0	0	2	0.0000	6	0.0001	-	5.17	-	-	4.86*	12.08***			+
BUT I	102	83	0.0010	99	0.0012	82	0.0018	0.83	1.43	0.58	1.52	5.59*	11.83***			+
NOT GOING TO	104	4	0.0001	20	0.0003	4	0.0001	0.20	0.34	0.58	11.75***	4.69*	0.60		+	
<P> AND THE	105	25	0.0003	25	0.0003	35	0.0008	0.99	2.41	0.41	0.00	11.51***	11.69***			+
WAS THE	106	38	0.0005	28	0.0004	6	0.0001	1.35	0.37	3.65	1.45	5.96*	11.61***			−
<P> LAST	107	20	0.0002	4	0.0001	8	0.0002	4.97	3.45	1.44	11.53***	4.42*	0.80		−	
AND I'M	108	18	0.0002	44	0.0006	21	0.0005	0.41	0.82	0.49	11.43***	0.55	4.81*	−		
COUPLE OF	109	17	0.0002	8	0.0001	18	0.0004	2.11	3.88	0.54	3.25	11.29***	3.21			+
GET TO	111	9	0.0001	29	0.0004	7	0.0002	0.31	0.42	0.74	11.21***	5.10*	0.35		+	
<P> LAST NIGHT	112	15	0.0002	2	0.0000	7	0.0002	7.45	6.03	1.23	11.16***	6.32*	0.22		−	
FROM WORK	114	1	0.0000	12	0.0001	0	0	0.08	-	-	11.05***	-	0.91	−		
GOT A	116	25	0.0003	17	0.0002	27	0.0006	1.46	2.74	0.53	1.48	10.96***	5.08*			+
NP1 TO	117	29	0.0004	9	0.0001	11	0.0002	3.20	2.11	1.52	10.94***	2.75	1.48	+		
<P> NPD1	118	12	0.0001	23	0.0003	0	0	0.52	-	-	3.59	-	10.92***			−
TO HER	118	12	0.0001	12	0.0001	0	0	0.99	-	-	0.00	-	10.92***			−
<P> TO	119	47	0.0006	28	0.0004	37	0.0008	1.67	2.28	0.73	4.74*	10.91***	1.98		−	
WHO I	119	3	0.0000	17	0.0002	3	0.0001	0.18	0.30	0.58	10.91***	4.66*	0.45		+	

Table C.7: Collocations Significant to $p < 0.001$ for Extraversion (cont.)

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> AND	1	698	0.0084	605	0.0076	361	0.0047	1.11	0.62	1.79	3.70	53.26****	85.38****			–
THAT I	2	127	0.0015	186	0.0023	70	0.0009	0.66	0.39	1.68	13.43***	50.15****	12.74***			–
I AM	7	116	0.0014	146	0.0018	59	0.0008	0.77	0.42	1.82	4.62*	34.84****	14.89***			–
<P> <P>	9	141	0.0017	61	0.0008	133	0.0017	2.23	2.27	0.98	29.73****	30.22****	0.02		–	
IN NP1	10	73	0.0009	20	0.0003	50	0.0007	3.52	2.60	1.35	30.20****	14.47***	2.78		–	
<P> AS	12	134	0.0016	68	0.0009	54	0.0007	1.90	0.83	2.30	19.64****	1.12	29.43****	+		
GOING TO	15	77	0.0009	152	0.0019	122	0.0016	0.49	0.83	0.59	27.82****	2.24	13.91***	–		
AND I	17	196	0.0024	208	0.0026	111	0.0014	0.91	0.55	1.64	0.93	26.35****	17.91****			–
I WILL	19	44	0.0005	55	0.0007	13	0.0002	0.77	0.25	3.14	1.66	26.32****	15.57****			–
SO <P>	20	53	0.0006	59	0.0007	15	0.0002	0.87	0.26	3.28	0.58	26.31****	19.76****			–
<EOP> <SOP> I	21	53	0.0006	77	0.0010	26	0.0003	0.66	0.35	1.89	5.37*	24.47****	7.51**			–
ALL OF	24	23	0.0003	52	0.0007	13	0.0002	0.43	0.26	1.64	12.59***	23.58****	2.12		+	
<SOP> I	28	55	0.0007	78	0.0010	28	0.0004	0.68	0.37	1.82	4.88*	22.66****	7.04**			–
IS NOT	29	29	0.0004	17	0.0002	3	0.0000	1.65	0.18	8.97	2.75	10.29**	22.55****			–
<P> AND I	30	93	0.0011	109	0.0014	48	0.0006	0.82	0.46	1.80	1.92	22.04****	11.44***			–
<P> SO	33	163	0.0020	219	0.0027	240	0.0031	0.72	1.14	0.63	10.40**	1.94	21.13****	–		
OK <P>	34	12	0.0001	39	0.0005	8	0.0001	0.30	0.21	1.39	16.05****	21.09****	0.53		+	
THEY DON'T	36	0	0	10	0.0001	14	0.0002	-	1.46	-	-	0.83	20.48****	–		
<P> SO <P>	37	37	0.0004	35	0.0004	7	0.0001	1.02	0.21	4.90	0.01	19.31****	20.25****			–
<P> HOWEVER	38	50	0.0006	14	0.0002	22	0.0003	3.44	1.63	2.11	20.20****	2.11	9.19**	+		
<P> HOWEVER <P>	39	48	0.0006	13	0.0002	16	0.0002	3.56	1.28	2.78	20.11****	0.44	14.44***	+		
HOWEVER <P>	39	48	0.0006	13	0.0002	17	0.0002	3.56	1.36	2.62	20.11****	0.70	13.17***	+		
IN NP1 <P>	40	40	0.0005	9	0.0001	28	0.0004	4.29	3.23	1.33	20.08****	10.98***	1.33		–	
<P> AFTER	43	51	0.0006	23	0.0003	14	0.0002	2.14	0.63	3.38	9.87**	1.88	19.70****	+		
NPD1 <P>	45	42	0.0005	40	0.0005	87	0.0011	1.01	2.26	0.45	0.00	19.67****	19.58****			+
<P> IT IS	46	34	0.0004	9	0.0001	6	0.0001	3.64	0.69	5.26	14.60***	0.49	19.59****	+		
HAVE TO	47	50	0.0006	87	0.0011	97	0.0013	0.55	1.16	0.48	11.51***	1.00	19.02****	–		
<P> THE	48	376	0.0045	256	0.0032	245	0.0032	1.42	0.99	1.42	18.77****	0.00	18.91****	+		
<P> BUT <P>	49	25	0.0003	10	0.0001	3	0.0000	2.41	0.31	7.73	6.11*	3.71	18.14****	+		
THERE'S NO	51	0	0	6	0.0001	12	0.0002	-	2.08	-	-	2.28	17.55****	–		
BUT <P>	52	27	0.0003	13	0.0002	4	0.0001	2.00	0.32	6.26	4.51*	4.68*	17.46****	+		
NP1 AND I	53	36	0.0004	16	0.0002	8	0.0001	2.17	0.52	4.18	7.19**	2.42	17.24****	+		
<P> I WILL	54	14	0.0002	26	0.0003	4	0.0001	0.52	0.16	3.25	4.10*	17.19****	5.16*			–
FIGURE OUT	56	13	0.0002	12	0.0001	0	0	1.04	-	-	0.01	-	17.07****			–
<P> MORE	57	24	0.0003	8	0.0001	3	0.0000	2.89	0.39	7.42	7.80**	2.17	17.06****	+		
AND I'M	58	20	0.0002	44	0.0006	13	0.0002	0.44	0.31	1.43	10.12**	16.63****	1.02		+	
<P> OK	59	4	0.0000	24	0.0003	4	0.0001	0.16	0.17	0.93	16.59****	15.09***	0.01		+	
MANAGED TO	60	11	0.0001	2	0.0000	19	0.0002	5.30	9.87	0.54	6.54*	16.57****	2.80		–	
THE OFFICE	61	3	0.0000	1	0.0000	16	0.0002	2.89	16.63	0.17	0.98	16.55****	10.77**			+
YES <P>	62	20	0.0002	40	0.0005	51	0.0007	0.48	1.33	0.36	7.54**	1.79	16.42****	–		

