THE UNIVERSITY *of* EDINBURGH

Title	Personality and language : the projection and perception of personality in computer-mediated communication
Author	Gill, Alastair James
Qualification	PhD
Year	2004

Digitisation notes:

- Page numbers 202,222,226 and 232 are missing in original pagination

Personality and Language: The projection and perception of personality in computer-mediated communication

Alastair James Gill

Doctor of Philosophy

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

2003

Abstract

Personality plays an important role in socialisation and collaboration, and people are able to form accurate impressions of other's personality from face-to-face interaction. However, the growing use of computer-mediated communication means that individuals are often faced with purely textual means for the projection or perception of personality. This thesis focuses on the projection and perception of personality in informal e-mail text of native English speakers. Here we examine the major personality traits of Extraversion (sociability) and Neuroticism (emotional stability), and also Psychoticism (tough mindedness). There are two hypotheses: Firstly, that personality is projected linguistically; Secondly, that personality can be perceived through language. Results were found supporting the two hypotheses, and the thesis has implications for the understanding of personality and language production and for methodology. These main findings are summarised as follows:

Personality is projected through language. Previous research has shown that content-analysis measures relate to personality. However, such top-down methods are often limited by constraints imposed by the analysis technique. Here it is shown that data-driven approaches from corpus linguistics—which provide more sophisticated information about context, syntax, and semantics—can give further characterisation of personality in language.

Personality is perceived through language. Personality can be perceived through face-to-face communication, internet chatroom environments and by observing strangers. Here, personality perception research is extended to the rating of short e-mail texts of around 200 words. It is shown that personality can be perceived, but as in other studies, this is mediated by each trait's observability and evaluativeness, and also by the environment.

Individual differences influence theories of language production. By using several different approaches to the analysis of personality language, it is shown that different personality traits influence different levels of language production.

New methodologies can inform individual differences. The adoption of techniques from computational corpus linguistics has revealed new features of personality language, and provided techniques more sensitive to smaller or non-standard data sets.

Acknowledgements

Gratitude is firstly due to my supervisors: To Jon Oberlander, for his encouragement and enthusiasm throughout this project, and also for his great generosity of time, energy and resources; To Richard Shillcock, for giving me the freedom to pursue various ideas, and for his insightful comments and discussions. Acknowledgement is also due to Elizabeth Austin, for her advice and guidance throughout, and also to my examiners Joe Levy and Helen Pain for the interesting (and somewhat lengthy!) viva and for their comments which helped smooth some of the rougher edges of my thesis.

I also gratefully acknowledge the Economic and Social Research Council, and the School of Informatics, for their financial support; Paul Rayson of UCREL, Lancaster University, for providing me with access to his corpus software; Betty Hughes and the staff of the ICCS/HCRC offices for making everything run smoothly behind the scenes; and also Tam Jardine for keeping me smiling, whatever the weather. I would also like to thank my colleagues who have who have provided comments, discussion, enthusiasm, encouragement and technical advice on various aspects of this project: James Curran, Mary Ellen Foster, Frank Keller, Mirella Lapata and Padraic Monaghan.

More generally, I also recognise those who, years previously, have shaped my current academic thinking: Steven Emsley for arousing my interest in language, and Chris Butler for recognising and directing this interest. I am also grateful for the encouragement and friendship they have offered since. Thanks to Chris in particular for his words of support which convinced me to persevere.

For their companionship along the journey I would like to acknowledge Conor Snowden and Tim Willis, in addition to my long-suffering friends, flatmates and officemates who have had to cope with me on a daily basis: their friendship and humour have allowed me to see the more amusing side of things.

Last but by no means least, I acknowledge my parents for their love and support over the many years, and who have made this possible. This work is dedicated to the memory of the paternal and maternal grandparents that I knew: To Mrs Edith Gill, I trust this is a fitting tribute to her vision and provision, and to Mr John James Lister, who taught me the important things not learnt through formal education.

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.

(*Alastair James Gill*)

"Personality is something we all once had."
Raymond Williams

To the memory of Mrs Edith Gill and Mr John James Lister

Table of Contents

List of Tables

Chapter 1

Introduction

In the introduction to this thesis we start by outlining our focus of interest, namely the relationship between personality and language. We then describe how this area of study relates to the wider fields of cognitive science, social psychology, and computational linguistics. We note also why the projection and perception of personality has particular relevance to computer-mediated communication.

The second half of the introduction describes in more detail the objectives of the thesis, and also where boundaries will be imposed upon the study. We conclude this first chapter with an outline of the structure of the rest of the thesis and of our main hypotheses.

1.1 Introduction to Personality and Language

We introduce our focus of study—personality and language—with the words of Louis Milic (Milic, 1966, p. 82):

> The fundamental assumption is that the style of a writer is an idiosyncratic selection of the resources of the language more or less forced upon him by the combination of individual differences summarized under the term "personality". This selection might be called a set of preferences except that this term suggests that the process is mainly conscious and willed. Although it is doubtless true that some part of the process of composition is deliberate and conscious, especially at the level of meaning, much of it is not fully conscious and it is this part which is of greater interest to the student of style. The reason is obvious: The unconscious stylistic decisions are less likely to be affected by the occasional and temporary characteristic of a given composition (its subject matter) and are more likely to reveal something the writer might be struggling to conceal. If we are interested in his personality, such information would naturally be of great interest [...]

Here Milic uses the term *personality* in the sense of referring to one individual in particular. In this thesis, we take a broader view—in line with the personality psychology literature—which describes personality in terms of traits which influence behaviour and capture the fundamental qualities of a person (Matthews and Deary, 1998). Examples of traits central to theories of personality are: Extraversion, which generally refers to sociability, and Neuroticism, which relates to a person's propensity to worry.[1] These traits are not purely theoretical constructs, but have been shown to have important practical implications for behaviour: Extraversion for the perception of social networks, and for co-operative situations (Casciaro, 1998; Koole et al., 2001), as well as task performance and information processing ability (see Matthews and Deary, 1998, for a review); Neuroticism for tasks involving interpersonal interaction, teamwork, and for counter-productivity in the work environment (Mount et al., 1998; Blackman, 2002b), and also for self monitoring (Jonassen and Grabowski, 1993). Personality psychology is therefore highly relevant to everyday life, as well as cognitive science and social psychology (Blass, 1984; Matthews, 1997).

[1] These descriptions are coarse characterisations; we will cover the exact definitions of these traits in the next chapter.

Like Milic, we are also concerned with how such individual differences in personality influence language behaviour in a way that is not 'conscious and willed', and this informs our first hypothesis of the thesis, that: *Personality is projected linguistically.* Furthermore, we also test the complementary hypothesis, which is that: *Personality can be perceived through language.*

1.1.1 Person Perception

Milic was primarily concerned with the written style of authors. However, we propose that personality affects language behaviour more generally (cf. e.g., Sanford, 1942; Ramsay, 1968): In this thesis we study a written form of language—computer-mediated communication (CMC), which is generally seen to share many similarities with spoken interaction (Bälter, 1998; Colley and Todd, 2002). Indeed, the study of the projection and perception of personality through e-mail is a particularly important issue, given the increasing popularity of the medium (Baron, 1998). In face-to-face interaction we are highly effective at judging people's personality (e.g., Funder and Dobroth, 1987; Funder and Colvin, 1988; Paunonen, 1989), or other characteristics, such as familiarity, gender, emotion or temperament (e.g., Cheng et al., 2001). However, e-mail is often used to make contact with people for the first time, but lacks many of the cues usually used for personality judgement in face-to-face situations. Additionally, given that a synchronic CMC environment is known to have implications for personality judgement (Hancock and Dunham, 2001a; Markey and Wells, 2002), we study the effects of asynchronous e-mail, upon person perception.

1.1.2 Language Generation

The study of personality and language can also be used to inform technological applications, for example in the user modelling of computer interfaces. Amichai-Hamburger (2002) proposes that the internet should be adaptive to the user's personality, and this may potentially be realised through automatically generating the language of web pages to match that of the user. In the human-computer interaction literature, there is evidence that computer users attribute personality to interfaces, and respond to it

in robust ways (e.g., Nass et al., 1995; Moon and Nass, 1996; Nass and Lee, 2000; Isbister and Nass, 2000). Even using manual linguistic manipulations and in a text-only environment, interfaces which used language associated with the personality of the interface user were preferred and rated as more attractive, credible and informative (Nass et al., 1995). Indeed the projection of personality by virtual-agents may lead to improved ratings of perceived social ability (cf. Burgoon et al., 2000).

The potential for modification and variation in communication style is a well-studied aspect of interpersonal interaction, particularly with regard to social perception (Bradac et al., 1976; Bradac and Mulac, 1984). Bradac (1990) states that 'all levels of language (i.e. phonology, syntax, semantics, and pragmatics) affect message recipients' beliefs about and evaluation of message sources' (p. 405), and this has been studied at various levels, for example, in terms of accent (e.g., Labov, 1972; Giles, 1973; Bell, 1984), and of other realisational properties of language (Bradac et al., 1980, 1988; Bradac, 1990).

Human responses to linguistic choices—whether they are of an artificial agent or human interlocutor—therefore have important implications for the communicative strategy adopted by interfaces, especially if they serve a pedagogic function (e.g., Person et al., 2001). In natural language generation (NLG) systems, attempts have been made to address this in two ways: Firstly, by developing systems which enable dynamic interface text generation tailored to the audience's level of expertise or previous knowledge of the topic (e.g., O'Donnell et al., 2001). However, this type of generation system is restricted to modifying content, and so can only manipulate *what* is said. Secondly, another approach is to modify *how* something is said. Work in natural language generation has focussed on the projection of linguistic style (DiMarco and Hirst, 1994; Hovy, 1996; Walker et al., 1997; Harrington, 2003), and also its detection (e.g., Argamon et al. 2003b; Finn and Kushmerick 2003; Koppel et al. 2003a; but cf. Reiter and Sripada 2002a,b), particularly in relation to giving embodied conversational agents more 'human-like' properties.

These 'human-like' properties are often framed in terms of personality, which is not surprising, given that personality psychology is concerned with describing the fundamental qualities of a person. Language generation systems which have resulted from

this research, for example may simply exhibit paranoia (Colby et al., 1971), or have used other pre-determined personality parameters (Walker et al., 1997) to determine social interaction strategies (e.g., implementing those of Brown and Levinson, 1987). Other NLG systems have responded to emotional information about the user (Ball and Breese 1998; Fleischman and Hovy 2002; see also Cañamero 1998 and Norman et al. 2003 for further discussion), whether this is assumed, or detected (Picard, 2000; de Vicente and Pain, 2002). Furthermore, attempts to incorporate the modelling of both personality and emotion have been proposed which would allow interaction between these levels to influence language generation (Moffat, 1991; Kshirsagar and Magnenat-Thalmann, 2002). Whilst these systems have developed increasingly advanced models of personality (and also emotion), there is relatively little research to demonstrate how personality systematically influences language production in humans. Therefore, as a result of our research we will be able to inform NLG systems to enable the more realistic generation of personality language.

1.2 Objectives

Personality, as we have just shown, is an interesting and important concept which influences behaviour, interaction, and interpersonal relationships. The expression of personality through language behaviour also has implications for natural language generation, embodied agents, person perception and impression formation. Additionally, given the growing use of computer-mediated communication, individuals are often faced with purely textual information mediating the projection or perception of personality.

Using computer-mediated communication as the domain of study, in this thesis we investigate two hypotheses: First, that personality is projected linguistically; Secondly, that personality can be perceived through language. The methodology of the thesis is based around the construction of a personality corpus. We collected this using experimental web techniques, and use it to test our hypotheses.

We test Hypothesis 1 using two main approaches: content analysis based upon psychological and psycholinguistic properties of the text, and empirical comparison

techniques from computational corpus linguistics to identify characteristic features. We will show that whilst the former analysis performs inconsistently across the different personality dimensions, the latter technique provides more reliable results with our data. We will conclude that a combination of these techniques gives a more comprehensive description of personality language for these traits.

Hypothesis 2 is tested using subjective rating of the salience of author personality in our texts. Here we will show that personality can be accurately perceived from asynchronous textual communication, but that accuracy is mediated by the personality trait in question. The usefulness of linguistic features in these ratings, and the interaction of judge personality are also evaluated.

1.3 Boundaries of the thesis

In addition to identifying the question that this thesis will address, it is also necessary to identify the boundaries of the study. In this way, we can focus on areas which are of primary interest without attempting to cover absolutely everything.

Although this thesis is about personality, it is not about *personality psychology*. The result is that although extensive reference will be made to personality traits, like Extraversion and Neuroticism, these will be used to inform the study, and so they will not be the *object* of study themselves. Therefore we do not intend to question the different theories of personality, or the traits themselves, although we will provide a brief overview of the background to these phenomena.

Similarly, although the focus of this study is language, it is not *language production*, whether this is in psycholinguistic or natural language generation terms. We will however, refer to work in such areas where relevant to the thesis, and we will frame our results with reference to the way they inform—and can be informed by—models of language production and generation.

Finally, in this thesis we study the ways in which personality is projected and perceived through language. Although we specifically look at e-mail data, we view this as being directly relevant to other computer-mediated environments, and also—to varying extents—informative for other forms of communication. Here our focus is on lan-

guage, with computer-mediated communication providing the means for investigation, rather than language providing a method for studying computer-mediated communication.

1.4 Structure of the thesis

The thesis is structured as follows. Chapter two provides a survey of the areas that this thesis draws upon: Personality and the theories relating to its measurement are introduced, and this is followed by a description of the traits which we will be using in this thesis. On the basis of personality theory, we then propose differences in linguistic behaviours which we predict will result from author personality (theoretical hypotheses). The perception of personality, and the role of different factors in the accuracy of judgement are also discussed. The rest of this chapter examines the previous findings for personality and its effect upon language behaviour (summarised in hypotheses from previous work), examines the methods which will be employed in the analysis of our personality corpus, and briefly discusses implications of computer-mediated technology on experimentation and communication.

Chapter three describes the construction of the e-mail personality corpus which forms the basis of the experimentation conducted in this thesis. This is informed and validated with reference to a previous study which used content analysis methods on texts encoded with author personality.

Chapter four builds upon the psychological content analyses carried out in the previous chapter in two ways: by adopting alternative statistical methods, and also by extending the analysis in order to measure psycholinguistic textual properties and lexical diversity.

Chapter five proposes the adoption of data-driven analysis techniques derived from computational corpus linguistics: here we examine different methods for dividing the corpus, annotation, and comparing the resulting subcorpora.

Chapter six extends the data-driven computational linguistic approaches of the previous chapter further by exploring different annotation methods. The use of grammatical and semantic categories avoids some of the problems associated with data sparsity,

and allows the greater generalisation of findings.

Chapter seven describes the experimental perception study which tests people's ability to accurately judge author personality on the basis of a brief e-mail text. Here we additionally describe the way that the judge's perceptions of the e-mail text's author is related to awareness of personality in general.

The final chapter presents a summary of the thesis and its conclusion. We discuss the boundaries of this work and the implications for future research.

1.5 Summary and Hypotheses for the thesis

In this chapter we have introduced the focus of study for this thesis—the relationship between personality and language—and have described its importance, particularly with reference to computer-mediated communication. The ways in which personality and language relate to the wider fields of cognitive science and computational linguistics were overviewed.

We then described in more detail the objectives and boundaries of the thesis, and have outlined the structure of the rest of the thesis. The main hypotheses which we have described in this chapter are summarised as follows. The major goals of the thesis are to test whether:

Hypothesis 1 Personality is projected linguistically.

Hypothesis 2 Personality can be perceived through language.

In the next chapter we review previous work in relevant areas which will allow us to furnish these hypotheses with more specific questions—both from theory and from empirical work—to be addressed by this thesis. We therefore turn to the literature review.

Chapter 2

Literature Review

This thesis is interdisciplinary in nature and draws upon the research and methodologies from a number of diverse areas. Therefore our overview of relevant literature reflects this variety. We do not attempt to cover all areas exhaustively, but wherever relevant, references are provided to enable the reader to pursue a topic in more detail.

The literature review is divided into two main sections: first we present work which informs this study on the basis of topic. Here we overview theories of personality, the perception of personality, and also the results of work which has examined the effects of personality on language. After each of these three sections we present hypotheses, based on theory or on previous findings, which we take forward for further examination in this thesis.