Table C.8: Collocations Significant to $p < 0.001$ for Agreeableness

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
AT ME	64	0	0	9	0.0001	11	0.0001	-	1.27	-	-	0.28	16.09****	–		
IN FACT	65	6	0.0001	16	0.0002	27	0.0004	0.36	1.75	0.21	5.09*	3.29	16.07****	–		
FAR TOO	68	4	0.0000	0	0	11	0.0001	-	-	0.34	5.40*	15.68****	3.94*			+
<P> OK <P>	69	4	0.0000	22	0.0003	3	0.0000	0.18	0.14	1.24	14.38***	15.59****	0.08		+	
IT IS	70	73	0.0009	46	0.0006	30	0.0004	1.53	0.68	2.26	5.24*	2.81	15.44****	+		
<P> WE	71	115	0.0014	99	0.0012	58	0.0008	1.12	0.61	1.84	0.69	9.31**	15.11***			–
I WOULDN'T	73	5	0.0001	3	0.0000	20	0.0003	1.61	6.93	0.23	0.44	14.74***	10.80**			+
FIND OUT	75	0	0	6	0.0001	10	0.0001	-	1.73	-	-	1.17	14.63***	–		
THERE ARE	76	48	0.0006	33	0.0004	16	0.0002	1.40	0.50	2.78	2.28	5.39*	14.44***			–
NOT HAVE	76	11	0.0001	4	0.0001	0	0	2.65	-	-	3.15	-	14.44***			–
<P> PERHAPS	77	6	0.0001	26	0.0003	14	0.0002	0.22	0.56	0.40	14.21***	3.21	3.92*	–		
IS TO	79	3	0.0000	15	0.0002	1	0.0000	0.19	0.07	2.78	9.17**	14.17***	0.90		+	
I HAVE	80	145	0.0018	170	0.0021	103	0.0013	0.82	0.63	1.31	3.00	14.13***	4.35*			–
MY FRIENDS	82	8	0.0001	23	0.0003	4	0.0001	0.34	0.18	1.86	8.13**	14.06***	1.08		+	
I THINK	83	74	0.0009	105	0.0013	55	0.0007	0.68	0.54	1.25	6.58*	14.02***	1.57		+	
BIRTHDAY <P>	84	6	0.0001	10	0.0001	25	0.0003	0.58	2.60	0.22	1.16	7.23**	13.98***			+
GOING TO BE	86	9	0.0001	31	0.0004	27	0.0004	0.28	0.91	0.31	13.61***	0.14	10.82**	–		
DIDN'T HAVE	89	2	0.0000	4	0.0001	16	0.0002	0.48	4.16	0.12	0.75	8.18**	13.47***			+
OF ME	90	4	0.0000	21	0.0003	10	0.0001	0.18	0.49	0.37	13.30***	3.58	3.13	–		
HAVE AN	92	10	0.0001	7	0.0001	0	0	1.38	-	-	0.43	-	13.13***			–
<P> THERE	93	83	0.0010	57	0.0007	39	0.0005	1.40	0.71	1.97	3.96*	2.74	13.11***	+		
LISTEN TO	96	2	0.0000	16	0.0002	3	0.0000	0.12	0.19	0.62	12.91***	9.27**	0.28		+	
TRIED TO	97	4	0.0000	13	0.0002	20	0.0003	0.30	1.60	0.19	5.35*	1.78	12.88***	–		
<P> <EOP> <SOP>	99	232	0.0028	305	0.0038	235	0.0031	0.73	0.80	0.92	12.78***	6.60*	0.90		+	
YESTERDAY <P>	100	31	0.0004	9	0.0001	30	0.0004	3.32	3.46	0.96	12.01***	12.75***	0.03		–	
WANTED TO	101	12	0.0001	35	0.0004	28	0.0004	0.33	0.83	0.40	12.60***	0.53	7.84**	–		
MONEY <P>	103	17	0.0002	3	0.0000	2	0.0000	5.46	0.69	7.89	10.32**	0.16	12.46***	+		
THIS WEEKEND	104	8	0.0001	4	0.0001	20	0.0003	1.93	5.20	0.37	1.22	12.27***	6.25*			+
<P> BECAUSE	105	33	0.0004	58	0.0007	25	0.0003	0.55	0.45	1.22	7.89**	12.25***	0.59		+	
PROCESS <P>	106	9	0.0001	0	0	2	0.0000	-	-	4.18	12.15***	2.85	4.31*	+		
RIGHT NOW	107	10	0.0001	31	0.0004	17	0.0002	0.31	0.57	0.55	12.06***	3.62	2.40		+	
<EOP> <SOP> SO	108	5	0.0001	21	0.0003	4	0.0001	0.23	0.20	1.16	11.18***	12.03***	0.05		+	
<SOP> SO	108	6	0.0001	21	0.0003	4	0.0001	0.28	0.20	1.39	9.38**	12.03***	0.27		+	
AND THAT	109	26	0.0003	20	0.0003	6	0.0001	1.25	0.31	4.02	0.58	7.42**	12.02***			–
<P> I JUST	110	15	0.0002	39	0.0005	15	0.0002	0.37	0.40	0.93	11.94***	10.14**	0.04		+	
THIS IS NOT	112	9	0.0001	1	0.0000	0	0	8.68	-	-	7.07**	-	11.82***	+		
I WILL BE	112	9	0.0001	9	0.0001	0	0	0.96	-	-	0.01	-	11.82***			–
HAVE ANY	113	1	0.0000	2	0.0000	12	0.0002	0.48	6.24	0.08	0.38	8.32**	11.81***			+
<P> AND WE	113	19	0.0002	8	0.0001	3	0.0000	2.29	0.39	5.88	4.22*	2.17	11.81***	+		
<P> <EOP>	114	244	0.0029	315	0.0040	246	0.0032	0.75	0.81	0.92	11.80***	6.06*	0.84		+	

Table C.9: Collocations Significant to $p < 0.001$ for Agreeableness (cont.)