Secondly, we present work which is relevant from a methodological stance. This discusses issues surrounding computer-mediated interaction and experimentation, and also gives an overview of linguistic analysis methods with specific relevance to corpora. The chapter then concludes with a summary and presentation of the hypotheses.

2.1 Introduction to Personality

2.1.1 Theories of Personality

In this thesis we refer to two main models and associated measurements of personality, Eysenck's three-factor model (Eysenck and Eysenck, 1991; Eysenck et al., 1985), and the five-factor model (Digman, 1990; Costa and McCrae, 1992b; Wiggins and Pincus, 1992; Goldberg, 1993). These are termed 'trait' approaches to personality, and reduce personality to a number of essential descriptive traits, or factors.

Each factor should be regarded as a scale ranging from 'low' to 'high', with a score possible at either extreme or anywhere in between. These factors are considered to be orthogonal and independent of each other, and therefore if an individual has a particular score on one factor, this does not necessarily predict any of their scores on any of the other factors. However in practice there may be some relationship between traits, for example, especially in the case of extreme scorers (cf. Eysenck, 1970; Matthews and Deary, 1998; Buckingham et al., 2001).

These traits traditionally assume a 'causal primacy' and 'inner locus', namely that traits influence behaviour, and that they relate to the fundamental, core qualities of the person (Matthews and Deary, 1998). Therefore, each trait is assumed to be relatively stable over time, and this distinguishes them from more transitory aspects of an individual like mood or emotion. We do not address these transitory aspects here, but for an overview integrating personality, mood, and the cognitive processing of emotion, see Rusting (1998).

We do not claim that 'traits' are the only way of describing personality, and that this approach or its assumptions are undisputed: these have been challenged by proposals that behaviour is learnt as a response to stimuli and that concepts of traits are constructed as a result of social situation. In response to such arguments Funder (2001) confidently sees the increase in recent research as indicating a growing acceptance and validity in trait approach. However Matthews and Deary (1998) are more reserved, placing emphasis on the theoretical basis of traits as being essential for establishing them as scientifically useful constructs.

Indeed, within personality psychology there is debate about the precise number of

traits which can be use to describe personality, hence the existence of the three- and five-factor models, amongst others. Essentially, the first two traits associated with major models—Extraversion (perhaps better described as Extraversion-Introversion) and Neuroticism (Emotionality-Stability)—are undisputed and central to theories of personality (Matthews and Deary, 1998; Lippa and Dietz, 2000). Indeed the EPI (Eysenck Personality Inventory; Eysenck and Eysenck, 1964) which solely measures these traits, is described by Costa and McCrae (1986) as the 'gold standard' for many researchers, with the three-factor EPQ-R equally respected (Kline, 1993b; Ferrando, 2003).

Where these two approaches diverge then, is most obviously represented in the number of factors which they claim describe personality, but perhaps more significantly in their theoretical basis: Eysenck claims a 'biological basis' for his model of personality (Eysenck, 1970; Eysenck and Eysenck, 1991), and has emphasised that a trait's validity is based upon this—along with for example, its cultural invariance, and relationship to social behaviour and illness (Eysenck, 1993); Proponents of the five-factor model do not make such claims, and instead using the 'lexical hypothesis' approach have derived factors which group statistically, with validity demonstrated by further replication of these factors (McCrae and Costa, 1987, 1997; Funder, 2001; cf. Block, 1995; Matthews and Deary, 1998; however attempts have been made to redress this lack of theoretical basis, e.g., Buss and Finn, 1987; Pytlik Zillig et al., 2002). Here we distinguish between the term 'lexical hypothesis' which describes a contribution to personality theory, and this thesis's concern with studying personality and its influence on language: The former approach studies words *which describe particular aspects of personality* and sets about deriving broad factors which represent traits related to personality; By contrast, the latter approach employed here is concerned with the words *which are used by speakers of different personality types* (work in this area is described in more detail in Section 2.2).

Turning now to describe the traits, and the subsequent factors which are measured: In the three-factor model there is one further trait, namely Psychoticism, whereas in the five-factor model, the remaining variance is described in terms of Openness, Agreeableness and Conscientiousness. Eysenck argues that Agreeableness and Conscientiousness are primary level traits which both form facets of Psychoticism (negatively

related[1]; Eysenck 1991, 1992, and which to some extent is supported by the early studies of McCrae and Costa 1987). Conversely, Costa and McCrae and others have argued that five factors are required to describe personality fully (Costa and McCrae, 1992a; McCrae and Costa, 1997; Digman, 1990; Goldberg, 1993).

Indeed, recent work looking comparing measures of normal and dysfunctional personality suggests many of the questions such as 'which model?' or 'how many traits?' are not productive. With reference to the three- and five-factor models, Larstone et al. (2002) instead note that 'each instrument is an imperfect measure of personality that shares components of variance with the other while also tapping specific dimensions' (cf. interpretations of Extraversion; Depue and Collins, 1999). In contrast, Kline (1993b) considers that EPQ 'Extraversion and Neuroticism are clearly identical to two of the big five factors and Psychoticism would appear to be a mixture [of the other traits]'.

It is not a goal of this thesis to prove or disprove any particular theory or model of personality, nor will it provide further discussion of the surrounding debate. For a concise and lively debate of trait approaches to personality, see Deary and Matthews (1993) and the associated peer commentaries, with more extensive and detailed information found in Matthews and Deary (1998). For a more general discussion of personality, individual differences and related concerns, the reader is directed to Cooper (1998). In the rest of the thesis we will draw on these two main models of personality, along with associated concepts. In the experimentation which we describe below, we have primarily adopted the three-factor EPQ-R (Eysenck et al., 1985; Eysenck and Eysenck, 1991) as our measurement of personality. We do not regard this as necessarily subscribing to one particular theory of personality, however here we outline our justification of this choice: *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 (cf. Dewaele and Furnham, 2000; Dewaele, 2002a). Although the current study is concerned with language, and thus interaction, we have chosen not to use a specifically interpersonal measure (e.g., Wiggins, 1979; Kiesler, 1983), since they appear easily in-

[1] A further proposal is that Openness forms a part of Extraversion, and low Conscientiousness a part of Neuroticism, but we will not address this claim here.

corporated into more general models of personality, which allow greater comparability (McCrae and Costa, 1989; Trapnell and Wiggins, 1990). In terms of *validity* of the personality model and its measurement, Kline (1993a) regards these three factors (E, N, P) as having 'considerable external validation' experimentally and correlationally (p. 304). This is desirable, since we shall be relating these personality traits to features of language behaviour. In summary, Kline (1993b) states (p. 454):

> If we want a reliable and valid measure of these three basic personality factors the EPQ is as good as can be desired. It represents a clear marker in personality space. Its only flaw [...] is that the factors are broad[...] In brief it is a benchmark personality test[...]

Practically, this broadness viewed by Kline as a disadvantage, for us represents an advantage: three rather than five factors provide a more reduced model of personality with which to work. In future the research can be extended to incorporate five factors if the variation associated with Psychoticism appears to be better described in terms of Openness, Agreeableness and Conscientiousness.

Below, we provide a brief outline and definition of Extraversion, Neuroticism, and Psychoticism, incorporating Eysenck's description based on a 'typical' individual (Eysenck and Eysenck, 1975). We also briefly include details relating to Eysenck's biological theories of the traits, although as noted above we do not necessarily subscribe to them. To ensure applicability to other theories of personality, we briefly discuss them in relation to the traits described in the five-factor model.

2.1.2 Personality Traits

2.1.2.1 Extraversion

Extraversion in one of the most salient and visible personality traits (Funder, 1995), and one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993, p. 367). In their original theorising from a biological perspective, Extraversion is regarded as related to the degree of inhibition and excitation present in the central nervous system. Eysenck and Eysenck (1975) view this as largely inherited, but which may be mediated by the ascending reticular formation. They describe the realisation of the trait thus (p. 9):

> The typical extravert is sociable, likes parties, has many friends, needs to have people to talk to, and does not like reading or studying by himself. He craves excitement, takes chances, often sticks his neck out, acts on the spur of the moment, and is generally an impulsive individual. He is fond of practical jokes, always has a ready answer, and generally likes change; he is carefree, easy-going, optimistic, and likes to "laugh and be merry". He prefers to keep moving and doing things, tends to be aggressive and lose his temper quickly; altogether his feelings are not kept under tight control, and he is not always a reliable person.
>
> The typical introvert is a quiet, retiring sort of person, introspective, fond of books rather than people; he is reserved and distant except to intimate friends. He tends to plan ahead, "looks before he leaps" and distrusts the impulse of the moment. He does not like excitement, takes matters of everyday life with proper seriousness, and likes a well-ordered mode of life. He keeps his feelings under close control, seldom behaves in an aggressive manner, and does not lose his temper easily. He is reliable, somewhat pessimistic, and places great value on ethical standards.

In their NEO-PI-R model of personality, Costa and McCrae (1992b) divide each of the personality dimensions into six further facets, each of which indicate a lower-level property of the trait. For the NEO-PI-R Extraversion dimension, these facets are: Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, and Positive Emotion.

2.1.2.2 Neuroticism

Neuroticism[2]—and also Extraversion—are regarded as 'clearly marked and outstandingly important dimensions' (Eysenck and Eysenck, 1964, p. 5), and form the core of personality descriptions. Like Extraversion, Eysenck and Eysenck (1975) view it as largely inherited, and is closely related to the degree of liability of the autonomic nervous system. They describe the trait as (pp. 9–10):

> [...] we may describe the typical high N[euroticism] scorer as being an anxious, worrying individual, moody and frequently depressed. He is

[2]Note that in this thesis we use the terms Neuroticism—and also Psychoticism—which refer to particular personality traits, and should therefore be regarded purely as technical descriptions with specific definitions. However to avoid possible negative associations, when communicating with audiences outside personality psychology, it is usual to adopt the alternative terms 'Emotionality' for Neuroticism and 'Tough-mindedness' for Psychoticism (Eysenck and Eysenck, 1975).

> likely to sleep badly, and to suffer from various psychosomatic disorders. He is overly emotional, reacting too strongly to all sorts of stimuli, and finds it difficult to get back on an even keel after each emotionally arousing experience. His strong emotional reactions interfere with his proper adjustment, making him react in irrational, sometimes rigid ways. [...] If the high N individual has to be described in one word, one might say that he is a *worrier*; his main characteristic is a constant preoccupation with things that might go wrong, and a strong emotional reaction of anxiety to these thoughts.The stable individual, on the other hand, tends to respond emotionally only slowly and generally weakly, and to return to baseline quickly after emotional arousal; he is usually calm, even-tempered, controlled and unworried.

In their interpretation of Neuroticism, Costa and McCrae (1992b) claim that it is related to psychological well-being, referring to their previous work (Costa and McCrae, 1984) which showed each of the six facets is significantly related to negative affect and lower life satisfaction. The six facets of Neuroticism are: Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, and Vulnerability.

2.1.2.3 Psychoticism

Psychoticism, as observed previously, is the most contentious trait of the three-factor model, and was a later addition to the existing traits of the EPI model, Extraversion and Neuroticism (Eysenck and Eysenck, 1964, 1975; Eysenck et al., 1985). It was conceived to be related to behavioural disorders, but designed to measure individuals belonging to a 'normal' population, rather than those displaying extreme pathological symptoms. Subsequent research has suggested that it is indicative of thoughtless or reckless personality and that high Psychoticism scorers are predisposed to personality traits associated with an excess of severe and threatening life events (Pickering et al., 2003). Its biological basis is traditionally regarded to be related to androgen levels in the individual, and this may explain the higher Psychoticism scores present in males (Kline, 1983).

It is described as follows (Eysenck and Eysenck, 1975, p. 11):

> A high [Psychoticism] scorer, then, may be described as being solitary, not caring for people; he is often troublesome, not fitting in anywhere. He

> may be cruel and inhumane, lacking in feeling and empathy, and altogether insensitive. He is hostile to others, even with his own kith and kin, and aggressive even to loved ones. He has a liking for odd and unusual things, and a disregard for danger; he likes to make fools of other people and to upset them. [...] Socialisation is a concept which is relatively alien to such people; empathy, feelings of guilt, sensitivity to other people are notions which are strange and unfamiliar to them.

The simplest interpretation of these two models maps the NEO-PI-R traits conscientiousness and agreeableness negatively onto EPQ-R Psychoticism. Costa and McCrae (1992b) describe the facets which compose these traits as: Trust, Straightforwardness, Altruism, Compliance, Modesty, and Tender-Mindedness, for Agreeableness; Competence, Order, Dutifulness, Achievement Striving, Self-Discipline, and Deliberation, for Conscientiousness. Here we can indeed observe the relationship between Psychoticism and NEO-PI-R Agreeableness and Conscientiousness, with this proposed inverse relationship providing a useful way of comparing findings across personality models.

In addition to Conscientiousness and Agreeableness, Costa and McCrae (1992b) also describe the facets of Openness as: Fantasy, Aesthetics, Feelings, Actions, Ideas, and Values. Although Eysenck proposes that Openness is related to Extraversion, for simplicity of interpretation, here we will disregard such a possible link (Eysenck, 1991, 1992). In addition to the three personality dimensions of Extraversion, Neuroticism, and Psychoticism, the EPQ-R model also incorporates a 'Lie Scale', which measures an individual's tendency to avoid admitting undesirable characteristics. Although this is not regarded as a personality trait, it can be incorporated into personality studies as an indicator of deception, for example, spouse perceptions of a partner's deceptiveness (Gomà-i-Freixanet, 1997).

2.1.3 Language hypotheses from Theory

On the basis of the personality descriptions of the EPQ-R (Eysenck and Eysenck, 1975), we propose the following realisations of personality through language.[3]

[3]Note that in contrast to the hypotheses of Furnham (1990) (see Section 2.2.1), here we are solely concerned with language behaviour which can be realised in a written form, equivalent to his categories of *Fluency*, *Morphology and syntax*, and *Conversational behaviour*. We therefore ignore features which are unique to spoken language, for example, the category *Voice (frequency, intensity and quality)*.

Extraversion We expect the language of high Extraverts to reflect their sociability by referring to other people, and to express their activity by using more words describing action, and by saying more. We also expect them to use language which suggests positive affect.

Neuroticism The language of high Neurotics, we expect to be highly emotional—particularly expressing negative affect, but also positive affect—and this is also revealed through intensified language (e.g., adjectives and adverbs). Since the individual is a worrier, we also expect this self-preoccupation to be expressed through an increased reference to self.

Psychoticism We expect highly Psychotic individuals to reflect their lack of sociability and detachedness by making fewer references to themselves or to others, and to demonstrate their harshness and toughness by avoiding emotional words. Since they are creative and enjoy unusual things, we predict that they will adopt a more unusual language use, realised both in words and constructions (i.e., lexically and syntactically).

2.2 Personality and Language

In reviewing the work which has looked at personality and language, there are a number of observations which can be made: Firstly, that there has not been a great deal of work in this area, and that which has been done tends to use incommensurable[4] approaches and is spread across different disciplines; Secondly, the majority of work has focussed on speech; Thirdly, research has tended to focus on traits relating to Extraversion-Introversion, and to a lesser extent Neuroticism-Stability, rather than others from the three-factor (Psychoticism), or five-factor (Conscientiousness, Agreeableness, Openness) models of personality.

The main explanation of this appears to be the interdisciplinary nature of this research question, with it touching upon the fields of personality theorists, social psychologists of language, and also psycholinguists and sociolinguists (Furnham, 1990).

[4] Although Furnham (1990) describes the work in the field as 'inconsistent', here we use the term 'incommensurable' as a less value-laden alternative.

Indeed the reason why little work has addressed this question may be that from a personality perspective, language is regarded as not important, interesting, or not being high-enough level behaviour (indeed the relative value of verbal or non-verbal behaviour is not always apparent; cf. Eckman et al., 1980; O'Sullivan et al., 1985); whereas from a social or linguistic perspective, other factors (e.g., social or situational) are seen as more important to language behaviour, with debate surrounding the definition and stability of personality traits, and with greater interest relating more to the inference of *perceived*, rather than *actual* personality from speech. The incommensurability of approaches results from the different methodologies adopted by these disciplines, and indeed the variety of approaches available in each (e.g., personality theory and measurement, or level or type of linguistic analysis).

The focus of these studies appears to be determined by saliency both within and outside the fields: Speech is the most ubiquitous form of language, and includes paralinguistic features, such as pronunciation, intonation or loudness, and which can be seen to vary readily across individuals due to, for example, social or geographical reasons; similarly, Extraversion is a highly salient personality trait (Funder, 1995), and therefore draws the focus of investigation, rather than, for example Neuroticism, which is equally central to major theories of personality, but is less salient (Lippa and Dietz, 2000).

2.2.1 Previous hypotheses

Reflecting the focus on Extraversion more generally, Furnham (1990) has proposed the following features based on a knowledge of the characteristics of Extraverts and Introverts. Extravert language is less formal, has a more restricted rather than elaborated code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely, which he defines as how correct or unusual are the words used. With specific relevance to speech, he also proposes that Extraverts will tend to use an accent which is more non-standard (or local), rather than standard (or received), that they talk at a faster speed, and that their speech will contain more dysfluencies.

Other speculations based on an intuitive knowledge of the personality types, sug-

gest more higher-level features of Extravert-Introvert language: Extraverts are individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts monopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (Teiger and Barron-Teiger, 1998).

2.2.2 Previous Findings

In the following review of the literature which has examined the area of personality and language, we divide the results into four areas of research (cf. Scherer, 1979; Furnham, 1990): voice, fluency, morphology and syntax, and conversational behaviour. For each section, we first present findings for native English speakers, although where appropriate, we also report studies from other languages or from non-native speakers. Due to the particular emphasis on research into Extraversion, the majority of findings reported are for this trait, although other traits are discussed where relevant. We do however, give preference to studies utilising recognised measures of personality (e.g., three- or five-factor model), and which report statistically significant results.

Finally, given that the interest of this thesis is in written language, e-mail and computer-mediated communication, we will focus our attention on the linguistic levels most relevant. Therefore, the review is as follows: we briefly touch on acoustic findings for 'voice'; describe mainly results relating to speech rate and errors in 'fluency'; primarily cover use of grammatical features in 'morphology and syntax'; finally, our most extensive review concentrates on 'conversational behaviour'. In addition to interpersonal behaviour, here in the final section, we also include findings relating to the content and topic of language. In all these cases, findings from studies of spoken language will be reported where relevant to written language. For further information: a general overview of issues relating to the area can be found in Furnham (1990), work on speech and personality perception is covered in more detail in Scherer (1979), and for discussion of Extraversion findings—particularly with relevance to second language learning—consult Dewaele and Furnham (1999). See also Pennebaker and King (1999) and Smith (1992) for work relating individual differences to content analysis.

2.2.2.1 Voice

Here we provide a brief summary for findings relating to voice. Findings specifically looking at the speech of American Extraverts found that they were perceived to talk louder and with a more nasal voice (Scherer, 1978). For English as second language, we find that high Extravert speakers score lower for pronunciation (Busch, 1982).

2.2.2.2 Fluency

The majority of findings which relate to fluency report an Extravert advantage, and are reported in terms of speech rate. Extraverts have higher speech rates (Siegman, 1987), which is the case in both informal and formal settings (Dewaele, 1998; Dewaele and Furnham, 2000). Extraverts also show an inverse relationship with silence quotient (derived from silent pauses and speech rate, Siegman, 1978; but cf. Dewaele, 1998 for issues of silent pauses and measurement of speech rate), and in more complex verbal tasks, Introverts' pauses were significantly longer before speaking, than was the case for the Extraverts (Ramsay, 1968). Extravert children and teenagers showed greater verbal fluency for simple and complex recall tasks (Tapasak et al., 1979). Additionally it has been found that in formal situations Extraverts show less hesitation ('er'), but also make a higher proportion of semantic errors (Dewaele and Furnham, 2000).

2.2.2.3 Morphology and Syntax

Here we describe findings mainly relating to grammatical features. Extraverts show higher counts of pronouns, adverbs, verbs and total number of words (Cope, 1969; taking 'zestful' to be a synonym for Extravert, cf. Furnham, 1990; Dewaele and Furnham, 1999). These characteristics of Extravert language are also found for non-native speakers: Using factor analysis of syntactic tokens, Dewaele and Furnham (2000) describe this as implicit language (preference for pronouns, adverbs and verbs), which contrasts with the explicit language characteristics of Introverts (nouns, modifiers and prepositions). This finding relates to both informal and formal situations, and mirrors previous analyses of the individual linguistic categories (Dewaele, 1996b,a). Additionally, Heylighen and Dewaele (1999) more generally note that introvert language

features tend to be closely related to those of formal language.

An additional finding is that Extraverts demonstrate lower lexical richness in formal situations (controlling for length; Dewaele, 1993; Dewaele and Furnham, 2000). Although Cope (1969) also notes a lower lexical diversity (measured as type-token ratio; TTR), for Extravert native English speakers; however this would appear to be less reliable given that Extraverts also use a greater total number of words, and thus may be explained as an effect of length (cf. Gill, 1998). Low type-token ratio is also related to language produced in anxiety promoting situations (Howeler, 1972), and is also implicated in the perception of greater anxiety (Bradac, 1990).

2.2.2.4 Conversational behaviour (and content)

Firstly, we look at the results which relate personality to interpersonal aspects of language. As would be expected, Extraverts show greater desire to communicate and initiate interactions (McCroskey and Richmond, 1990), which is also found in computer-mediated communication (Yellen et al., 1995). In terms of conversational behaviours, analysis of speech acts shows that Extraverts initiate more individual and group laughter (Gifford and Hine, 1994). Gifford and Hine also found that Extraverts talk more, with other studies finding that they use a greater total number of words (Campbell and Rushton, 1978; Carment et al., 1965; although cf. Thorne, 1987 who found no significant differences in talk time or number of speech acts between Extraverts and Introverts). Perceptions of transcribed texts found that longer texts were regarded as displaying greater dominance and competence (Berry et al., 1997).

In terms of individual utterances, studies of second language speakers have shown that the length of the longest utterances produced by Extraverts is actually shorter, especially in informal conditions (Dewaele, 1995; Dewaele and Furnham, 2000). Additionally, Dewaele (2002b) finds that in L3 English production, Psychoticism and Extraversion showed a strong negative relationship to communicative anxiety, whilst Neuroticism showed a positive relationship.

We now turn from interpersonal language behaviour to looking at how personality affects the content of language used. In a study of conversational dyads, coding of the speech acts found that introverts used more hedges and problem talk, namely express-

ing qualification, and dissatifaction with one's own activities, but that Extraverts expressed more pleasure talk, agreement, and compliments, with content focusing more on extracurricular activities (Thorne, 1987). Extraverts have also been shown to use more self-referent statements (Gifford and Hine, 1994).

Here we also report findings from the Linguistic Inquiry and Word Count (LIWC; Pennebaker and Francis, 1999) text analysis program. These are discussed in relation to conversation behaviour because LIWC is primarily concerned with content. Although some syntactic features (e.g., pronouns, and verbs of various tenses) are included in the analysis, these are not derived from a part-of-speech analysis of the data. Indeed, no explicit claim is made that LIWC offers syntactic analysis. Note that we cover the results for LIWC studies in some detail, since this is an analysis method we will adopt in this thesis. For more specific details and discussion about the the LIWC text analysis method, see Section 2.5.3.4.

In a perception study Berry et al. (1997) found that transcribed texts rated as higher in Dominance used fewer positive emotion words and self referents, and texts regarded as displaying greater Competence used fewer self referents and negations, and more present tense verbs (these texts also show lower lexical diversity, but this appears to be a length effect, see Gill, 1998, for a discussion).

Pennebaker and King (1999) have applied LIWC analysis to texts written by authors for whom (five-factor) personality information was available. Using factors derived from the LIWC features, they found that: Extraversion showed a strong negative relationship with Making Distinctions (reflected in factor loadings of greater use of discrepancies, exclusive, tentative words and negations; fewer inclusive words) and a positive relationship with The Social Past (factor loadings showing greater use of past tense, present tense, and social words; fewer positive emotion words); Neuroticism correlated positively with Immediacy (the factor being composed of greater first person singular and discrepancies; fewer articles or longer words); Agreeableness—like Neuroticism—correlated positively with Immediacy, although conversely Openness showed a strong negative relationship with this factor; Conscientiousness—like Extraversion—showed a strong negative relationship with Making Distinctions.

Relationships between the personality dimensions and individual LIWC variables,

shows that high Extraverts use more social and positive emotion words, and fewer negations, tentativity, exclusive, inclusive, causation, negative emotion words, and articles; High Neurotics use more first person singular and negative emotion words, and fewer positive emotion words, and articles; High Openness scorers use more articles, longer words and insight words, and fewer first person singular, present tense, and causation words; High Agreeableness scorers use more first person singular and positive emotion words, and fewer articles and negative emotion words; High Conscientiousness scorers use more positive emotion words, and fewer negations, negative emotion, causation, exclusive words, and discrepancies.

2.2.3 Personality language hypotheses from previous work

On the basis of the previous findings reported above, we summarise them in terms of the following hypotheses, with particular reference to the features of computer-mediated communication (see Section 2.4.2). Here, note that we describe 'voice' features in terms of 'realisation', and additionally describe the hypotheses from LIWC separately from the conversational features:

2.2.3.1 Extraversion Hypotheses

Realisation Extravert 'loudness' will be realised in the increased use of capital letters and exclamation marks; Worse pronunciation will result in worse spelling and more typographical errors.

Fluency Higher speech rate of Extraverts will be realised in longer sentences; shorter pauses and less hesitation will result in more ellipses (as in the punctuation feature '...') and hyphens (-) being used to separate clauses rather than the full stop.

Grammatical Extravert language will contain more adverbs, pronouns, and verbs (i.e., more 'implicit'), and have a lower lexical density (TTR); it will contain fewer nouns, modifiers and prepositions (less 'explicit'), and be less formal.

Conversational Extraverts will write more; initiate more laughter, perhaps indicating

this by explicit references ('ha') or by exclamation; they will refer to themselves more; they will use more terms indicating pleasure and agreement, pay more compliments; they will use fewer hedges and references to problems, and will show less anxiety during the communication.

LIWC Extraverts will show a greater use of social and positive emotion words, and use fewer negations, tentativity, exclusive, inclusive, causation, negative emotion words, and articles; In terms of factors, Extraversion will have a negative relationship with Making Distinctions and positive relationship with The Social Past factors.

2.2.3.2 Neuroticism Hypotheses

Conversational High Neurotics will use a lower lexical density (TTR), and show greater anxiety during communication, realised explicitly through references to 'worry' or 'stress'.

LIWC High Neurotics will use more first person singular and negative emotion words, and fewer positive emotion words, and articles; Neuroticism will also correlate positively with the Immediacy factor.

2.2.3.3 Psychoticism Hypotheses

Conversational We expect high Psychotics to show less anxiety during communication, for example, through fewer explicit references to 'stress' or 'worry'.

LIWC Here we predict Psychoticism will show an inverse relationship to Agreeableness and Conscientiousness: based on Agreeableness, we expect fewer first person singular and positive emotion words, and more articles and negative emotion words, and also a negative correlation with the factor Immediacy; on the basis of the findings for Conscientiousness, we expect fewer positive emotion words, and more negations, negative emotion, causation, exclusive words, and discrepancies, and a positive relationship with the factor Making Distinctions.

2.3 Perception of Personality

In this section we overview personality perception, including the effects of different media. However, we do not cover studies which look at the specific effect of linguistic features upon person perception; these have been covered previously in the section reviewing findings for personality and language (Section 2.2.2).

Personality judgement data can be gathered in several ways. On the one hand, targets' self-reports of personality, together with ratings of these targets by peers (such as spouses or colleagues), have been compared with each other for agreement. On the other hand, strangers have been called upon to make personality judgements, after being exposed to various different kinds of information about the target individuals.

For many years, studies investigating personality perception had been stalled due to proposals that the errors in perception should be the focus of investigation, rather than accuracy (Cronbach, 1955). However, recently there has been an increased interest in measuring accuracy of perception (Funder, 1987; Kenny and Albright, 1987). In turn, studies have focused upon the investigation of factors which influence accuracy (Kenny, 1994; Funder, 1995), and here we focus upon one model to describe these factors. Funder's (1995) Realistic Accuracy Model views accuracy of judgement as a function of the relevancy, availability, detection, and utilisation of relevant behavioural cues. Furthermore, he outlines (pp. 658–63) a 'path to accurate judgement', which grounds these processes in terms of the quality of the 'judge', 'target', 'trait', and 'information' in the study. We therefore adopt these categories overviewing the area.

2.3.1 Judges

Good and bad judges are distinguished by their differing use of the cues which are available to them, for example, Funder (1995) proposes that knowledge about personality and the way it is revealed in behaviour would favour better socialised judges. This therefore implies that Extraverts make better judges than Introverts, because they 'have more experience in social settings than introverts', and Funder cites studies which have shown this to be the case for non-verbal cues in social interaction (Akert and Panter, 1988), and in determining the authenticity of suicide notes (Lester, 1991).

Given that this model is concerned with perception of personality, and has emphasised the role of judge personality, Funder also acknowledges the implications of judge ability and motivation: specifically, he notes the importance of intelligence, and also that the judge considers making an accurate decision to be important. Recent work by Lippa and Dietz (2000) found a more complicated picture in the judgement of Extraversion and Neuroticism: for the former trait intelligence shows correlation with greater accuracy, however judge Openness appears negatively related to accuracy in rating Neuroticism, although greater agreement is found for female—than male—judges rating Neuroticism.

2.3.2 Targets

Good targets are proposed to be those whose behaviour gives numerous and informative clues to their personality. In particular, Funder (1995) again notes the relevance of social behaviour—this time in the targets—since those with higher levels of social behaviours in particular exhibit more potential clues about their personality, relative to people who are less active (e.g. Borkenau and Liebler, 1992).

Additionally, people who are high self-monitors (Snyder, 1974, 1987), and adjust their behaviour to changes in the social environment, are predicted to be harder to judge accurately than low self-monitors, who are supposed to act consistently across different situations. Indeed, this is regarded as similar to the difficulty found in rating individuals with dishonest or socially undesirable behaviours who are likely to try and conceal them, leading to difficulty in accurate judgement on the basis of their overt social behaviour (Funder, 1995).

Although Furnham (1990) notes that self-monitoring has shown impressive reliability and validity, he also observes that it is multidimensional (Furnham and Capon, 1983) and correlates with Extraversion and Neuroticism (Gabrenya and Arkin, 1980; Luu et al., 2000, which is not discussed by Funder, 1995). This therefore complicates matters with regard to Extraversion, however, we propose that additional availability of cues in the behaviour of Extraverts outweighs the difficulties presented by self-monitoring; in the case of Neuroticism, we would expect lower Neurotics to make better targets due to the reduced effects of self-monitoring.

2.3.3 Traits

Distinguishing between the different personality dimensions has shown that, even in judgements by close acquaintances, much greater agreement is found for ratings of Extraversion than for Neuroticism in both the EPQ (Gomà-i-Freixanet, 1997), and in the five factor models of personality (McCrae and Costa, 1987). For the EPQ, we find additionally that Psychoticism displays the lowest agreement in judgements and that additionally agreement for Lie-scale ratings is slightly higher than for Neuroticism (Gomà-i-Freixanet, 1997); for the other traits of the five-factor model, generally Openness shows similar levels of agreement to Extraversion (in the case of self-reports and mean peer reports, it is actually higher), whereas Agreeableness shows low Agreement similar to that of Neuroticism, with Conscientiousness located somewhere between these groups (McCrae and Costa, 1987). Additionally, self-ratings were shown to be more informative in predicting behaviour for Extraversion than for Neuroticism (Spain et al., 2000).