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
BANK HOLIDAY	116	0	0	0	0	8	0.0001	-	-	-	-	11.40***	11.70***			+
TURN UP	116	0	0	1	0.0000	8	0.0001	-	8.31	-	-	6.47*	11.70***			+
<P> EITHER	116	0	0	3	0.0000	8	0.0001	-	2.77	-	-	2.56	11.70***	–		
USED TO	117	22	0.0003	14	0.0002	37	0.0005	1.52	2.75	0.55	1.51	11.66***	5.06*			+
LIKE A	118	34	0.0004	22	0.0003	49	0.0006	1.49	2.31	0.64	2.17	11.60***	3.96*			+
NP1 <P>	119	311	0.0038	223	0.0028	243	0.0032	1.34	1.13	1.19	11.55***	1.80	4.05*	+		
OF THE	120	326	0.0039	287	0.0036	226	0.0029	1.10	0.82	1.34	1.27	5.11*	11.50***			–
TOMORROW <P>	121	11	0.0001	16	0.0002	31	0.0004	0.66	2.01	0.33	1.12	5.47*	11.48***			+
SO I	122	54	0.0007	75	0.0009	89	0.0012	0.69	1.23	0.56	4.24*	1.80	11.47***	–		
<P> OTHER	123	5	0.0001	0	0	8	0.0001	-	-	0.58	6.75**	11.40***	0.94		–	
THANK GOD	123	2	0.0000	0	0	8	0.0001	-	-	0.23	2.70	11.40***	4.32*			+
SEE IT	123	5	0.0001	1	0.0000	12	0.0002	4.82	12.47	0.39	2.77	11.40***	3.52			+
<P> AND THE	124	42	0.0005	25	0.0003	15	0.0002	1.62	0.62	2.60	3.77	2.16	11.38***	+		
FROM WORK	125	1	0.0000	12	0.0001	1	0.0000	0.08	0.09	0.93	11.37***	10.55**	0.00		+	
TO TALK TO	125	1	0.0000	12	0.0001	2	0.0000	0.08	0.17	0.46	11.37***	7.54**	0.42		+	
IF I'M	125	1	0.0000	12	0.0001	6	0.0001	0.08	0.52	0.15	11.37***	1.81	4.35*	–		
<P> EVERYONE	125	1	0.0000	12	0.0001	6	0.0001	0.08	0.52	0.15	11.37***	1.81	4.35*	–		
<P> YES	126	13	0.0002	20	0.0003	34	0.0004	0.63	1.77	0.35	1.76	4.23*	11.36***			+
WHO I	127	3	0.0000	17	0.0002	12	0.0002	0.17	0.73	0.23	11.33***	0.68	6.48*	–		
OFF <P>	129	21	0.0003	15	0.0002	38	0.0005	1.35	2.63	0.51	0.80	11.23***	6.32*			+
TO BED	131	4	0.0000	19	0.0002	11	0.0001	0.20	0.60	0.34	11.18***	1.86	3.94*	–		
TO TALK	131	5	0.0001	21	0.0003	10	0.0001	0.23	0.49	0.46	11.18***	3.58	2.09		+	
LISTENING TO	131	4	0.0000	19	0.0002	10	0.0001	0.20	0.55	0.37	11.18***	2.50	3.13	–		
DID <P>	132	25	0.0003	11	0.0001	6	0.0001	2.19	0.57	3.87	5.09*	1.31	11.13***	+		
I HAD	134	98	0.0012	111	0.0014	64	0.0008	0.85	0.60	1.42	1.35	11.03***	4.87*			–
PEOPLE WHO	135	12	0.0001	33	0.0004	12	0.0002	0.35	0.38	0.93	10.97***	9.40**	0.03		+	

Table C.10: Collocations Significant to $p < 0.001$ for Agreeableness (cont.)

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> <P> <P>	1	123	0.0013	16	0.0002	25	0.0003	6.45	1.42	4.55	75.66****	1.20	63.37****	+		
<P> <P>	2	215	0.0023	61	0.0008	126	0.0014	2.96	1.87	1.58	66.07****	17.11****	17.12****	+		
<P> AND	3	624	0.0066	605	0.0076	846	0.0096	0.86	1.27	0.68	6.47*	19.96****	53.05****			+
OK <P>	7	14	0.0001	39	0.0005	3	0.0000	0.30	0.07	4.32	17.09****	40.27****	6.90**			−
I WILL	8	49	0.0005	55	0.0007	11	0.0001	0.75	0.18	4.12	2.21	36.52****	23.16****			−
<P> I	10	1188	0.0125	1237	0.0155	1387	0.0158	0.81	1.02	0.79	28.40****	0.16	34.66****	−		
<P> OK	16	5	0.0001	24	0.0003	1	0.0000	0.17	0.04	4.63	17.11****	28.59****	2.61		+	
GOING TO	18	142	0.0015	152	0.0019	84	0.0010	0.78	0.50	1.56	4.38*	27.16****	10.90***			−
THAT I	20	123	0.0013	186	0.0023	173	0.0020	0.55	0.84	0.66	26.42****	2.63	12.80***	−		
<P> OK <P>	21	4	0.0000	22	0.0003	1	0.0000	0.15	0.04	3.70	17.09****	25.79****	1.70		+	
I AM	25	95	0.0010	146	0.0018	84	0.0010	0.55	0.52	1.05	21.72****	23.60****	0.09		+	
<P> <EOP> <SOP>	27	308	0.0032	305	0.0038	221	0.0025	0.85	0.66	1.29	4.23*	23.04****	8.43**			−
<P> <EOP>	28	321	0.0034	315	0.0040	231	0.0026	0.85	0.66	1.29	3.93*	22.59****	8.60**			−
<P> AND I	29	78	0.0008	109	0.0014	139	0.0016	0.60	1.16	0.52	12.07***	1.28	22.43****	−		
<P> HOWEVER	32	17	0.0002	14	0.0002	52	0.0006	1.02	3.37	0.30	0.00	19.69****	21.42****			+
IN NP1	33	70	0.0007	20	0.0003	62	0.0007	2.93	2.81	1.04	21.31****	18.62****	0.06		−	
<P> ONE OF	34	2	0.0000	15	0.0002	22	0.0003	0.11	1.33	0.08	13.67***	0.73	21.09****	−		
<P> I HATE	36	6	0.0001	11	0.0001	31	0.0004	0.46	2.55	0.18	2.50	8.05**	20.48****			+
TO TALK	37	6	0.0001	21	0.0003	2	0.0000	0.24	0.09	2.78	11.68***	20.23****	1.80		+	
<EOP> <SOP>	38	337	0.0035	335	0.0042	256	0.0029	0.84	0.69	1.22	4.86*	19.83****	5.72*			−
<P> HOWEVER <P>	40	16	0.0002	13	0.0002	48	0.0005	1.03	3.35	0.31	0.01	18.06****	19.32****			+
KIND OF	41	65	0.0007	58	0.0007	23	0.0003	0.94	0.36	2.62	0.12	19.28****	17.77****			−
THE GAME	42	0	0	12	0.0001	13	0.0001	-	0.98	-	-	0.00	19.05****	−		
THE WAY <P>	43	19	0.0002	7	0.0001	1	0.0000	2.28	0.13	17.58	3.84*	5.67*	18.42****			−
HOWEVER <P>	45	18	0.0002	13	0.0002	48	0.0005	1.16	3.35	0.35	0.17	18.06****	16.57****			+
<P> LAST	46	9	0.0001	4	0.0001	28	0.0003	1.89	6.34	0.30	1.19	17.96****	11.76***			+
SHE WAS	47	17	0.0002	26	0.0003	48	0.0005	0.55	1.67	0.33	3.82	4.65*	17.90****			+
<P> TO	48	38	0.0004	28	0.0004	75	0.0009	1.14	2.43	0.47	0.27	17.87****	15.38****			+
FRIENDS <P>	49	2	0.0000	18	0.0002	8	0.0001	0.09	0.40	0.23	17.70****	5.00*	4.33*	−		
ANYMORE <P>	52	9	0.0001	16	0.0002	1	0.0000	0.47	0.06	8.33	3.41	17.48****	6.76**			−
LAST NIGHT	53	44	0.0005	28	0.0004	74	0.0008	1.32	2.39	0.55	1.33	17.22****	10.21**			+
<P> AS	54	93	0.0010	68	0.0009	136	0.0015	1.15	1.81	0.63	0.74	16.89****	11.79***			+
I WAS	55	190	0.0020	237	0.0030	258	0.0029	0.67	0.99	0.68	16.77****	0.02	16.30****	−		
THAT MY	56	6	0.0001	17	0.0002	27	0.0003	0.30	1.44	0.21	7.60**	1.41	16.13****	−		
ANYWAY <P>	58	38	0.0004	34	0.0004	76	0.0009	0.94	2.03	0.46	0.08	12.57***	16.03****			+
<P> I WILL	60	20	0.0002	26	0.0003	6	0.0001	0.65	0.21	3.08	2.20	15.53****	6.91**			−
<P> THE	61	371	0.0039	256	0.0032	251	0.0029	1.22	0.89	1.37	5.81*	1.78	14.94***	+		
<P> ACTUALLY	62	13	0.0001	4	0.0001	25	0.0003	2.73	5.66	0.48	3.56	14.93***	4.84*			+
BY THE WAY	63	16	0.0002	5	0.0001	1	0.0000	2.68	0.18	14.81	4.28*	3.32	14.82***	+		
A FEW WEEKS	64	0	0	2	0.0000	10	0.0001	-	4.53	-	-	5.06*	14.65***			+