These differences in agreement appear to demonstrate that different properties of the personality traits have implications for their judgement and perception. In response to such findings, Funder has proposed that good traits are highly 'visible' (easily observable), and demonstrate low 'evaluativeness' (are not related to judgements of desirability or undesirability). Using Extraversion and Neuroticism as examples to which lay perceivers of personality show sensitivity, he notes that Extraversion is highly visible and revealed by 'frequent positive social interaction' (Funder and Dobroth, 1987; Funder and Colvin, 1988; Paunonen, 1989), but relatively low in evaluativeness (but cf. Scherer, 1979 who notes the desirability of Extraversion, at least in American culture, and Eysenck et al., 1993, who note higher E scores for Canadian males compared to English counterparts). However, Neuroticism is lower in visibility (characterised by, e.g., internal worrying thoughts or feelings), and is regarded as more 'evaluative', i.e., affectively charged, or related to desirability. It may thus lead to: the concealment of undesirable behaviour from observers; a distortion of self-perception, leading to lower target-judge agreement; or a greater reluctance to pass judgement on such behaviours, leading to reduced inter-judge agreement. When less evaluative measures of Neuroticism are used, agreement increases (John and Robbins, 1993).

2.3.4 Information

The amount and relevance of target information available to the judges influences their agreement. Close acquaintances agree better with each other and with the target, than do relative strangers (Funder and Colvin, 1988; Paunonen, 1989; Paulhus and Bruce, 1992), although both predict target behaviour equally well, when they know the target in a relevant context (Colvin and Funder, 1991). Indeed, certain types of information can be more or less diagnostic of personality: for example, a person talking about their thoughts and feelings, rather than about hobbies, leads to more accurate judgement of their personality (Andersen, 1984), with similar behaviour in unstructured situations being most informative (Funder and Colvin, 1991); conversely, reduced accuracy resulted from judgements based on highly scripted tasks, and one-to-one interactions with judges reduced agreement, even when the target believed they had conveyed a similar impression in all cases (DePaulo et al., 1987).

Judgements by close acquaintances (especially when taken as a composite measure) generally also better predict target behaviour than judgements by other peers (Kolar et al., 1996). At the other extreme, studies have investigated personality perception of strangers on the basis of minimal cues, at so-called *zero-acquaintance*. Here there appears to be interaction between the available information and the visibility of the trait being judged. This has been demonstrated using solely linguistic or visual cues; From exposure to transcribed interactions, self-other agreement was shown for ratings of Extraversion and Introversion (Gifford and Hine, 1994). Alternatively, Albright et al. (1988) found that, on the basis of physical appearance, Extraversion and Conscientiousness, but not emotional stability (Neuroticism), Agreeableness, or Culture (Openness), could be reliably rated. However, the judgements of Extraversion appeared to be mediated—or influenced—by judgements of the physical attractiveness of the target. Judgements made at zero-acquaintance appear readily influenced by stereotypes, which judges may attend to in the absence of readily available cues. For example, perceptions of target nationality or gender (Gallois and Callan, 1986) may influence accuracy, in addition to ideas about personality (McCrae and Costa, 1987).

2.3.5 Perception Hypotheses

On the basis of previous perception studies and the properties of the traits themselves, we present the following hypotheses (note that we also refer to the perception of the traits in a computer-mediated environment; this is discussed next, Section, 2.4.3):

Extraversion This trait will be the most easily perceived due to its high visibility and low evaluativeness. We therefore expect it to show the highest levels of inter-judge and target-judge agreement, even in CMC at zero-acquaintance.

Neuroticism We expect that agreement will be lowest for Neuroticism, due to its high evaluativeness and low visibility, which we predict will be most affected by the lack of information available in the CMC and zero-acquaintance conditions.

Psychoticism Since we propose that Psychoticism is visible, but evaluative, we expect agreement to be higher than for Neuroticism, but lower than for Extraversion. We also expect that the conditions will only have moderate lowering effect upon agreement.

2.4 Implications of Computer-mediated Environment

In this section we overview the issues surrounding the computer-mediated environment which have implications for this thesis. We first look at studies which have examined experimentation and methodological issues over the computer and internet. We then cover studies which have focussed on language use in the CMC environment.

2.4.1 Computer-Mediated Experimentation

Increasing familiarity and use of the internet and computers in general has resulted in increasing possibilities of using these resources for psychological experimentation (Schumacher and Morahan-Martin, 2001; Epstein and Klinkenberg, 2001).[5] The bene-

[5]For example, see any of the on-line psychology labs which have appeared: Web Experimental Psychology Lab, Ulf-Dieter Reips, University of Zurich, `http://www.psychologie.unizh.ch/genpsy/Ulf/Lab/WebExpPsyLab.html`; Language Experiments, Christoph Scheepers, University of Glasgow and Martin Corley, University of Edinburgh, `http://www.hcrc.ed.ac.uk/web_exp` (Keller et al.,

fits of computerised experimentation are that the electronic form of the materials allow easy modification, and the results in electronic form also allow easy data processing. In the administration of experiments over the internet, we also have additional benefits associated with this technology. For example: the access to a greater number of more diverse potential participants, than can be found amongst psychology undergraduates; and also flexibility of participation via the internet is not limited by physical access or time (see Reips, 2000 and Epstein and Klinkenberg, 2001, for further details and discussion).

Conversely, many of these benefits also harbour potential pitfalls, especially in the case of web experiments: Unlimited access denies the experimenter control over who—or even how many times an individual—participates; similarly, there is little control over the conditions of participation—crowded internet cafe, or quiet library—or different presentation of the materials due to technological variation; there are also issues of self-selected and unrepresentative samples (Reips, 2000; Epstein et al., 2001; Zelenski et al., 2003). For example, the internet is seen as male-dominated (Sussman and Tyson, 2000), with females showing greater discomfort and lower levels of competence (Schumacher and Morahan-Martin, 2001); additionally males and females use the internet for different purposes (Hamburger and Ben-Artzi, 2000), and although this apparently is also related to personality (Swickert et al., 2002), in a wide-ranging UK sample, this was not found to be the case when gender and age were controlled for (Hills and Argyle, 2003).

In light of these possible difficulties, comparison of computerised with traditional 'pencil and paper' questionnaires have shown that they give equivalent results (Knapp and Kirk, 2003), with similar comparison of web and lab experiments supporting the validity of this new method in a wide variety of studies. For example, replicability has been shown in psycholinguistic experiments (Keller and Alexopoulou, 2001; Corley and Scheepers, 2002), and even in manipulations requiring millisecond accuracy (McGraw et al., 2000), although Krantz and Dalal (2000) note that some methods transfer better than others.

Questionnaires measuring personality and individual differences are regarded as

2002); PsychExps, Ken McGraw, University of Mississippi, `http://psychexps.olemiss.edu` (McGraw et al., 2000).

being well suited for administration via the internet, since they are easily encoded in HTML (the language used to write WWW pages; Hewson et al., 1996), and the anonymity offered by the the internet potentially leads to greater disclosure and honesty (Buchanan and Smith, 1999; Buchanan, 2000). Additionally, recent studies have shown the validity and comparability of 'pencil and paper' with computerised and web administered personality questionnaires in normal populations (Buchanan, 2001; Fox and Schwartz, 2002), and also in personality disorder groups (Pinsoneault, 1996; Weber et al., 2003).

2.4.2 CMC and language

Computer-mediated communication, and more specifically e-mail, is considered to be a form of communication located between the domains of speech and writing: it shares properties of both media (Bälter, 1998; Baron, 2001). For example, it is a written form with interlocutors physically separated, it is durable and often utilises complex syntactic constructions; however, e-mail is often unedited, makes extensive use of first and second pronouns, present tense and contractions, and is informal. Additionally it has also developed its own stylistic features (Baron, 1998). Colley and Todd (2002) refer to stylistic "emailisms" described by Petrie[6] which are common to e-mail, but rare in other forms of writing. These include trailing dots, capitalisation, excessive use of exclamation marks and question marks; however use of 'emoticons' was found to be rare. Study of a bulletin board corpus (e-mails posted to the web) using a multi-dimensional analysis similar to that of Biber (1993), found that the language genre was most like that of 'public interviews and letters, personal as well as professional' (Collot and Belmore, 1996).

Computer-mediated communication provides impoverished cues, and is less rich than face-to-face communication (Panteli, 2002), therefore information has to be communicated using alternative means. Werry (1996) notes that in internet relay chat (interactive electronic communication) innovative linguistic strategies are adopted to represent the intonational or paralinguistic features of face-to-face discourse, with this

[6]The study which Colley and Todd (2002) refer to was published on-line, and downloaded by them in 2000, however the link which they publish no longer works.

finding mirrored in coordination devices employed in task-based interaction in a CMC environment (Hancock and Dunham, 2001b).

Although CMC lacks cues compared to face-to-face interaction, it still provides rich information about the communicator, for example Panteli (2002) found that the construction of text-based messages conveyed the social cues indicating status differences in organisations. Additionally, several studies have shown gender to be communicated in a CMC environment: in mailing lists, messages written by females used more interactional features, and communicated more information, whereas males were more critical (Herring, 1996); in e-mails to friends, females preferred social and domestic topics, whereas males preferred impersonal and external topics (Colley and Todd, 2002); interlocutors and judges were consistently able to identify author gender from e-mails, with female messages found to be characterised by more modal auxiliaries, intensifying adverbs, mention of emotions, sharing of personal information, questions, compliments, apologies, and self-derogatory remarks. Conversely males were found to give more opinions and use more insults (Thomson and Murachver, 2001). Additionally style matching was found for interlocutors of the minority gender style when communicating with those belonging to the norm group, regardless of their own gender (Herring, 1996; Thomson et al., 2001).

Additional properties of the CMC environment are that it enables and encourages increased communication, for example, in computer-mediated task-based group meetings, introverts provided more original solutions than in the face-to-face meetings (although in the latter environment they provided more comments), in each case extraverts showed greater participation in both environments (Yellen et al., 1995). This behaviour is mirrored with second language learners, with students who are less forthcoming in class being more inclined to contact their teacher by e-mail (Bloch, 2002).

2.4.3 CMC and personality judgement

When the availability of information for personality judgements is reduced, we find that accuracy is also reduced. For example, judges who are better acquainted with the target generally provide more accurate personality ratings, as discussed above (section 2.3.4). Whether or not subject and judge have prior knowledge of each other,

technology also has an impact on what information is available in a communicative situation. Zero-acquaintance judgements are particularly vulnerable to technological artifacts. For example, interviews conducted over the telephone were found to result in reduced self-interviewer and peer-interviewer agreement than face-to-face interviews (Blackman, 2002a). Furthermore in text-based computer-mediated environment (CMC) judgements of gender, accuracy was reduced by expectations of linguistic stereotypes for the male and female writers (Savicki et al., 1999). For judgements of personality in CMC (following one-on-one interactions in an internet chat room), consensus was found between judges for a target's Extraversion, Agreeableness, and Openness, but target-judge agreement was only found for Extraversion and Openness (Markey and Wells, 2002).

Impressions of personality formed following task-oriented synchronous computer-mediated communication found that they were less detailed but more intense compared with those from face-to-face communication. Specifically, in the CMC environment, judges seemed less able to rate their partners for Extraversion, Neuroticism, and Agreeableness, relative to face-to-face interaction. Across both environments, Conscientiousness, Agreeableness, and Extraversion were the most rateable (Hancock and Dunham, 2001a).

2.5 Linguistic Analysis Methods

This section provides the background to the analysis techniques which we will use in this thesis. First we introduce corpus linguistics and associated methodology and annotation methods, then we cover the top-down and bottom-up approaches in more detail.

2.5.1 Introduction to Corpus Linguistics

Corpus linguistics is a methodology—rather than a branch of linguistics—which can be applied to the study of any language phenomenon (Rayson, 2003), and is based on the study of samples of 'real life' language use, namely a *corpus* (McEnery and Wilson, 1996). A corpus is simply a collection of texts (usually in electronic form) which is

used for the study of language, and the choice of texts included in the corpus is largely determined by the research question (Sinclair, 1991). For example, the London-Oslo-Bergen Corpus is used by Biber (1995) for its representativeness of written English in general, and so could be termed a Reference Corpus (also termed a Sample Corpus; Sinclair, 1991), whereas an example of a more specific collection would be the International Corpus of Learner English, as used by Aarts and Granger (1998). Both of these form corpora resources, however for some studies where such data may not already exist, the researcher may choose to create their own Specialised Corpus (Hunston, 2002), for example, collections of doctor-patient interactions (Thomas and Wilson, 1996) or writings about 'thoughts and feelings' (Pennebaker and King, 1999), even though this may not overtly be termed 'a corpus'.

2.5.2 Corpus Linguistics Methodology

In the analysis of various different corpus types, Biber et al. (1998) identifies a methodology common to corpus linguistic studies. This is summarised by Rayson (2003, p. 13) as:

1. Question: A research question or model is devised
2. Build: Corpus design and compilation
3. Annotate: Computational analysis of the corpus
4. Retrieve: Quantitative and qualitative analyses of the corpus
5. Interpret: Manual interpretation of the results or confirmation of the accuracy of the model

Within this common methodology, Rayson (2003) observes variation in approaches, especially with regard to the research question, and how this determines the study: specifically, from a theory-driven perspective, he notes that the focus of study can be specific linguistic variables and their (different) behaviour or function in language generally (e.g., Greenbaum et al., 1996; Johansson and Oksefjell, 1996; Butler, 2001), or language styles or genres and how these are realised through the different use of linguistic variables (e.g., Biber, 1986, 1988, 1993, 1995). Furthermore, Rayson (2003) proposes that in data-driven studies, there is an iterative process whereby the results of the annotation and retrieval stage can inform the question, which leads to further

annotation and retrieval. Rayson views the data-driven method as one not used to investigate a particular directional hypothesis. However, in this thesis, we take a more moderate approach, which allows the initial motivation of a theoretical research question (namely, how is personality projected through language), but adopt a data-driven approach, whereby this is modified and the process of annotation and retrieval are subject to iteration based on results of previous analyses. We also note that this approach is different again from that of Biber who sequentially uses the initial results of textual styles to then focus on individual linguistic behaviours, without iteration.

2.5.3 Annotation

2.5.3.1 Basic corpus processing and annotation

At the most basic level, we can analyse the corpus at the word level—that is without explicit annotation (Hunston, 2002)—however decisions still need to be made in the way the corpus is presented for analysis: For example, whether non-standard spellings, or spelling 'mistakes' are corrected, or whether punctuation is included in the analysis. Indeed, some of these features which are often 'cleaned up' and standardised, may well be characteristic features in themselves (especially in a CMC environment, cf. Section 2.4.2). Furthermore, there are different definitions of 'words'—whether they are simply delineated by white space (sometimes referred to as a 'word-form')—or whether a more sophisticated account which considers the base, or uninflected forms ('lemmas') is required (Sinclair, 1991).

Although not considered a primary method of 'annotation' (Leech, 1997b), lemmatisation does fit with the category-based methodology of annotation proposed by Hunston (2002), since word-forms are categorised by lemmas, and that this 'adds value to a corpus, making it easier to retrieve information and increasing the range of investigations that can be done on the corpus'. Therefore, here we take a broad view of annotation, which we regard as forming a continuum: lemmatisation is at the mild end of the scale, adding the least additional information to the corpus, and using the lowest-level (i.e., lemma or base-word) categories; below we describe other, higher-level, methods of annotation.

2.5.3.2 Part-of-Speech tagging

Tagging a corpus usually refers to adding grammatical category ('part-of-speech'; POS) information to the words of the corpus (for a review of the complicated issues and processes involved, see Leech, 1997a). Tagging is different to parsing, since grammatical categories are assigned to words individually; no analysis is made at, for example, the sentence or clause level (note, however, that contextual information of tag sequences is often used to determine individual POS tag probabilities; Hunston, 2002). The value of POS information is that it allows both the more specific study of word behaviour, or more general patterns to be found in a corpus, like the different concordances of a word (lemma) according to its grammatical category (Sinclair, 1991), or syntactic distribution across register (Biber et al., 1998). Additionally, the combination of parts of speech can also be used to indicate syntactic patterns or constructions (Aarts and Granger, 1998), or indicate an author's style (Milic, 1966; Koppel et al., 2003a), or classify texts according to author gender (Koppel et al., 2003b).

There are many different approaches to implementing part-of-speech tagging, generally these are regarded as having at least 95 percent accuracy; however this can vary according to individual features of the corpus and the tagger (Manning and Schütze, 1999). Additionally the labels—or tags—given to grammatical categories vary according to the tagger, for example, the CLAWS tagger uses the CLAWS tagset (Rayson, 2003), whereas the MXPOST tagger uses the PENN tagset (Ratnaparkhi, 1996).

2.5.3.3 Semantic Annotiation

An alternative approach to grammatical tagging is to annotate words according to their meaning: this is known as semantic annotation. One system which implements this is the UCREL Semantic Analysis System (USAS; Wilson and Rayson, 1993; Garside and Rayson, 1997; Piao et al., 2003) which has been applied to the analysis of doctor-patient interactions (Thomas and Wilson, 1996), as well as early modern English texts (Archer et al., 2003).

The USAS semantic tags encode meaning by assigning a letter representing the general discourse field, and then specify subdivisions of this using numbers, with additionally strength optionally indicated by pluses or minuses. From the number of

possible tags which could be assigned, disambiguation is carried out using a number of techniques, for example, POS information, likelihood of tags, domain of discourse, contextual rules, or local probabilistic disambiguation (Rayson, 2003). Although the broadness of categories does not allow for finesse in the categorisation of meanings, Hunston (2002) acknowledges that such methods allow the automatic annotation of large amounts of data which would be too difficult and time-consuming to do consistently by hand (cf. Thomas and Wilson, 1996).

2.5.3.4 Analysis of Content

An additional method of annotating corpora is to perform content analysis. Although this is similar to that of semantic tagging, described above, there are a number of differences: Firstly, content analysis generally gives a score for certain concepts for the text or corpus as a whole, rather than applying annotations to the corpus, and provides a new way of describing texts (Kilgarriff, 2001). Although there are many different analysis systems (see Smith, 1992; Pennebaker et al., 2003, for an overview), here we describe one method in particular, which we have already mentioned, and will refer to in more detail later in the thesis.

The Linguistic Inquiry and Word Count (LIWC; Pennebaker and Francis, 1999)[7] text analysis program was originally designed to examine the relationship between disclosure and language use features with health and well-being (Pennebaker et al., 1997; Pennebaker, 1997; Graybeal et al., 2002; cf. Oxman et al., 1988 who uses alternative content analysis programs for a similar purpose). However, this method has since been applied to investigate a variety of linguistic behaviours in many different genres, including suicidal and non-suicidal poets (Stirman and Pennebaker, 2001), deception (Newman et al., 2003), and gender (Mehl and Pennebaker, 2003). Here we make particular note of this method since it has been used to analyse linguistic features related to individual differences, including personality (Pennebaker and King, 1999), and we use it for comparison purposes in our analysis.

LIWC essentially works by counting the number of words in a text which belong

[7]Note that a more recent version of the program has been released (LIWC2001; Pennebaker et al., 2001), however since the analysis in this thesis was undertaken using the original version of LIWC, we describe this version here.

to its pre-defined dictionaries, and then outputting the frequency of words occuring in each of these dictionary categories as a percentage of the text as a whole (with the exception of categories 'word count' and 'words per sentence' which are expressed as the raw count). The dictionary categories are grouped under four main dimensions: The first, Linguistic Dimensions, measures features such as 'word count' and 'unique words' which are values calculated directly from the text, in addition to basic linguistic features which are included in pre-defined dictionaries, such as various different pronoun categories (e.g., 'first person singular', 'total second person'), 'negations', and 'numbers'. The other dimensions of Psychological Processes, Relativity, and Personal Concerns are further sub-divided into groups of dictionaries, and the dictionaries themselves: 'Affective and Emotional Processes' (e.g., 'positive feelings', 'anxiety and fear'), 'Cognitive Processes' (e.g., 'causation', 'insight'), 'Sensory and Perceptual Processes' (e.g., 'seeing', 'hearing'), 'Social Processes' (e.g., 'communication', 'friends'); 'Time' ('past tense verb'), 'Space' (e.g., 'inclusive'), 'Motion'; 'Occupation' (e.g., 'school'), 'Leisure Activity' (e.g., 'sports', 'music'), 'Money and Financial Issues', 'Metaphysical Issues' (e.g., 'death and dying'), 'Physical States and Functions' (e.g., 'body states, symptoms', 'sex and sexuality').[8] In contrast, these latter dictionary categories are largely concerned with psychological and traditional content analysis concepts, with these derived from theoretical sources. The end result is that LIWC contains around 70 dictionary categories, between them containing over 2,000 words. Furthermore, Pennebaker and colleagues note that in contrast to other text analysis programs, both the constituent dictionaries and the LIWC analysis have been independently rated and validated by judges (Pennebaker and Francis, 1999; Pennebaker and King, 1999).

2.5.3.5 Analysis of Psycholinguistic properties

Here we discuss the MRC Psycholinguistic Database in relation to the content analysis annotation of corpora. Although the Psycholinguistic Database is machine readable resource (Coltheart, 1981; Wilson, 1987), in this thesis we implement it in a novel

[8] Additionally there is also an experimental dimension consisting of 'swear words', 'non fluencies', and 'fillers'.

content analysis technique which measures the psycholinguistic properties of texts. Its original purpose is described thus (Wilson, 1987, p. 1):

> 'It is designed to be of use to psycholinguists in selecting stimulus materials for testing; for use by researchers in Artificial Intelligence as a source of information required for natural language processing and cognitive simulation, and for computer scientists who wish to use word lists and syntactic information in the design of text processors.'

Our content analysis technique which implements the MRC Psycholinguistic Database—like the LIWC—uses a dictionary lookup technique. However, in addition to the different focus of analysis which the two approaches provide, these approaches differ in three important ways: Firstly LIWC relies upon a pre-defined dictionary based on human judgements, whereas the MRC Psycholinguistic Database is built upon empirically derived data collected from several different psycholinguistic studies; Secondly, the MRC database also includes part-of-speech information (for example, noun, verb, etc.), meaning that words can be disambiguated according to their syntactic function, allowing for more accurate categorisation; Thirdly, presumably as a result of their method of derivation, the resources differ in size, and therefore linguistic coverage. The MRC Psycholinguistic Database contains around 150,000 words, of which it has psycholinguistic information for about 40,000, whereas the LIWC dictionary contains just over 2,000 words or stems.

It is apparent that the MRC Psycholinguistic Database therefore provides a useful additional resource to that of LIWC analysis, allowing the calculation of psycholinguistic properties, such as abstractness/concreteness, frequency, and imageability, across a much wider coverage of words from a text. This technique therefore gives a more generalisable picture of the properties of a particular personality text, and allows greater flexibility in applying the results, than the relatively restrictive word count method employed by LIWC.

2.5.4 Top-down Methods

In their study of personality and its relationship to language, Pennebaker and King (1999) apply the LIWC analysis to written assignments (e.g., 'thoughts and feelings'

and 'coming to college') from around 800 undergraduate students. On the basis of previous reliability results, 15 of the 72 LIWC features were included in factor analysis, which gave a four-factor solution. These factors are: Immediacy, Making Distinctions, The Social Past, and Rationalisation.[9] The factors were then used to calculate factor scores for each text, with these then correlated with author personality scores from the five-factor model.

The approach adopted by Pennebaker and King (1999) is very similar to the factor analysis methodology used by Biber to determine the dimensions of language with particular reference to genre (e.g., Biber, 1986, 1988, 1993, 1995). With reference to the latter technique, the main steps are described (Rayson 2003, p. 55; cf. Kilgarriff 2001) as follows:

1. review previous research to identify important linguistic features
2. collect texts
3. count occurrences of features in the texts
4. perform factor analysis: clustering of features into groups of features that co-occur with a high frequency in particular texts
5. interpret factors as dimensions
6. for each factor, compute a factor score for each text
7. compute an average factor score for texts in each genre
8. interpret the textual dimensions in the light of relations among genres given by the factor scores

This approach has been applied to the study of systematic variation associated with genres of spoken and written language, for example, in English (Biber, 1986, 1988), and other languages (Biber, 1995).

Sixty-seven linguistic features were identified as potentially important for the study of English (Biber, 1995), and these were grouped into the following 16 grammatical and functional categories: tense and aspect markers; place and time adverbials; pronouns and pro-verbs; questions; nominal forms; passives; stative forms; subordination features; prepositional phrases, adjectives, and adverbs; lexical specificity; lexical classes; modals; specialised verb classes; reduced forms and discontinuous structures;

[9]For further details of the LIWC features which loaded on each of these factors, or of their relationship to personality dimensions, see the discussion of personality language (Section 2.2.2.4).

co-ordination; negation. The dimensions (and their names) resulting from the different analyses have varied, however Biber (1995; cf. Biber, 1993) gives the following six dimensions derived from the written data of the London-Oslo-Bergen, and the spoken data of the London-Lund corpora: Involved versus Informational Production, Narrative versus Non-narrative Discourse, Situation-dependent versus Elaborated Reference, Overt Expression of Argumentation, Non-abstract versus Abstract Style, and On-line Informational Elaboration Marking Stance; additionally, a further, seventh dimension is proposed, namely Academic Hedging.

The similarity in the approaches of Pennebaker and King (1999) and Biber (1995) are apparent. Referring to Rayson's summary of the methodology (above), we note that the factor analysis and interpretation into dimensions, albeit relating them differently to personality or genre, (steps 4–8) is similar. However, the main differences between the two approaches is in the first three stages of data collection; that is, the choice of feature and determination of their occurrence. In terms of data collection: Biber uses pre-existing corpora, however this is clearly not possible for Pennebaker and King since they use a particular genre of written text, and require the authors to complete individual difference questionnaires; even given these constraints they still manage to accumulate a relatively large data set.

The greatest difference between the studies appears to be in terms of the selection of features and identifying their occurrence: Biber approaches this from a linguistic perspective, based on analysis of spoken and written texts, functional studies of linguistic features, and descriptive grammars; In contrast, Pennebaker and King's mainly psychological approach uses LIWC, and although it does analyse some linguistic features, it is informed mainly by the psychological content analysis literature, emotional scale ratings, dictionaries and a thesaurus. Additionally, the differences of approach at the feature selection level have implications for the counting of these features: Biber generally uses categories which are well recognised and well defined linguistic categories, and is therefore able to draw upon research from natural language processing (NLP), and corpora resources. As a result, Biber's feature counting program which is used to tag the input texts is relatively sophisticated. It uses a large scale dictionary of over 50,000 words derived from the Brown corpus (Kucera and Francis, 1967),

and a number of context-dependent disambiguating algorithms; In the case of LIWC, since this contains words with psychologically derived properties, this relied upon the manual selection and rating of these features. This resulted in the relatively modest predefined dictionaries which together total around 2,000 words, with feature counting (rather than tagging) using a 'pattern-matching' technique which ignores contextual information. Although they do not regard this as affecting LIWC's validity (Berry et al., 1997), they have since addressed this to some extent in LIWC2001 (Pennebaker et al., 2001).

In summary of the two approaches, the main differences between Pennebaker and King's and Biber's studies result from their motivation: although both seek to determine informative linguistic dimensions, they start from different disciplines (psychology versus linguistics), and seek to relate these dimensions to different variables (author individual differences versus written or spoken genre). This results in different techniques and resources being available, which favours Biber in terms of available corpora, and linguistic feature analysis programs and dictionaries; Pennebaker and King had to build their own collection of individual differences texts, and implemented their program without the use of large-scale dictionaries, or semantic or syntactic contextual disambiguation techniques available from statistical NLP. The two approaches were, however able to apply the factor analysis methodology and derive linguistic dimensions for their selected features and corpora.

2.5.5 Bottom-up Methods

Bottom-up, or data-driven, methods are characterised by reliance upon the data to inform the theory, rather than to impose the theory (by way of selecting specific features, or combinations of features for the analysis) upon the data. As in any experimentation, the researcher has to make descisions which may potentially have repercussions for the eventual results; however these are kept to a minimum, and are relatively transparent compared with those of the top-down approach. This therefore means that the methodology of such studies is more easily replicated.

One of the most fundamental analyses of a text is to calculate frequency profiles for its constituent features, the simplest of which are the words themselves. The in-

terpretation of these words or features can be aided by additional simple statistical information such as the percentage of text or vocabulary (Sinclair, 1991). Below, however we describe more sophisticated methods of corpus analysis.

2.5.5.1 Keyword Analysis

Viewed in isolation, raw frequency counts of features in a corpus can inform us about which features are most common. However, this does not tell us whether the frequency pattern is expected, or conversely is unusual in any way. To get a better idea of whether these features are usual, or in some way characteristic of the corpus under investigation (termed the Research Corpus), we can compare this to another corpus, typically one which is considered representative of language in general (Reference Corpus). This reference corpus can either be larger than the research corpus, for example, Scott (2001) compares newspaper editorials with a larger sample of newspaper text, or of a similar size, for example, Aarts and Granger (1998) compare non-native English speaker texts to a reference corpus of texts written by native English speakers.

Therefore, with the benefit of a reference corpus, we are able to not just identify a feature as having a high frequency, but as having an *unusual* frequency. These features are special to our research corpus, and are characteristic of it; in some way encapsulating its essence (Scott, 1997). In their most basic form, these features are words, and Scott (1997) refers to these characteristic words as 'key words'. Although words are the most basic feature of a corpus, anything which can be encoded, and counted in a corpus can be a 'feature' (punctuation, POS tag, etc.). For example, in a comparison of native and non-native speakers of English, Aarts and Rayson (1998) compare words and also major grammatical tag categories, such as verbs, nouns and adjectives in order to study lexical verbs in more detail, whereas Argamon et al. (2003a) use word and POS tags to characterise the genre and gender of authors, and Milic (1966) finds particular grammatical categories useful in authorship attribution. Additionally, having derived a number of words which are characteristic of—or *overused* by—our research corpus, we may then want to identify which of these are in some way particularly characteristic of this corpus. One way of doing this would be to compare the relative-frequencies (i.e., the percentage of a feature's occurrence in its corpus) of

these characteristic features for the two corpora, by for example, calculating the ratio of these relative-frequencies (e.g., Damerau, 1993, who uses it to identify characteristic phrases of specialist texts). Rayson (2003) observes that although the relative-frequency ratio takes into consideration the differences is size of the two corpora, it is not sensitive to the differences between the raw frequencies of the features. Additionally, the relative-frequency ratio does not give a statistical significance which would allow us to identify how likely the difference in feature distributions between the corpora is to have occurred by chance. And although Manning and Schütze (1999) note it does not fit well into the hypothesis testing paradigm, they suggest that relative-frequency ratio can be interpreted as a likelihood statistic.

There has been much debate about the relative virtue of various statistical tests in the measurement of relative difference in word or feature occurrence across corpora, with some of these being more suitable than others depending upon the situation. For example, X^2 (chi-squared) is widely noted as a possible test (Butler, 1985; Woods et al., 1986; Oakes, 1998), however there are limitations for its use with smaller corpora, with Dunning (1993) proposing G^2 (Log-likelihood) as an alternative. Here we use the notation X^2 and G^2 to refer to the chi-square and log-likelihood tests respectively, with χ^2 reserved for the chi-square distribution against which we can compare both the chi-square and log-likelihood tests. In his overview of a range of statistical tests, Kilgarriff (2001) finds the Mann-Whitney test to be most suitable for measuring the significance of different word frequencies; however this test is not suitable for low frequency features, and the corpora to be compared must be the same size. Rayson (2003) evaluates the general suitability of tests given the often unpredictable nature of data (low frequencies, different corpora sizes) and favours the G^2 test 'in general'. In his empirical evaluation, he demonstrates the suitability of the G^2 test for corpus comparison studies, and although the critical values (indicating significance) are generally regarded as being directly comparable to χ^2 (Dunning, 1993), Rayson recommends adopting a higher critical value of 15.13 for the 0.01% significance level 'if a statistically significant result is required for a particular item' (p. 155). This higher critical value allows the lowering of the Cochran rule (Cochran, 1954) to include expected values of 1 or more, rather than the 5 or more cases usually required by this statistic (Butler, 1985; see also Section 5.3.2 for further discussion in relation to this thesis).

2.5.5.2 Concordance and Collocation

Concordancing is the viewing of a target word—generally one which has been selected for its 'keyness'—in the context of its occurrence. Generally this process involves the displaying of all occurences of the target word and contexts, so that comparisons can be made, and patterns established (Hunston, 2002). In a corpus tagged for part-of-speech, it may also be possible to view target words seperately in their contexts, depending upon their grammatical category, which can aid interpretation (e.g., using Key Word in Context; KWIC, Christ, 1994 or Wmatrix, Rayson, 2001). In some cases, this will allow the comparison of observed linguistic behaviour, with models or theories (e.g., Stubbs, 1995; Johansson and Oksefjell, 1996; Butler, 2001). However, the availability of information derived from concordances may also lead to the further iteration and analysis, in a data-driven approach (Rayson, 2003).