Table C.11: Collocations Significant to $p < 0.001$ for Conscientiousness

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
<P> I SHOULD	66	10	0.0001	5	0.0001	27	0.0003	1.68	4.89	0.34	0.93	14.53***	9.48**			+
<P> DAMN	68	1	0.0000	13	0.0002	2	0.0000	0.06	0.14	0.46	14.43***	10.14**	0.42		+	
MY FRIENDS	69	6	0.0001	23	0.0003	5	0.0001	0.22	0.20	1.11	13.85***	14.38***	0.03		+	
<P> DESPITE	70	4	0.0000	3	0.0000	21	0.0002	1.12	6.34	0.18	0.02	13.47***	14.03***			+
THAT <P> I	71	21	0.0002	18	0.0002	3	0.0000	0.98	0.15	6.48	0.00	13.42***	13.83***			−
HOW I	72	14	0.0001	16	0.0002	2	0.0000	0.73	0.11	6.48	0.72	13.82***	9.22**			−
ARE GOING	72	5	0.0001	16	0.0002	2	0.0000	0.26	0.11	2.31	8.16**	13.82***	1.11		+	
LISTEN TO	72	8	0.0001	16	0.0002	2	0.0000	0.42	0.11	3.70	4.31*	13.82***	3.41		+	
<P> I HOPE	73	9	0.0001	13	0.0002	1	0.0000	0.58	0.07	8.33	1.61	13.42***	6.76**			−
I GET	74	19	0.0002	42	0.0005	23	0.0003	0.38	0.50	0.76	13.41***	7.67**	0.75		+	
<P> THEN <P>	75	11	0.0001	0	0	9	0.0001	-	-	1.13	13.40***	11.61***	0.08		−	
NOT GOING TO	76	11	0.0001	20	0.0003	4	0.0000	0.46	0.18	2.55	4.48*	13.28***	2.88		+	
DO IS	77	0	0	6	0.0001	9	0.0001	-	1.36	-	-	0.34	13.19***	−		
THING TO	78	1	0.0000	3	0.0000	13	0.0001	0.28	3.93	0.07	1.43	5.79*	13.15***			+
LONG AS	78	1	0.0000	4	0.0001	13	0.0001	0.21	2.94	0.07	2.49	4.17*	13.15***	−		
IN NP1 <P>	79	35	0.0004	9	0.0001	34	0.0004	3.26	3.42	0.95	12.17***	13.13***	0.04		−	
NEXT WEEK	80	10	0.0001	3	0.0000	0	0	2.80	-	-	2.84	-	13.10***			−
THE WAY	81	58	0.0006	37	0.0005	23	0.0003	1.31	0.56	2.33	1.72	4.82*	13.04***			−
THEM <P> I	82	1	0.0000	12	0.0001	8	0.0001	0.07	0.60	0.12	13.01***	1.25	6.75**	−		
<P> SHE	83	60	0.0006	55	0.0007	99	0.0011	0.91	1.63	0.56	0.23	8.78**	12.92***			+
SORT OF	85	21	0.0002	15	0.0002	45	0.0005	1.17	2.72	0.43	0.23	12.88***	10.89***			+
THIS MORNING	85	37	0.0004	15	0.0002	45	0.0005	2.07	2.72	0.76	6.14*	12.88***	1.52		−	
LIKE A	86	36	0.0004	22	0.0003	57	0.0006	1.37	2.35	0.58	1.40	12.80***	6.55*			+
OKAY <P>	87	9	0.0001	9	0.0001	29	0.0003	0.84	2.92	0.29	0.14	9.19**	12.68***			+
ARE GOING TO	88	4	0.0000	15	0.0002	2	0.0000	0.22	0.12	1.85	8.87**	12.58***	0.53		+	
ANYTHING <P>	89	24	0.0003	4	0.0001	15	0.0002	5.03	3.40	1.48	12.54***	5.74*	1.46		−	
<P> I WAS	90	53	0.0006	82	0.0010	84	0.0010	0.54	0.93	0.58	12.43***	0.23	9.68**	−		
THIS MORNING <P>	91	23	0.0002	7	0.0001	29	0.0003	2.76	3.75	0.73	6.41*	12.36***	1.24		−	
<P> I WANT	92	10	0.0001	28	0.0004	22	0.0003	0.30	0.71	0.42	12.34***	1.44	5.59*	−		
AND I	93	173	0.0018	208	0.0026	191	0.0022	0.70	0.83	0.84	12.33***	3.37	2.83		+	
AND THEN	95	62	0.0007	75	0.0009	43	0.0005	0.69	0.52	1.33	4.59*	12.23***	2.14		+	
MANAGED TO	97	14	0.0001	2	0.0000	17	0.0002	5.87	7.70	0.76	8.13**	12.12***	0.57		−	
HIM <P> I	98	7	0.0001	12	0.0001	1	0.0000	0.49	0.08	6.48	2.36	12.09***	4.61*			−
THAT SHE	100	5	0.0001	12	0.0001	21	0.0002	0.35	1.59	0.22	4.33*	1.68	11.86***			+
IN A	103	71	0.0007	77	0.0010	110	0.0013	0.77	1.29	0.60	2.45	3.05	11.76***			+
I WOULD	104	39	0.0004	64	0.0008	35	0.0004	0.51	0.50	1.03	11.33***	11.72***	0.02		+	
<P> IN	105	83	0.0009	58	0.0007	110	0.0013	1.20	1.72	0.70	1.15	11.64***	6.17*			+
TO MAKE	106	27	0.0003	50	0.0006	40	0.0005	0.45	0.72	0.62	11.62***	2.32	3.65		+	
WHAT I	108	43	0.0005	45	0.0006	21	0.0002	0.80	0.42	1.90	1.08	11.46***	6.11*			−
<P> ACTUALLY <P>	109	9	0.0001	3	0.0000	19	0.0002	2.52	5.74	0.44	2.18	11.45***	4.47*			+