A concept related to concordance is that of collocation: this is the patterning of two or more words together. Sinclair (1991) distinguishes collocation based upon whether the collocate occurs before (*upward*) or after (*downward*; usually less frequently) the target word (node). He describes the systematic variation thus: 'Upward collocation, of course, is the weaker pattern in statistical terms, and the words tend to be grammatical frames, or superordinates. Downward collocation by contrast gives us a semantic analysis of a word' (Sinclair, 1991, p. 116). More generally, collocations which contain function (grammatical) words and content (lexical) words are sometimes referred to as *colligation* (Hunston, 2002).

The study of collocations, or word sequences, has been used to identify domain-specific vocabulary (Damerau, 1993), differences between native and non-native speakers (Milton, 1998), and genre (Stubbs and Barth, 2003). However, collocations are not limited to words alone, it is also possible to apply such contextual occurrence information to grammatical tags: For example, POS n-grams (sequences of n length; but generally 2 or 3) are used to compare sentence-initial features and prepositional patterns in native and non-native speakers (Aarts and Granger, 1998), and in conjunction with other features to extract multiword units (Dias, 2003), categorise texts according to style (Koppel et al., 2003a), and gender (Koppel et al., 2003b; Argamon et al., 2003a) and to determine authorship of texts (Milic, 1966), amongst others (Collins,

1996; Pedersen, 2001). Indeed, the co-occurence of words and their distribution in language has important implications for lexical processing (McDonald, 2000; Levy and Bullinaria, 2001; Monaghan et al., 2004),

The co-occurrence of words within a collocation can also be tested to determine whether a particular pattern is statistically significant. As is the case for comparison of frequencies between corpora—which can be regarded as an analogous problem (Kilgarriff, 2001)—there is also dispute as to the most appropriate test for determining patterning of collocations. As noted above, in smaller samples, the G^2 statistic is regarded as approximating better to the χ^2 distribution, than the X^2 statistic (Dunning, 1993). However, this approximation may be violated in sparse n-gram data (Pedersen, 1996; Evert and Krenn, 2001), and therefore Pedersen et al. (1996) instead propose the use of Fisher's exact test. Daille (1995) in an evaluation of statistical tests to extract domain-specific terminology, found that overall 'the 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' (p. 1).

Furthermore, there are issues in n-gram calculation which are similar to those of corpus annotation, such as which words or features (such as punctuation) should be included or excluded from the analysis. For example, if technical terminology or expressions—and therefore content word n-grams—are the focus of the investigation, a stop-list containing function words can be used (e.g., Damerau, 1993); whereas if more general features of style—which may be expressed through colligations—are the focus, all word categories can be analysed (cf. Stubbs and Barth, 2003). There are also additional options, such as n-gram frequency of occurrence, whether a collocation can take into account words of n-words distance from the node word, and indeed how long the n-gram itself should be. A description of n-gram calculation software, and the options involved in measuring collocation can be found in Banerjee and Pedersen (2003), with an alternative application of n-grams designed for lexicographical use described in Kilgarriff and Tugwell (2001).

2.6 Summary and Presentation of Hypotheses

In this chapter we have overviewed the literature from a variety of sources which will inform this thesis. We presented this according to two main criteria, either topic or methodological relevance. Firstly, we discussed theories of personality and its influence upon language, and issues of personality perception; we then summarised features of computer-mediated interaction and methods of linguistic analysis, with particular reference to methodology.

We now present a summary of hypotheses based upon the literature reviewed in this chapter. These are: firstly hypotheses of personality language use derived from theory; secondly, hypotheses of personality language features use based upon previous findings; and thirdly hypotheses of personality perception based upon previous findings. We present these for each dimension in turn (note that 'voice' features are described in terms of 'realisation' and that e-mail features are included where appropriate, and also that the LIWC hypotheses are presented separately from the conversational features as Lexis). In the subsequent chapters we will compare our findings with the hypotheses for each of these categories.

2.6.1 Extraversion Hypotheses

Theory We expect the language of high Extraverts to reflect their sociability by referring to other people (1), and to express their activity by using words associated with actions (2; cf. Grammatical Hypotheses, increased use of verbs), and by saying more (3). We also expect them to use language which suggests positive affect (4).

Realisation Extravert 'loudness' will be realised in the increased use of capital letters (1) and exclamation marks (2); Worse pronunciation will result in worse spelling and more typographical errors (3).

Fluency Higher speech rate of Extraverts will be realised in longer sentences (1); shorter pauses and less hesitation will result in more ellipses (...)(2) and hyphens (-)(3) being used to separate clauses rather than the full stop.

Grammatical Extravert language will contain more adverbs (1), pronouns (2), and verbs (3)(i.e., more 'implicit'), and have a lower lexical density (TTR) (4); it will contain fewer nouns (5), modifiers (6) and prepositions (7)(less 'explicit'), and be less formal (8).

Conversational Extraverts will write more (1) (cf. Theory hypotheses); initiate more laughter, perhaps indicating this by explicit references ('ha') or by exclamation (2); they will refer to themselves more (3); they will use more terms indicating pleasure and agreement (4), pay more compliments (5); they will use fewer hedges (6) and references to problems (7), and will show less anxiety during the communication (8).

Lexis Extraverts will show a greater use of social (1), inclusive (2), and positive emotion words (3), and use fewer negations (4), tentativity (5), exclusive (6), causation (7), negative emotion words (8), and articles (9); In terms of factors, Extraversion will have a negative relationship with Making Distinctions (10).

Perception Extraversion will be the most easily perceived due to its high visibility and low evaluativeness, we therefore expect it to show the highest levels of inter-judge (1) and target-judge agreement (2), even in CMC at zero-acquaintance.

2.6.2 Neuroticism Hypotheses

Theory The language of high Neurotics, we expect to be highly emotional—particularly expressing negative affect (1), but also positive affect (2; cf. Lexis which predicts fewer positive emotion words)—and this is also revealed through intensified language (e.g., adjectives [3] and adverbs [4]). Since the individual tends to focus more on themselves, we also expect this self-preoccupation to be expressed through an increased reference to self (5).

Conversational High Neurotics will use a lower lexical density (1) (TTR), and show greater anxiety during communication, realised explicitly through references to 'worry' or 'stress' (2).

Lexis High Neurotics will use more first person singular (1) and negative emotion words (2), and fewer positive emotion words (3) and articles (4); Neuroticism will also correlate positively with the Immediacy factor (5).

Perception We expect that agreement (within judges [1], and target-judge [2]) will be lowest for Neuroticism, due to its high evaluativeness and low visibility, which we predict will be most affected by the lack of information available in the CMC and zero-acquaintance conditions.

2.6.3 Psychoticism Hypotheses

Theory We expect highly Psychotic individuals to reflect their lack of sociability and detachedness by making fewer references to themselves (1) or to others (2), and to demonstrate their harshness and toughness by avoiding emotional words (positive [3] and negative [4]). Since they are creative and enjoy unusual things, we predict that they will use more unusual language, realised both in words and constructions (i.e., lexically and syntactically). This we predict would result in for example, the use of less frequent words (5), a higher type-token ratio (6), and use of passive constructions (7).

Conversational We expect high Psychotics to show less anxiety during communication, for example, through fewer explicit references to 'stress' or 'worry' (1).

Lexis Here we predict Psychoticism will show an inverse relationship to Agreeableness and Conscientiousness: based on Agreeableness, we expect fewer first person singular (1) and positive emotion words (2), and more articles (3) and negative emotion words (4), and also a negative correlation with the factor Immediacy (5); on the basis of the findings for Conscientiousness, we expect fewer positive emotion words (=2), and more negations (6), negative emotion (=4), causation (7), exclusive words (8), and discrepancies (9), and a positive relationship with the factor Making Distinctions (10).

Perception Since we propose that Psychoticism is visible, but evaluative, we expect agreement to be higher than for Neuroticism, but lower than for Extraversion

(inter-judge [1], and target-judge [2]). We also expect that the conditions will only have moderate lowering effect upon agreement.

In these hypotheses we are aware of a number of interactions: Firstly, the Extraversion Theory hypothesis predicts a greater use of 'action' words, and the Grammatical hypothesis predicts an increased use of verbs. In this case the parallel hypotheses are complimentary. However, in the second instance for Neuroticism these are contradictorary. Here the theory hypothesis predicts more emotionality expressed through a greater number of negative *and* positive emotion words, whereas the Lexis hypothesis predicts the use of more negative emotion words but *fewer* negative emotion words. In this case, we test both hypotheses to investigate whether the theoretical or LIWC-based hypothesis can be accepted.

In the next chapter we introduce the initial stage of testing our hypotheses, namely collecting personality language data. The rest of the following chapter is then concerned with conducting initial content analyses, and validating our corpus against previous work.

Chapter 3

Personality Corpus Collection and Validation

We now describe the first stages taken towards addressing our hypotheses, namely building and validating our e-mail corpus which forms the basis of the experimental work in the rest of the thesis. We start by discussing the need for data collection and our choice of methodology and experimental design. In order to test the validity of our data, we use content analysis methods and factor analysis to derive comparable factors to a previous study. The rest of the chapter discusses similarities and differences found between the two analyses, and proposes possible explanations. The chapter concludes with a summary and discussion of the appropriateness of multiple-dimension analysis techniques to our data.

3.1 Introduction to Data Collection

In this thesis we investigate the relationship between personality and language. The starting point therefore, is to analyse the language use of individuals for whom we have personality information. As discussed in the previous chapter (Section 2.5.1), many corpora represent a general selection of language, although for example, the British National Corpus has some sociobiographic information available which allows more specific analyses (e.g., Rayson and Hodges, 1997). Despite the wealth of corpora available, as Hunston (2002) notes, it is sometimes necessary to build a specialised corpus in order to address a particular research question. We therefore describe this process in more detail.

3.1.1 Methodological approach

Since we are to collect our own data, it is important that this language is as informative as possible in exhibiting speaker or author personality. In studies which investigate sociolinguistic phenomena, spontaneous spoken language between close family or friends is often regarded as providing the most 'natural' form of data (e.g., Labov, 1972; Chambers and Trudgill, 1980). In contrast, spoken and written language which is used in more formal contexts is generally regarded as more formulaic, and is largely determined by situational conventions (Brown and Yule, 1983). It is therefore unsurprising that a large number of studies investigating personality and language have focussed particularly upon speech (see e.g., Scherer, 1979, for a review).

However, speech is time-consuming to transcribe, and since we do not intend to investigate the paralinguistic cues, much of the possible variation features may be lost in transcription. Additionally, it is often difficult to collect speech data in naturalistic settings and also control for potential audience effects (cf. Trudgill, 1974; Bell, 1984). Computer-mediated communication—and in particular e-mail—is widely regarded as having much of the spontaneity of speech but in a written form (Bälter, 1998; Baron, 1998; Colley and Todd, 2002, see Section 2.4.2 for further discussion). Therefore, we select e-mail as our focus of investigation.

Using 'real life' sent e-mails for experimentation purposes can be difficult due to

ethical concerns, and may account for the study of more easily available, ubiquitous forms (e.g., 'junk' e-mail Orăsan and Krishnamurthy, 2002). Even if access to sent e-mails could be gained, these texts would vary in terms of their topic and recipient; additionally, we would also require personality information from the e-mail's author. Therefore, in order to enable greater control over the data collection, we collected the e-mail data as part of an experimental task in which background and personality information is collected from the participant, along with pre-defined e-mail writing tasks which specify the topic and purpose of the e-mail and also the recipient (past week activities and plans for forthcoming week, written to a good friend). Additionally, by collecting the data using an on-line HTML form, the participant can remain anonymous if desired (since there is no record of the sender's e-mail address, unless they optionally specify it).

3.2 Method

3.2.1 Participants

One-hundred and five current or recently graduated university students participated in this experiment, of which 37 were males, and 68 females. The mean age of subjects was 24.34, with 53 studying (or having studied) at an undergraduate level, and 52 at a postgraduate level. All participants spoke English as their first language.

A sociobiographical questionnaire and Eysenck Personality Questionnaire-Revised (short version) (Eysenck et al., 1985) were administered to give information about the subjects' background and scores on the personality dimensions of Psychoticism (Mean score: 2.90, SD 1.7; Normative score: M = 3.08, F = 2.35), Extraversion (Mean score: 7.91, SD 3.3; Normative score: M = 6.36, F = 7.60), Neuroticism (Mean score: 5.51, SD 3.2; Normative score: M = 4.95, F = 5.90), and Lie Scale (Mean score: 3.48, SD 2.2; Normative score: M = 3.86, F = 2.71). Further information about participants can be found in Tables A.1 and A.2.

3.2.2 Materials

The experiment was conducted on-line via the author's departmental web page, and used an HTML form which subjects filled in and then submitted over the internet.[1]

On the web page a short introductory section outlined what the experiment was about and how long it would take, stated that any responses would be treated confidentially, gave the author's e-mail address, and finally thanked the subject for participating. The rest of the experiment was broken down into two main sections: The first section was concerned with the collection of demographic and personality information; the second collected linguistic samples from the participant. Preliminary versions of the materials were pilotted to evaluate ease of use, time required for the tasks, and to identify bugs in the code.

3.2.2.1 Collection of sociobiographic information

The first part of this section is concerned with the collection of sociobiographical information. Since the study required that participants were native English speakers (and to save unsuitable candidates unnecessarily completing the experiment), the first question required subjects to check a box to confirm that they were 'Native Speakers of English'. For extra emphasis, this question was presented in larger typeface to to the rest and was also emboldened. The rest of the questions comprising the 'Background Information' section were as follows: 'Name' (which was optional); 'Age'; 'Gender'; 'Nationality'; 'Place of Birth'; 'Place where you grew up'; 'Course of Study'; 'Level of Study'; 'Number of Years in University/Higher Education'; 'University currently attended' (to be specified if not Edinburgh); and 'Job and location if graduated' (again optional). To the right of each of this list of background questions a text box was supplied to allow subjects to type in their response, with the exception of 'Gender' and 'Level of Study' for which click buttons were used to allow a choice of response between 'Male/Female' or 'Undergraduate/Postgraduate' respectively.

The second part of this section provides an online version of the EPQ-R short

[1]The experiment can be found at the following URL: `http://www.cogsci.ed.ac.uk/~agill/experiments/ag-expt1.html`. If you experience problems in accessing this page, please contact the author.

scale questionnaire (Eysenck et al., 1985, see Section 2.1.1 for our justification of this choice). As far as possible this replicates the questionnaire as is presented in its paper form, retaining the exact wording of the questions and numbering. However, since the original paper version of the questionnaire asks respondents to circle the 'YES/NO' presented to the right of the question, the online version instead uses click 'radio' buttons to the left of the 'YES' or 'NO'. In the web version, the trail of full stops ('.....') which lead from the question to the 'YES/NO' on the paper version have been omitted, as have the age and gender questions at the top of the questionnaire, since this information had already been collected in the 'Background Information' section. Before the first question the following instructions were presented: 'Please answer *ALL* of the following questions, *clicking* the answer which you feel best describes you. Answer the questions honestly and do not spend too much time thinking about them'. After the last question information regarding the source of the questionnaire was given: 'Questionnaire based on Eysenck, et. al., (1985). For more information, please consult *Personality and Individual Differences*, 6: 21-29'.

3.2.2.2 Collection of linguistic data

The second section consists of the two message writing tasks. Before the tasks themselves were introduced, the following disclaimer was used to reassure participants and to reduce potential discomfort or reservations that they may have about writing about themselves and related events and experiences: 'If during either of the following writing tasks, you worried about writing anything too personal, simply substitute names of people and places as appropriate.' The writing task was then completed using a large scrollable text box which subjects could type into, with the following instructions provided for the first writing task:

> 'Imagine you haven't seen a good friend for quite some time, and in order to keep them up to date with your news you decide to write them an e-mail.
>
> In the message you should write about **what has happened to you, or what you have done in the past week**, trying to remember and write down as much as possible, as quickly as possible.

> Your message should be written in normal English prose (that is, standard sentences, although don't worry if your grammar is not perfect).
>
> Once you have started writing a sentence, you should complete it and not go back to alter or edit it. Also, don't worry too much about spelling, and don't bother addressing it to anyone or signing it. Just write down the main body of the text.
>
> You should spend 10 minutes on this task.'

Since the second writing task was very similar to the first, many of the general instructions were omitted, with the resultant instructions simply giving details of the task and summarising the form that the writing should take:

> 'Again writing to the same friend, you should say **what your plans are for the next week**. As before, you should write in sentences and not go back to alter or edit them, except for spellings if necessary.
>
> You should spend 10 minutes on this task.'

3.2.3 Procedure

3.2.3.1 Recruitment of subjects and internet methodology issues

Potential subjects were contacted using an e-mail written by the experimenter which was then sent to students at the University of Edinburgh's School of Informatics, and to other contacts of the author. The e-mail briefly introduced the author, explained broadly what the experiment was looking at (how writing style varies across individuals), that it could be completed anonymously, and requirements for subjects (that they spoke English as their first language and that they are current students or recent graduates). Furthermore the e-mail requested that recipients forwarded the e-mail onto contacts who may also be interested in participating in the study. This 'word of mouth' approach seemed relatively successful in promoting the legitimacy of the study by using known contacts, whilst at the same time avoided unnecessary use of "spam" (Buchanan, 2000).

3.2.3.2 Presentation of on-line materials

In order to make the online study as transparent as possible, and in order to reduce possible interfering effects, the HTML form which was presented to subjects via the

internet was designed so as to be as simple as possible. All 'normal' text was presented as black default typeface[2] on a white background. Italicised and emboldened typeface were used for emphasis, whilst a larger size of font was used to reflect the hierarchical structure of the webpage, for example headings, new sections, etc. Blue text (as is usually found to mark internet 'links') was used in some special cases, firstly for the author's e-mail address, to indicate that this could be clicked upon to compose a message, and secondly for the text instructing participants to scroll down past the white space to the next section of the experiment. To ensure easy navigation, these white spaces between sections were designed so as not to fill the whole of the web browser screen, which may have led to a 'white out' and the participant losing their place in the experiment. Similarly, the sections themselves (with the exception of the personality questionnaire which was not altered in order to retain its original structure and layout) were designed so that all of the section would be visible at once without needing to scroll through it.[3]

3.2.3.3 Submission and debriefing

The experiment is completed after the subject has supplied this second written sample, and so after the second text box, the subject is thanked and is instructed how to submit their experiment data. Once the subject has pressed the 'submit' button, if they have completed all necessary information, they are directed to a second web page which thanks them for participating in the experiment and gives the URLs of related departmental sites and the experimenter's e-mail address if they require more information about the experiment in which they have just participated. If the subject has not successfully filled in all the required fields in the questionnaire, they are then directed to a page which informs them that they have been unable to submit their data, and advises them of fields which they need to fill in in order to submit the questionnaire.

[2] Although this is generally Adobe Times in Netscape, this may vary across browsers, or by personal specification.

[3] This was found to be the case using a default Adobe Times 14 point typeface on a Netscape browser window sized 825x800 viewed using a Sun Microsystems 21 inch monitor, and appeared to transfer well to other environments.

3.2.4 Preparation of the corpus

Pre-editing of the e-mail texts was kept to a bare minimum in order to retain as much individuality as possible (for example, non-standard words and spellings to imitate sounds). Although such informal linguistic strategies, along with relaxed attitude to typographical errors, are regarded as a feature of e-mail (Baron, 1998; Colley and Todd, 2002), a distinction was made between intentional non-standard spellings for communicative effect and spelling errors. The reason for this was because of the sensitivity of the word list approaches used by some of the dictionary analysis techniques to spellings, leading to an incorrect non-classification of incorrectly spelt words. Therefore, a basic spell-check was carried out (using the standard emacs spell-checker; Stallman, 1994) and then the resulting texts were hand corrected by the author to ensure unintentional spelling errors had been corrected. Copies of texts at each stage of editing were retained for reference, or future analysis if required (Sinclair, 1991).

This slightly more relaxed approach was used in preference to the relatively strict text cleaning-up regime outlined by Pennebaker and Francis (1999). This however does not appear to affect the ability of LIWC to analyse our data: the resulting percentage of dictionary words captured by LIWC for our data (77.88 percent for the Past text and 79.04 percent for the Next text) places them very close to the mean of 78.9 percent reported by Pennebaker and Francis for 'Control Writing'.

3.3 Results

3.3.1 Factor Analysis of LIWC data

The written e-mail data collected from the 105 subjects resulted in two sets of data: The *Past* (texts written about last week) and *Next* (texts written about next week) were taken directly from the subjects' experimental submission. These 210 texts (two from each author) were each retained and analysed separately using LIWC, as recommended by Pennebaker and Francis (1999).

In their study Pennebaker and King (1999) outline a number of considerations used to select which of the original 72 LIWC variables would be retained for factor analysis:

Firstly, only LIWC variables which showed a mean reliability of .60 or greater in previous validation studies by Pennebaker and King would be used. These studies used multiple writing samples produced by a number of subjects in three different writing and topic contexts, and consisted of daily diaries by in-patients at an addiction centre, daily class assignments by summer school students, and 'published abstracts by prominent social psychologists', and which Pennebaker and King claim show that 'word category usage is remarkably stable across time and writing topic' (p. 1300). The second criterion for factor analysis was that the LIWC variable did not substantially overlap with any of the other variables, therefore Prepositions were excluded (due to overlapping with inclusive and exclusive words), as were first-person plural (*we, us*, and *our*)[4] (because of overlap with Social words). Thirdly, categories which did not refer to features or meaning of specific words were excluded (Total Word Count, Words per Sentence and Dictionary Words[5]) since they provide a relatively abstract linguistic description of the text. Current Concern Words (also terms Personal Concerns) were also excluded due to their topic dependency rather than process dependence. The final selection criteria for the factor analysis required the LIWC variables to have mean usage levels of at least 1% per essay.

The 15 LIWC variables to be used by Pennebaker and King for factor analysis were derived for the means of their four texts for: 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. Psychometric information for these variables in the current study can be found in Table 3.1.

In order to ensure comparability with Pennebaker and King's factor analysis, these same 15 LIWC variables were selected from the current data for factor analysis. Since we use the LIWC program and standard dictionaries, we adopt the same three adoption criteria as Pennebaker and King. However, Pennebaker and King (1999)'s fourth selection criterion, that of LIWC variables included for factor analysis having a mean usage of greater than 1 percent, uncovered two variables in the current study which

[4]And presumably other super-ordinate categories of which first-person plural is a hyponym, e.g., Total first person and Total pronouns.

[5]Also Unique Words and Question marks.

Dimension	Example words	*Mean*	*SD*
Words of more than 6 letters	N/A	12.69	2.52
First Person Singular	*I, my, me*	6.51	2.13
Negations	*no, never, not*	1.69	.74
Articles	*a, an, the*	6.17	1.50
Positive Emotions	*happy, pretty, good*	3.10	1.11
Negative Emotions	*hate, worthless*	.99	.65
Causation	*because, effect, hence*	.68	.48
Insight	*think, know, consider*	1.65	.88
Discrepancy	*should, would, could*	2.18	.87
Tentative	*maybe, perhaps, guess*	2.62	1.00
Social Processes	*talk, us, friend*	6.34	2.03
Past tense verbs	*walked, were, had*	4.56	1.28
Present tense	*walk, is, be*	11.12	2.14
Inclusive	*with, and, include*	6.32	1.58
Exclusive	*but, except, without*	3.55	1.27

Note. All means are expressed as percentage of total words within the texts ($n = 105$).

Table 3.1: Psychometric information for Factor Analysis Sample

did not match their inclusion criterion: 'Negative Emotions' and 'Causation' words, which across past and next texts had means of .99% and .68% respectively (across the different text types, they scored means of 1.15% and .79% for past texts and .85% and .57% for next texts). In order to take into account this discrepancy, the factor analysis was carried out twice: Once with all of the 15 LIWC variables; and then again with the remaining 13 LIWC variables excluding 'Negative Emotions' and 'Causation' words. Note that in this analysis, we do not use Cronbach's alpha to measure the consistency of LIWC variables in our past and next writing samples, but rather rely upon the values derived from Pennebaker and King's validation studies. Since we only analyse two relatively short written texts, and given the variety of language, the value of such a measure is debatable. Indeed Pennebaker and King note that the Cronbach alpha coefficients of their four writing samples used for their factor analysis are lower than those for the validation studies, stating that 'these patterns are not surprising, however, given the limited number of writing samples' (p. 1302; across the three conditions of 'Inpatients', 'Summer School', and 'Abstracts', the number of writing samples per author

was 18, 10, and 15 respectively).

Exploratory factor analysis was undertaken using the mean of the two e-mail texts, in the same way as Pennebaker and King (who used the mean of their four writing samples). Diagnostic tests (Bartlett's test of sphericity and Kaiser-Meyer-Olkin, KMO, measurement of sampling adequacy) indicate similar suitability of the current data for a factor model to the data of Pennebaker and King: for the 15 variables KMO = .580, Bartlett's test of sphericity = 333, $p < .001$; for the 13 variables KMO = .600, Bartlett's test of sphericity = 278, $p < .001$; whilst Pennebaker and King report KMO = .633, Bartlett's test of sphericity = 2,831, $p < .001$.

3.3.1.1 Analysis including 15 LIWC variables

Turning first to the 15 LIWC data, examination of the scree plot indicated that a four factor solution would best fit the data, since five factors had eigenvalues above 1. Principal-components analysis extracted four factors, and varimax rotation was used to aid interpretation of the factors. All 15 variables had communalities greater than .37.

Rotated factor loadings are shown in the Table 3.2, and appear broadly comparable to those of Pennebaker and King (1999), which we include in the Appendix (Table C.1). Note that only factor loadings greater than .4 are shown to aid interpretation, rather than .2 (cf. Pennebaker and King, 1999). Dictionary variables loaded on the first factor (eigenvalue = 2.92), like that of Pennebaker and King's Immediacy (eigenvalue = 3.35) included present-tense verbs, fewer longer words, the total number of first-person singular words (*I, me, and my*) (showing a primary rather than secondary loading), and fewer articles (*a, an,* and *the*). Differences from Pennebaker and King included an additional similar secondary loading for fewer articles on factor four, whilst insight words (*understand* and *realise*) (found on Pennebaker and King's fourth factor 'Rationalization') also make an appearance in factor one, whilst discrepancies are absent. In all, Pennebaker and King's first factor 'Immediacy' accounted for 22.4% of the variability, whilst in the present study the first factor explained 19.4% of the variance.

The second factor (eigenvalue = 1.91) included the same variable loadings as Pennebaker and King's 'Making Distinctions' (eigenvalue = 1.47), namely discrepancies (*would, should,* and *could*) (featuring as a primary rather than secondary loading), ex-

Dictionary	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)
Present tense	.788			
Words > 6 letters	−.669			
First-person Sing.	.587			
Insight	.561			
Articles	−.522			−.506
Exclusive		.697		
Negations		.651		
Tentative		.604		
Discrepancies		.553		
Inclusive		−.546	−.465	
Past tense			.732	
Social			.676	
Positive emotion			.569	
Negative emotion				.737
Causation				.707

Note. Only loadings of .40 or above are shown. $N = 105$.

Table 3.2: Rotated Factor Loadings for Exploratory Analysis of 15 LIWC Variables.

clusive words (*but, without,* and *except*), tentative words (*perhaps,* and *maybe*), negations (*no, not,* and *never*), and fewer inclusion words (*and, with*). Additionally, inclusion words also have a secondary negative loading on the third factor. Factor two accounted for 12.8% variance compared to Making Distinction's 10.3% of the variance.

Factor three (eigenvalue = 1.73) like Pennebaker and King's 'The Social Past' (eigenvalue = 1.47) shows high use of past tense verbs and social references, however it also shows a secondary negative loading of inclusion words and an absence of present tense verbs. Whilst as in Pennebaker and King (1999)'s analysis positive emotion words also load on this factor, here the loading has a *positive*, rather than *negative* relationship. This factor accounts for 11.5% variance, compared to 9.8% for The Social Past.

Factor four (eigenvalue = 1.44) compares to 'Rationalization' (eigenvalue = 1.29) in the inclusion of causation words (*because, reason*); however there is also the inclusion of a secondary (negative) loading of articles, whilst insight words are instead loaded on factor one. Although negative emotion words are similarly included in this factor, as with positive emotion words for factor three, it is found that they load inversely compared to the relationship found in Pennebaker and King's analysis: their relationship being positive rather than negative. Variance accounted for by this factor is 9.6% compared to the 8.6% variance for Rationalisation.

3.3.1.2 Analysis including 13 LIWC variables

When the dictionary categories which did not meet Pennebaker and King's 1% mean inclusion criterion (negative emotion words and causation words) are excluded and factor analysis is carried out on the remaining 13 LIWC variables, the scree plot indicates that a three factor solution is appropriate for the data, since four factors had a eigenvalue greater than 1. Again, principal-components analysis was used to extract three factors and varimax rotation enabled interpretation of the factors. With the exception of insight words and tentative words, whose communalities were .26 and .35 respectively, all other variables had communalities greater than .37.

The rotated factor loadings are shown in Table 3.3. Although it is obvious that

Dictionary	Factor 1: Making Distinctions (21.9% variance)	Factor 2: Immediacy (14.5% variance)	Factor 3: The Social Past (11.8% variance)
Exclusive	.697		
Negations	.598		
Discrepancies	.593		
Tentative	.581		
Inclusive	−.561		−.457
Present tense		.812	
Articles		−.710	
First-person Sing.		.622	
Words > 6 letters		−.556	
Insight			
Past tense			.755
Social			.658
Positive emotion			.577

Note. Only loadings of .40 or above are shown. $N = 105$.

Table 3.3: Rotated Factor Loadings for Exploratory Analysis of LIWC Dictionaries using 13 LIWC Variables

items loading on factors one and two have swapped, when this is taken into account the resultant distribution of variables across factors is again very similar to those found by Pennebaker and King.[6] Factor one includes the same dictionary categories as Pennebaker and King's second factor 'Making Distinctions', namely exclusive words, tentative words, negations, and fewer inclusive words. As before discrepancies load primarily onto this factor rather onto Immediacy.

Factor two, like Pennebaker and King's first factor 'Immediacy' includes first-person singular words, fewer articles, fewer longer words and more present tense verbs. However, once again this factor is the primary (and only) loading for present tense verbs, and discrepancy words are omitted.

Again the third factor shows similarity to Pennebaker and King's third factor 'The Social Past', with past-tense verbs and social words both loading onto it. However present-tense verbs are absent and inclusive words show a secondary loading, whilst

[6]In this analysis, the eigenvalues were as follows: Factor 1 = 2.84; Factor 2 = 1.89; Factor 3 = 1.54. The variance which these factors accounted for is 21.9%, 14.5%, and 11.8% respectively.

positive emotion words are again loaded inversely (*positively* rather than *negatively*) onto this factor.

Insight, which in the previous four factor analysis had loaded onto factor one, and in Pennebaker and King's factor model had appeared in their fourth factor 'Rationalization' along with causation words and negative emotion words, failed to load onto any of the other three factors when the positive emotion and causation words were excluded from the analysis.

3.3.2 Correlation of LIWC factors to Personality

Here we compare the correlation of the LIWC factors and their constituent 15 LIWC categories entered into the factor analysis with the personality variables from the current study with the findings of (Pennebaker and King, 1999, reproduced as Table C.2). We only refer to the 15 LIWC categories factor structure (Table 3.4) since this is very similar to the 13 LIWC category analysis (Table 3.5), and also includes results for Negative Emotion and Causation (the correlation of individual categories is the same for both sets of results).

Overall we find that these correlations with personality are similar to those found by Pennebaker and King (1999), however here very few reach significance. For Extraversion there are no significant relationships, however for Neuroticism we find a positive relationship with Inclusive words (part of 'Making Distinctions') (.26; $p =< .001$), which contrasts with that of the previous study ($-.01$), and the 'Social Past' factor here shows a significant negative correlation with Neuroticism ($-.21; p < .05$), although the original only shows a weak positive relationship.

Since Pennebaker and King (1999) use the five-factor model, we have taken Psychoticism to be inversely related to Agreeableness and Conscientiousness. Here we find that Psychoticism shows a stronger relationship to both First-Person Singular (loading on the 'Immediacy' factor) ($-.23; p < .05$) and Negative Emotion (part of 'Rationalisation') ($.20; p < .05$) than is predicted by the previous findings.