Table C.12: Collocations Significant to $p < 0.001$ for Conscientiousness (cont.)

Table C.13: Collocations Significant to $p < 0.001$ for Conscientiousness (cont.)

Feature	Rank	High Freq.	High R.Freq.	Mid Freq.	Mid R.Freq.	Low Freq.	Low R.Freq.	High-Mid R.F. Ratio	Low-Mid R.F. Ratio	High-Low R.F. Ratio	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
CASE <P>	111	5	0.0001	2	0.0000	16	0.0002	2.10	7.25	0.29	0.85	11.06***	6.94**			+
<P> ETC	111	17	0.0002	2	0.0000	8	0.0001	7.13	3.62	1.97	11.06***	3.29	2.65	+		
THE WEEKEND <P>	112	7	0.0001	1	0.0000	13	0.0001	5.87	11.78	0.50	4.07*	11.05***	2.32		−	
<P> OUR	113	3	0.0000	15	0.0002	3	0.0000	0.17	0.18	0.93	10.99***	9.96**	0.01		+	
IS TO	113	3	0.0000	15	0.0002	15	0.0002	0.17	0.91	0.19	10.99***	0.07	9.69**	−		
<P> OTHER	114	9	0.0001	0	0	6	0.0001	-	-	1.39	10.96***	7.74**	0.39		−	
FUN <P>	115	19	0.0002	25	0.0003	8	0.0001	0.64	0.29	2.20	2.22	10.95***	3.80		+	
<P> ANYWAY <P>	116	29	0.0003	27	0.0003	56	0.0006	0.90	1.88	0.48	0.15	7.69**	10.94***			+
A BIT	117	83	0.0009	37	0.0005	66	0.0008	1.88	1.62	1.16	10.92***	5.66*	0.85		−	
<EOP> <SOP> I	118	51	0.0005	77	0.0010	54	0.0006	0.56	0.64	0.87	10.89***	6.65**	0.48		+	

Appendix D

Publications

- Nowson, S., Oberlander, J., & Gill, A. (2005). Weblogs, genres and individual differences. *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1666–1671). Hillsdale, NJ: LEA.

k

k

k

k

k

k

Bibliography

Adenekan, S. (2005). *Academics give lessons on blogs.* BBC News Online. Available at http://news.bbc.co.uk/1/hi/education/4194669.stm

Anderson, K. (2005) *American media vs the blogs.* BBC News Online. Available at http://news.bbc.co.uk/1/hi/world/americas/4279229.stm

Anderson, K.H. (2004). *Student's use of weblogs: weblogs for collaboration in an educational setting.* Masters Dissertation, Department of Information Science, University of Bergen.

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. (2003). Gender, genre, and writing style in formal written texts. *Text, 23* (3).

Askehave, I., & Swales, J.M. (2001). Genre identification and communicative purpose: a problem and a possible solution. *Applied Linguistics, 22* (2), 195-212.

Ball, C. (1994). Automated text analysis: Cautionary tales. *Literacy and Linguistic Computing, 9* (4), 295-302.

Bälter, O. (1998). *Electronic Mail in a Working Context*. PhD Thesis, Royal Institute of Technology, Stockholm.

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.

Baron, N. (2001). Commas and Canaries: the role of punctuation in speech and writing. *Language Sciences, 23* (1), 15-67

BBC Online (2004) *'blog' picked as word of the year.* BBC News Online. Available at http://news.bbc.co.uk/1/hi/technology/4059291.stm

BBC Online (2005) *Blogs respond to London blasts.* BBC News Online. Available at http://news.bbc.co.uk/1/hi/technology/4659679.stm

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

Biber, D. (2004). *Towards a typology of web registers: a multi-dimensional analysis*. Invited lecture, Conference of Corpus Linguistics: Perspectives for the future. Heidelberg University.

Boyd, C. (2004). *Web logs aid disaster recovery*. BBC News Online. Available at http://news.bbc.co.uk/1/hi/technology/4135687.stm

Buchanan, T. (2001). *Online implementation of an IPIP Five Factor Personality Inventory [On-line]*. Available at http://users.wmin.ac.uk/∼buchant/wwwffi/introduction.html

Buchanan, T., Johnson, J.A., & Goldberg, L.R., (2005). Implementing a five-factor personality inventory for use on the internet. *European Journal of Psychological Assessment, 21* (2), 115-127.

Buchanan, T., & Reips, U.-D. (2001). Technological biases in online research: Personality and demographic correlates of Macintosh and Javascript use. Poster presented at *Psychology and the Internet: A European Perspective*. Farnborough, UK.

Burger, J.D., & Henderson, J.C. (2006). An Exploration of Observable Features Related to Blogger Age. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Burt, C.D.B. (1994). Prospective and retrospective account-making in diary entries: A model of anxiety reduction and avoidance. *Anxiety, Stress & Coping: An International Journal, 6* (4), 327-340.

Campbell, A., & Rushton, J. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology*, 17, 31-36.

Campbell, D. (2005) *Judiciary will need to know bloggers' rights.* The Scotsman online. Available at http://news.scotsman.com/topics.cfm?tid=956&id=60872005

Carment, D.W., Miles, C.G., & Cervin, V.B. (1965). Persuasiveness and persuasibility as related to Intelligence and Extraversion. *British Journal of Social and Clinical Psychology*, 4, 1-7.

Cho, N. (1996). Linguistic Features of Electronic Mail: Results from a pilot study. Paper presented at the Australia and New Zealand Communication Association Annual Conference, Brisbane, July 1996.

Cloninger, S.C. (1996). *Personality: Description, Dynamics and Development*. W. H. Freeman and Company.

Cohn, M.A., Mehl, M.R., & Pennebaker, J.W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, 15, 687-693.

Colley, A., & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*. 21, 380-392

Collot, M. and Belmore, N. (1996). Electronic language: A new variety of English. In S. Herring, (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*. Amsterdam: Benjamins.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, 497-505.

Costa, P., & McCrae, R. R. (1992). *Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Costa, P.T., McCrae, R.R., & Dye, D.A. (1991). Facet scales for Agreeableness and Conscientiousness: A revision of the NEO Personality Inventory. *Personality and Individual Differences, 12*, 887-898.

Crawford, K. (2005). *Have a blog, lose your job?* CNN Money online. Available at `http://money.cnn.com/2005/02/14/news/economy/blogging/index.htm?cnn=yes`

Crowston, K., & Williams, M. (2000). Reproduced and emergent genres of communication on the World Wide Web. *The information Society*, *16* (3), 201-216.

Crystal, D., (2001). *Language and the Internet.* Cambridge, Cambridge University Press.

Cubranic, D., Holmes, R., Ying, A.T.T., & Murphy, G.C. (2003) *Tools for lightweight knowledge-sharing in open-source software development.* Workshop on Open-Source Software, held as part of the International Conference on Software Engineering 2003.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management, 29* (4), 433-448.

De Raad, B., & Perugini, M. (2002) *Big Five Assessment*. Seattle, WA: Hogrefe and Huber.

Dewaele, J.-M. (1995). Variation dans la longueur moyenne d'énoncés dans l'interlangue française [variation in the mean length of utterances in french inter-language]. In L. Beheydt, (Ed), *linguistique appliquée dans les années 90 [Special Issue]*, volume 16 of *ALBA Papers*, 43-58. ALBA.

Dewaele, J.-M. (1996a). How to measure formality of speech? A model of synchronic variation. In K. Sajavaara and C. Fairweather, (Eds), *Approaches to second language acquisition [Special Issue]*, volume 17 of Jyväskylä Cross Language Studies, 119-133. Jyväskylä University.

Dewaele, J.-M. (1998). Speech rate variation in 2 oral styles of advanced French interlanguage. In V. Regan, (Ed), *Contemporary approaches to second language acquisition in social context: Cross-linguistic perspectives*, 113-123. University College Academic Press, Dublin.

Dewaele, J.-M. (2002). Individual difference in L2 fluency: the the effect of neurobiological correlates. In V. Cook (Ed.), *Portraits of the L2 user*, 219-250. Clevedon: Multilingual Matters.

Dewaele, J.-M. (2005). Investigating the Psychological and Emotional Dimensions in Instructed Language Learning: Obstacles and Possibilities, *The Modern Language Journal*, 89 (iii), 367-380.

Dewaele, J.-M., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, *49* (3), 509-544.

Dewaele, J.-M., & Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Difference, 28*, 355-365.

Digman, J., (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41* (1), 417-440.

Dunning, T.E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19* (1), 61-74.

Eckert, P. (1997). Gender and sociolinguistic variation. In J. Coates (Ed.), *Language and Gender: a reader*. Oxford: Blackwell, 64-75.

Efimova, L., & de Moor, A. (2005). Beyond personal webpublishing: An exploratory study of conversational blogging practises. *Proceedings of the 37th Annual HICSS Conference.* Big Island, Hawaii.

Efimova, L., & Fiedler, S. (2004). Learning webs: learning in weblog networks. in P. Kommers, P. Isaias, and M.b.Nunes (Eds.), *Proceedings of the IADIS International Conference Web Based Communities 2004* (pp. 490–494). Lisbon, Portugal, IADIS Press.

Efimova, L., Fiedler, S., Verwijs, C., & Boyd, A. (2004). *Legitimised theft: distributed apprenticeship in weblog networks*.

Elgersma, E., & de Rijkem M. (2006). Learning to Recognize Blogs: A Preliminary Exploration. *EACL 2006 Workshop: NEW TEXT - Wikis and blogs and other dynamic text sources*.

Eysenck, H. (1970). *The biological Basis of Personality*. Springfield, IL: Thomas.

Eysenck, H. (1993). From DNA to social behaviour: conditions for a paradigm of personality research. In J. Hettema and I. Deary (Eds.), *Foundations of personality*. Kluwer, Dortrect.

Eysenck, H., & Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire - Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H. & Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6, 21-29.

Feuer, J. (1992). Genre study and television. In R.C. Allen (Ed.), *Channels of Discourse, Reassembled: Television and Contemporary Criticism*. London: Routledge, pp. 138-59.

Finn, A., & Kushmerick, N. (2005). Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to StyleAnalysis and Synthesis*.

Funder, D. (2001). Personality. *Annual Review of Psychology, 52* (1), 197-221.

Furnham, A. (1990). Language and personality. In H. Giles & W. Robinson (Eds.), *Handbook of Language and Social Psychology*. Wiley, Chichester.

Gifford, R., & Hine, D.W. (1994). The role of verbal behaviour in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality*, 28, 115-132.

Gill, A.J. (2004). *Personality and Language: The projection and perception of personality in computer-mediated communication*. Unpublished Doctoral Thesis, University of Edinburgh.

Gill, A., Harrison, A. & Oberlander, J. (2004). Interpersonality: Individual differences and interpersonal priming. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, (pp. 464–469). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gill, A., & Oberlander, J. (2002). Taking care of the linguistic features of Extraversion. *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 363–368). Hillsdale, NJ: LEA.

Gill, A., & Oberlander, J. (2003). Perception of e-mail personality at zero acquaintance: Extraversion takes care of itself; Neuroticism is a worry. *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 456–461). Hillsdale, NJ: LEA.

Glance, N.S., Hurst, M., & Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. in *Proceedings of WWW 2004*. New York, US.

Goldberg, L.R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.) *Review of personality and social psychology*, 2, 141-165. Beverley Hills, CA: Sage.

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48* (1), 26-34.

Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P., (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*, 93-104.

Groom, C.J., & Pennebaker, J.W. (2005). The language of love: sex, sexual orientation, and language use in online personal advertisements. *Sex Roles: A Journal of Research, 52* (7-8), 447-461.

Gruber, H. (2000). Scholarly Email Discussion List Postings: a single new genre of academic communication? In L. Pemberton & S. Shurville (Eds) *Words on the Web*, 36-43. Exeter: Intellect Books.

Henning, J. (2003). *The Blogging Iceberg*. Perseus Development Corp. Available at `http://www.perseus.com/blogsurvey/`

Herman, D., Jahn, M., & Ryan, M.-L. (Eds.) (2005). *The Routledge Encyclopedia of Narrative Theory*. London, Routledge.

Herring, S. (2000). Gender difference in CMC: findings and implications. *CPSR Newsletter*, *18* (1), Winter 2000.

Herring, S.C., Kouper, I., Scheidt, L.A., &Wright, E. (2004b). Women and children last: The discursive construction of weblogs. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff and J. Reyman (Eds), *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*, at `http://blog.lib.umn.edu/blogosphere/`, University of Minnesota.

Herring, S.C., Scheidt, L.A., Bonus, S., & Wright, E. (2004a). Bridging the gap: A genre analysis of weblogs. *Proceedings of the 37th Annual HICSS Conference.* Big Island, Hawaii.

Herring, S.C., Scheidt, L.A., Bonus, S., & Wright, E. (2005). Weblogs as a bridging genre. *Information, Technology & People*, *18* (2), 142-171.

Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: an empirical measure. *Foundations of Science*, *7*, 293–340.

Hsu, W.H., Weninger, T., Pydimarri, T., & Paradesi, M.S.R. (2006). Collaborative and Structural Recommendation of Friends using Weblog-based Social Network Analysis. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Huffaker, D., & Calvert S.L. (2005). Gender, Identity and Language Use in Teenage Blogs. *Journal of Computer-Mediated Communication*, 10(2).

Huffaker, D. (2004). *Gender similarities and differences in online identity and language use among teenage bloggers*. Masters Dissertation, Graduate School of Arts and Sciences, Georgetown University.

Jesdanun, A. (2005). *Blog-linked firings prompt calls for better policies.* Associated Press, CNN.com. Available at `http://www.cnn.com/2005/TECH/internet/03/06/firedforblogging.ap/index.html`

Jonassen, D.H., & Grabowski, B.L. (1993). *Handbook of Individual Differences Learning & Instruction*. Lawrence Erlbaum Associates.

Juola, P. (2003) The Time Course of Language Cgange. *Computers and Humanities, 37* (1).

Karlgren, J., & Cutting, D. (1994). Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of Coling 94*, Kyoto.

Keller, F., Lapata, M., & Ourioupina, O. (2002). Using the web to overcome data sparseness. In Jan Hajic and Yuji Matsumoto, (Eds.), *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 230–237). Philadelphia.

Kessler, B., Nunberg, G., & Schütze, H. (1997). Automatic detection of text genre. In P.R. Cohen and W. Wahlster, (Eds), *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguists*, 32-38. Somerset, NJ: Association or Computation Linguistics.

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6* (1), 231-245.

Kline, P. (1993). *The Handbook of Psychological Testing*. Routledge, London.

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

Krishnamurthy, S., (2002). The multidimensionality of blog conversations: The virtual enactment of September 11. Paper presented at *Internet Research 3.0*, Maastricht, The Netherlands.

Labov, J. (1990). The interaction of sex and social class in the course of linguistic change. *Language Variation and Change*, *2*, 205-254.

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *Vol.5(3)*, 37-72.

Levelt, W.J.M. (1989). *Speaking. From Intention to Article*. MIT Press, Cambridge, Mass..

Li, D. (2005) *Why Do You Blog: A Uses-and-Gratifications Inquiry into Bloggers' Motivations*. Unpublished Masters Thesis, Marquette University.

Lloyd, L., Kaulgud, P., & Skiena, S. (2006). Newspapers vs. Blogs: Who Gets the Scoop?. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Louwese, M., McCarthy, P.M., McNamara, D.S., & Graesser, A.C. (2004). Variation in language and cohesion across written and spoken registers. *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1035–1040). Hillsdale, NJ: LEA.

Manning, C.D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Markey, P. and Wells, S. (2002). Interpersonal perception in internet chat rooms. *Journal of Research in Personality, 36*, 134–146.

Marlow, C., (2004). Audience, structure and authority in the weblog community. Presented at *The International Communications Association Conference*, New Orleans.

Mateas, M. (1997). An Oz-Centric Review of Interactive Drama and Believable Agents. *Technical Report CMU-CS-97-156*, Carnegie Mellon University.

Matthews, G., Deary I., & Whiteman, C. (2003). *Personality traits: Second Edition*. Cambridge: Cambridge University Press.

McCrae, R., Costa, P.T., Ostendorf, F., Angleitner, A., Hrebickova, M., Avia, M.D., Sanz, J., Sanchez-Bernardos, M.L., Kusdil, M.E., Woodfield, R., Saunders, P.R., & Smith, P.B. (2000) Nature over Nurture: temperament, personality and life span development. *Journal of Personal Social Psychology*, *78* (1), 173-186

McCrae, R., & Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, *52* (1), 81-90.

McCrae, R., & Costa, P. (1997). Personality trait structure as a human universal. *American Psychologist*, 52, 509-516.

McCroskey, J., & Richmond, V. (1990). Willingness to communicate: A cognitive view. *Journal of Social Behaviour and Personality*, 5, 19-37.

Mihalcea, R., & Liu, H. (2006). A Corpus-based Approach to Finding Happiness. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Miller, C.R. (1984). Genre as social action. *Quarterly Journal of Speech, 70*, 151-67.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger, (Ed), *Learner English on Computer, Studies in Language and Linguistics*, pages 186-198. Addison Wesley Longman, New York.

Mischel, W. (1968). *Personality and Assessment*. New York: Wiley.

Mishne, G. (2005). Experiments with Mood Classification in Blog Posts. In *Style2005 - the 1st Workshop on Stylistic Analysis Of Text For Information Access*, at SIGIR 2005.

Mortensen, T.E. (2004). Personal Publication and Public attention. In L. Gurak, S. Antonijevic, L. Johnson, C. Ratliff and J. Reyman (Eds), *Into the Blogosphere; Rhetoric, Community and Culture of Weblogs*, at `http://blog.lib.umn.edu/blogosphere/`, University of Minnesota.

Mortensen, T.E. (2004). *Dialogue in slow motion: The pleasure of reading and writing across the web.* Keynote, Blogtalk 2.0, Vienna.

Mortensen, T., & Walker J. (2002) Blogging thoughts: personal publication as an online research tool. In A. Morrison (Ed.) *Researching ICTs in Context*, InterMedia Report, 3/2002, Oslo.

Mulac A., Bradac, J.J., & Gibbons, P. (2001). Empirical support for gender-as-culture hypothesis: an intercultural analysis of male/female language differences. *Human Communication Research*, 27, 121-152.

Naiman, N., Frölich, M., & Stern, H.H. (1975). *The good language learner: A report.* Toronto, Canada: Ontario Institute for Studies in Education.

Naiman, N., Frölich, M., Stern, H.H., & Todesco, A. (1978). *The good language learner*. Toronto, Canada: Ontario Institute for Studies in Education.

Nardi, B.A., Schiano, D.J., & Gumbrecht, M. (2004). Blogging as social activity, or, Would you let 900 million people read your diary? in *Proceedings of Conference of the ACM 2004*.

Nass, C., & Lee, K., M., 2000. Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In *Proceedings of CHI 2000*.

Nichols, P.C. (1998). Black women in the rural south: conservative and innovative. In J. Coates (Ed.), *Language and Gender: a reader*. Oxford: Blackwell, pp. 55-63.

Nilsson, S. (2003). *The function of language to facilitate and maintain social networks in research weblogs*. Dissertation, Umea Universitet, Engelska lingvistik.

Nowson, S., Oberlander, J., & Gill, A. (2005). Weblogs, genres and individual differences. *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 1666–1671). Hillsdale, NJ: LEA.

Oberlander, J., & Gill, A. (2004). Individual difference and implicit language: personality, parts-of-speech and pervasiveness. *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 1035–1040). Hillsdale, NJ: LEA.

Oberlander J., & Gill, A.J. (2005). Language with character: A stratified corpus comparison of individual difference in e-mail communication. *In submission*.

Orlowski, A, (2003a). *Google to fix blog noise problem.* The Register. Available at `http://www.theregister.co.uk/2003/05/09/google_to_fix_blog_noise/`

Orlowski, A, (2003b). *Webloggers deal Harvard blog-bores a black eye.* The Register. Available at http://www.theregister.co.uk/2003/08/13/webloggers_deal_harvard_blogbores/

Orlowski, A, (2003c). *Most bloggers "are teenage girls" - A survey.* The Register. Available at http://www.theregister.co.uk/content/6/30954.html

Pateli, N. (2002). Richness, power cues and email text. *Information & Management, 40* (2), 75-86.

Pedersen, T. (1996). Fishing for exactness. In *Proceedings of the South Central SAS User's Group (SCSUG-96) Conference*, pages 188-200, Austin, TX.

Pedersen, T., Kayaalp, M., & Bruce, R. (1996). Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 455-460, Portland, OR.

Peng, C.-Y.J., Lee, K L., & Ingersoll, G M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96* (1), 3-14.

Pennebaker, J.W., & Francis, M.E. (1999). *Linguistic Inquiry and Word Count: LIWC*. Mahwah, NJ: Erlbaum.

Pennebaker, J.W., Francis, M.E., & Booth, R.J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah, NJ: Erlbaum Publishers.

Pennebaker, J.W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*, 1296–1312.

Pennebaker, J.W., & Lay, T.C., (2002). Language Use and Personality during Crises: Analyses of Mayor Rudolph Giuliani's Press Conferences. *Journal of Research in Personality*, *36*, 3, 271-282.

Pennebaker, J.W., Mehl, M.R., & Niederhoffer, K.G. (2003). Psychological Aspects of Natural Language Use: Our Words, Our Selves. *Annual Review of Psychology*, 54, 547-577.

Popper, K. (1957). *The Poverty of Historicism*. London: Routledge and Kegan Paul.

Rainie, L. (2005). *The state of blogging.* Pew Internet & American Life Project. Available at http://www.pewinternet.org/PPF/r/144/report_display.asp

Rall, T. (2005) *But who watches the watchdogs?* Yahoo News. Available at http://news.yahoo.com/news?tmpl=story&u=/ucru/20050223/cm_ucru/butwhowatchesthewatchdogs

Ramsay, R. (1968). Speech patterns and personality. *Language and Speech, 11* (1), 54-63.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

Rayson, P. (2001) *WMatrix: a web-based corpus processing environment*. Computing Department, Lancaster University.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Rayson, P., & Leech, G., & 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, Vol.2(1),* 133-152.

Lynn, R. (1993). Sex differences in competitiveness and the valuation of money in twenty countries. *Journal of Social Psychology, 133*, 507-511.

Rittenbruch, M., Mansfield, T., & Cole, L. (2003). *Making sense of "syndicated collaboration"*.

Rosenbloom, M. (Ed.) (2004). The Blogosphere. *Communications of the ACM, Special Issue, 47 (12)*. ACM Press, New York

Santini, M. (2005). Clustering Web Pages to Identify Emerging Textual Patterns. *RECITAL 2005*, Dourdan.

Santini, M. (2006). Interpreting Genre Evolution on the Web. *EACL 2006 Workshop: NEW TEXT - Wikis and blogs and other dynamic text sources*.

Schaffer, K.F. (1981). *Sex Roles and Human Behaviour*. Cambridge, MA: Winthrop Publishers.

Scherer, K. (1979). Personality markers in speech. In K.R. Scherer and H. Giles, (Eds), *Social Markers in Speech*, pages 147-209. Cambridge University Press, Cambridge.

Schler, J., Koppel, M., Argamon, S., & Penebaker, J. (2006). Effects of Age and Gender on Blogging. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Segerstad, Y.H.. (2002) *Use and adaptation of written language to the conditions of computer-mediated communication*. Göteborg: Department of Linguistics, Göteborg University.

Shepherd, M., & Watters, C. (1999) The Functionality Attribute of Cyber Genres. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, USA.

Shepherd, M., Watters, C., & Kennedy, A. (2004). Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering, 3*, (3&4), 236-251.

Siegman, A.W. (1978). The meaning of short pauses in the interview. *Journal of Nervous and Mental Disease*, 166, 387-406.

Siegman, A.W. (1987). The tell-tale voice: Nonverbal messages of verbal communication. In A. Siegman & S. Feldstein, (Eds.), *Nonverbal behaviour and communication*, pages 642-654. Erlbaum, Hillsdale, NJ.

Stam, R. (2000). *Film Theory*. Oxford: Blackwell.

Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2000a). Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, *26* (4).

Stamatatos, E., Fakotakis, N., and Kokkinakis, G. (2000b). Text Genre Detection Using Common Word Frequencies. In *Proceedings of The 18th Int. Conference on Computational Linguistics*.

Swales, J.M. (1990). *Genre Analysis*. Cambridge: Cambridge University Press.

Tannen, D. (1990). *You Just Don't Understand: Women and Men in Conversation*. New York: Harper Collins.

Thompson, E.P. (1982). *The Makings of the English Working Class*. Penguin, Harmondsworth.

Thomson R., & Murachver T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology*, *40* (pt. 2), 193-208.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, 53, 718-726.

Tong, R.M., & Snuffin, M. (2006). Weblogs as Market Indicators: Tracking Reactions to Issues and Events. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Twist, J. (2005). *Looming pitfalls of work blogs.* BBC News Online. Available at `http://news.bbc.co.uk/1/hi/technology/4115073.stm`

Ward, M. (2003). *A blog for everyone.* BBC News Online. Available at `http://news.bbc.co.uk/1/hi/technology/3078541.stm`

Weintraub, D. (2003). *Why a blog*. The Sacramento Bee. Available at `http://www.sacbee.com/content/politics/columns/weintraub/story/6414174p-7366437c.html`

Werry, C. (1996). Linguistic and interactional features of internet relay chat. In S.C. Herring (Ed), *Computer Mediated Communication: Linguistic, social and cross-cultural perspectives*, 47-63. Amsterdam: John Benjamin.

Wiggins, J, & Pincus A. (1992). Personality: Structure and assessment. *Annual Review of Psychology*, *43* (1), 473-504

Wilson, M. (1987). MRC Psycholinguistic Database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.

Wu, Y., & Tseng, B.L. (2006). Important Weblog Identification and Hot Story Summarization. *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, AAAI Technical report SS-06-03.

Yates, J., & Orlikowski, W.J. (1992). Genres of organizational communication: A structurational approach to studying communication and the media. *Academy of Management Review*, *17* (2), 299-326.

Yates, S. (1996). Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. In S.C. Herring (Ed.) *Computer Mediated Communication: Linguistic, social and cross-cultural perspectives*, 29-46. Amsterdam: John Benjamin.

Yates, S., & Graddol, D. (1996). ‘‘I read this chat is heavy’’: the discourse construction of identity in CMC. Centre for Language and Communication, Open University: Ms 1996.

Yellen, R., Winniford, M., & Sanford, C. (1995). Extraversion and introversion in electronically-supported meetings. *Information & Management*, *28* (1), 63-74.