LIWC factor	EPQ-R Dimension Psychoticism	Extraversion	Neuroticism
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
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**
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.
$*p < .05$. $**p < .001$, two tailed.

Table 3.4: LIWC Factors and Simple Correlations with EPQ-R Scores using E-mail data and 4 LIWC factor model

	EPQ-R Dimension		
	Psychoticism	Extraversion	Neuroticism
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
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**
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 3.5: LIWC Factors and Simple Correlations with EPQ-R Scores and E-mail data using 3 LIWC factor model.

3.4 Discussion

Here we discuss the results of our analyses. Firstly we cover the results of factor analysis, and then we turn to the correlation of the factors with our measure of personality. In each case we compare our findings with the previous study. Secondly, we turn to an evaluation of the multi-dimensional methods used in this chapter and the previous study; we discuss the appropriateness of these techniques for our data. We follow this with a review of our hypotheses based on the findings of the previous study.

3.4.1 E-mail factor structure

The overall similarity between the factor analyses of this study and that of Pennebaker and King (1999) appears to be good, and on the whole variables appear to load on the same factors across both studies, despite some minor variations. The variations include, for the 15 variable analysis: Present Tense Verbs not loading on factor three 'The Social Past', Inclusive words differentially loading on this factor, Discrepancy not primarily loading on factor one 'Immediacy', Articles loading on 'Rationalization', and Insight loading on factor one instead of factor four. When two of the variables found by Pennebaker and King to load on the fourth factor were excluded from the analysis, this produced the same factor structure for the remaining three factors (even resulting in the previously differently loaded Insight being omitted from the remaining three factors, as in the original study).

This high degree of similarity between the current factor analyses and those undertaken by Pennebaker and King appears to have two main implications: Firstly the factors derived from the LIWC variables appear fairly robust, since they have been largely replicated using the current data; and secondly, we note the similarity—and thus comparability—between the e-mail data used in the current analysis and the written texts used by Pennebaker and King. Such similarity, therefore implies that the e-mail data collected here are representative of the genre of personal written texts as a whole, and is comparable with Collot and Belmore (1996)'s analysis of CMC who found similarity with Biber's dimension of 'public interviews and letters' (Biber, 1993).

However, despite these apparent similarities between analyses and texts, significant discrepancies between the factor analysis findings of the current study and that of Pennebaker and King (1999) exist, namely the inversion of the loadings for positive emotion words and negative emotion words on their respective factors. The most important of these involves the positive emotion words since it fulfilled the 1% mean criterion for inclusion and is therefore to be viewed as a more reliable measure. We propose that this is due to the differences in the topic of the written texts analysed in the two studies: Pennebaker and King appear to overtly tap emotional writing in the topics they assign ('thoughts and feelings'), unlike the current study ('what has happened to you, or what you have done').

3.4.2 Correlation between LIWC measures and personality

Comparing the correlations found in the current study with those of Pennebaker and King (1999), we firstly note that the correlations for both studies are relatively modest. In the present study, although many of the correlations are of an equivalent magnitude, fewer of them reach significance due to the smaller sample size of 105 participants versus the 841 of Pennebaker and King, which results in a smaller corpus.

The main findings to which Pennebaker and King draw attention, we did not find occurred significantly in the current data, namely the relationship between 'Immediacy' and Openness, 'Making Distinctions' and Extraversion, and 'Making Distinctions' and Conscientiousness. However, we did find for Neuroticism a positive relationship with Inclusive words, and a negative relationship with the 'Social Past' factor; for Psychoticism, we find a positive relationship with First-Person Singular and Negative Emotion words.

3.4.3 Top-down analysis techniques

In this chapter we have broadly replicated the factors derived in the previous study, although some differences were found when correlating these factors to our measure of personality. However, here we note potential criticisms of such a 'top-down' methodological approach.

The first relates to the choice and representativeness of the study corpus, and suggests that this choice predicts the findings (Baayen, 1997). Although potentially a criticism of any corpus study, this is waged particularly against the multi-dimensional approach, because the implication is that the the resulting dimensions are in some way representative of language as a whole (rather than for a particular type of corpus). In this study we note the general replicability of the dimensions which we derived, suggesting at least that our corpus is comparable to that of Pennebaker and King. However, we suggest that some of the minor variations found between our results and the previous study may be due to variation between study corpora.

A second frequently cited criticism of the factor analysis technique relates to the way that initial choices made by the researcher have significant implications for the subsequent results: For example, the selection of linguistic features for inclusion in the factor analysis determines the dimensions—and the loadings—which are derived (Altenberg, 1989). Furthermore, Altenberg also notes that difficulty in interpretation of these dimensions may be due to problems in the selection and measurement of these original features, for example if they are 'ill-defined, functionally heterogeneous, [or] stylistically skewed' (p. 171). In the current study, we have used the same criteria as Pennebaker and King which selected the same variables for inclusion in our analysis. We therefore acknowledge that these relatively narrow criteria for inclusion may indeed have influenced the factors which we derived.

Additionally, we also note that our results depend heavily upon the features and analysis performed by the LIWC program, and the validity of its subjectively defined dictionaries. Indeed, Pennebaker and King (1999) note that in some of their validation studies, 'the types of words people used varied tremendously depending on the [...] topic' (p. 1300), and acknowledge in conclusion that 'the factor structure of language may well be dependent on writing topic, setting, or implicit writing rules' (p. 1309). We propose that such topic differences between writing to a friend and writing about feelings connected with starting university, may well result in the differences which we found for positive and negative emotions words, compared with Pennebaker and King's 1999 study. In subsequent chapters we will address issues connected with content analysis techniques in more detail.

Extraversion											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	o	o							
Real.	o	o	o								
Fluency	o	o	o								
Gramm	o	o	o	o	o	o	o	o			
Conv.	o	o	o	o	o	o	o	o			
Lexis	o	o	o	o	o	o	o	o	o	o	
Percept.	o	o									

Neuroticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	o	o	o						
Conv.	o	o									
Lexis	o	o	o	o	o						+Inclusive words; −The Social Past
Percept.	o	o									

Psychoticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	o	o	o	o	o				
Conv.	o										
Lexis	$\oplus^{w}$	o	o	$\oplus^{w}$	o	o	o	o	o	o	
Percept.	o	o									

Table 3.6: Review of hypotheses

Note. o indicates an hypothesis; $\oplus$ confirmation of hypothesis; $\ominus$ inverse of hypothesis; $\odot$ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{w} LIWC factor analysis. Please refer to Section 2.6 for a full description of the hypotheses.

3.4.4 Review of the hypotheses

Here we relate our findings to those proposed in the previous chapter on the basis of the previous LIWC analysis (Pennebaker and King, 1999). Overall they show a general similarity, however few of these relationships reach statistical significance. Significant relationships which we note are Neuroticism's positive correlation with inclusive words, and negative correlation with the Social Past factor; Psychoticism also shows a negative relationship with first-person singular references. These are indicated in our summary of the hypotheses (Table 3.6). In our previous discussion, we note that these lower levels of significance are due to the relatively smaller size of our corpus. However, we also observe that given this smaller size of corpus, and also its different genre, our factor analysis successfully replicated that of Pennebaker and King (1999).

3.5 Conclusion

In this section we have described the collection of an e-mail corpus of personality texts. We have seen an overview of the different methods incorporated in the design of the experiment used for data collection, and have justified this. By comparing this newly-collected personality corpus to work previously undertaken by Pennebaker and King (1999) using the LIWC text analysis program, we have shown a similar factor structure to be present in both sets of data, despite for example, the differences in writing task, and number of participants and samples. By comparing correlations of personality traits with these factors and the LIWC categories across both studies, we have demonstrated a similar pattern of personality language. However differences are present in the data, and we explain these by: firstly, the reduced number of participants and samples elicited; and secondly, the difference in topic of the texts elicited.

So far, the analysis of our e-mail corpus has shown its validity against a much larger and varied textual resource, and the similarity of broad features characteristic of different personality traits. However, the results derived so far in the form of broad factors or text analysis categories, do not tell us which features are most important for projecting or detecting a particular trait, and nor would they be sufficiently detailed or abundant to inform the automated generation of a personality text. Additionally we have presented some arguments demonstrating that top-down, multi-dimensional approaches may not be the most appropriate for the analysis of our data. We therefore suggest that more sensitive analysis techniques are required to determine such linguistic personality characteristics of e-mail. We explore these in the next chapter.

Chapter 4

Content Analysis of Personality Language

In this chapter we explore different analysis methods which extend the work of the previous chapter. There we largely replicated the factor structure and the correlations with personality of a previous study which indicates the comparability of our corpus. However, we also questioned the suitability of multi-dimensional approaches for our study. Therefore in this chapter we investigate statistical techniques which are better suited to this study. We additionally explore the more accurate measurement of lexical density, and different analyses of content.

The chapter concludes with a discussion of results in relation to our hypotheses, and also evaluates content analysis methods.[1]

[1] Work from this chapter (and also from Chapter 5) has been partially reported in Gill and Oberlander (2002).

4.1 Introduction

The previous chapter described the data collection of our personality corpus, and also the multi-dimensional analysis which we used to compare it with a previous study. Here we investigate alternative methods of analysis. We firstly concentrate on the LIWC content analysis program which we used in the previous chapter, and analyse these data using correlation and multiple regression techniques.

In the second part of the chapter we examine different methods of analysis, namely an alternative measure of lexical diversity to that used by LIWC, and also a measure of psycholinguistic properties. We conclude the chapter by evaluating these statistical and analytic methods, and summarising our findings. We relate these back to our hypotheses.

4.2 Psychological Analysis using LIWC

The results which we have derived so far in the previous chapter are in the form of broad factors of text analysis categories. These do not tell us which features are most important for projecting or detecting a particular trait, nor are they sufficiently detailed or abundant to inform the generation of text projecting personality. In the previous chapter, we also noted criticism of the multi-dimensional approach, and questioned its suitability for the analysis of our language data. Therefore we propose the adoption of statistical techniques which allow the retention of as much of the LIWC data as possible, without merging it together into larger and more general factors. This would allow us to be able to identify specific words or categories which would better allow the linguistic detection or projection of a particular personality trait.

4.2.1 Correlation Analysis

4.2.1.1 Method

In the previous chapter, we described how each of the Past and Next texts collected from the 105 participants were analysed using the LIWC program (Pennebaker and

LIWC variable	Example words	r	p
Sports	*football, game, play*	−.23	.018
Number	*one, thirty, million*	−.20	.045
Affective processes	*happy, ugly, bitter*	.19	.055
Word Count	No. of words in text	.18	.059
Certainty	*always, never*	.18	.061
Positive feelings	*happy, joy, love*	.18	.063
Anxiety	*nervous, afraid, tense, enemy*	.17	.083
Grooming	*wash, bath, clean*	−.17	.083

Table 4.1: Correlation of EPQ-R Extraversion Scores with LIWC Variables.

LIWC variable	Example words	r	p
Inclusive	*with, and, include*	.26	.008
Swear words	*damn, piss, shit*	−.23	.019
Grooming	*wash, bath, clean*	.19	.057
Past-tense verbs	*walked, were, had*	−.19	.058
Total second person	*you, you'll*	−.18	.062
Total First Person	*I, me, we*	.17	.078

Table 4.2: Correlation of EPQ-R Neuroticism Scores with LIWC Variables.

Francis, 1999). Here we take the mean scores of Past and Next texts which we calculated for each of the LIWC dictionary variables. Exploration of the data was carried out using simple correlations of each of the mean LIWC variable scores with EPQ-R scores for each participant (see Tables A.1 and A.2 for further information about participants).

4.2.1.2 Results

The results of the correlation analysis which demonstrated a significant relationship between EPQ-R personality score and LIWC variable at the relatively generous $p < .1$ level (indicating 10% level of chance for our findings) are shown for Extraversion, Neuroticism, and Psychoticism traits (Tables 4.1, 4.2, and 4.3).

From the correlation analysis, it is apparent that the most linguistic features are correlated with Psychoticism. Extraversion relates to fewer features, and Neuroticism demonstrating the fewest of all correlations; when the number of variables showing

LIWC variable	Example words	r	p
Motion	*walk, move, go*	−.28	.004
Total First Person	*I, me, we*	−.27	.006
Death	*dead, burial, coffin*	.25	.010
Certainty	*always, never*	.24	.014
First Person Singular	*I, my, me*	−.23	.016
Anger	*hate, kill, pissed*	.23	.017
Feeling	*touch, hold, felt*	−.23	.019
Swear words	*damn, piss, shit*	.21	.035
Total pronouns	*I, our, they, you're*	−.21	.035
Negative emotions	*hate, worthless*	.20	.042
Sadness	*grief, dry, sad*	.19	.057
Dictionary Words	Words captured by LIWC	−.18	.074
Money	*cash, taxes, income*	.17	.077
Humans	*boy, woman, group*	.17	.076
Cognitive processes	*cause, know, ought*	.17	.085
School	*class, student, college*	−.16	.097

Table 4.3: Correlation of EPQ-R Psychoticism Scores with LIWC Variables.

relationship at the $p < .05$ are considered, Psychoticism again shows by far the most features, whereas Extraversion and Neuroticism equally show relatively few.

Examination of the features characteristic of strongly Psychotic texts reveals an interpersonal distance characterised by the lack of pronouns—particularly first person—combined with an increase in more references to 'humans' in a more abstract sense. Additionally a negative, hostile impression is created through the use of Negative Emotion, Sadness and Swear words, and especially words relating to anger and death, whilst words discussing feelings are omitted. Words relating to thoughts, and especially certainty, and money are all used more, whilst words relating to education and actions are not. There was also a tendency to use words which are not covered by the LIWC dictionary.

Extraverts tended to write longer texts which contained fewer references to sports, numbers or grooming. Additionally Extraverts appeared more open in their ability to express emotions more generally—whether they were good or bad—and also were more likely to express certainty.

Highly Neurotic authors used fewer past-tense verbs and more inclusive words in

their texts. They also showed a distinctive pattern in their use of pronouns, showing a preference for first person references and avoid the use of second person references. Furthermore, unlike high Psychotics there was less evidence of swearing, and unlike the Extraverts a greater reference to grooming.

In summary of these findings from the correlation of personality measures with the LIWC linguistic features: High Psychotics use an interpersonally distanced style, make more negative and aggressive references, and are more likely to use more unusual words; Extraverts write more, express more positive and negative emotions, and make fewer references to sports or numbers; High Neurotic authors use more inclusive words, and fewer swear words, make fewer references to the past, and talk about themselves rather than other people.

These findings generally appear intuitively related to the personality types described by our theory-driven hypotheses (we will discuss these in detail at the end of the chapter; cf. Eysenck and Eysenck, 1975), and provide a much greater detail of linguistic features than those derived from the more general factor analysis. However, this method of analysis using multiple correlations is not optimal. First such use of multiple measures encourages the likelihood of statistical error, since the individual tests do not take into account the combined chance of significance occurring, leading to false positives (Type I error); Secondly, from the individual correlations it is not possible to examine how several features may fit together to give the best possible combination characteristic of a personality type. We therefore introduce a further analysis technique.

4.2.2 Multiple Regression Analysis

4.2.2.1 Data and Statistical Requirements

Multiple regression is an analysis technique which allows the relationship between variables to be investigated. Although similar to correlation, multiple regression shows the degree to which one or more *independent variables* can explain the *dependent variable* (Oakes, 1998). It therefore provides the best fit available, and maximum variance explained, from the combination of a number of predictor variables with the

predicted, dependent variable. Several methods can be used to select these predictor variables: in the current study, a 'stepwise' analysis was regarded as the most suitable, since variables are entered if they show a significant relationship with the independent variable—similar to the 'forward' selection procedure—but once in the equation, they are removed if they do not correlate significantly enough—as in a 'backward' elimination method—in the end retaining the equation which best explains the variance of the independent variable (Norušis and SPSS Inc., 1994). As Hinton (1995) notes, rather than focusing too closely upon the individual significant relationships between independent and dependent variables, it is better to view the r^2 value—or coefficient of determination—since this describes the total variance which the equation explains. In multiple regression analysis, it is assumed that the variables under investigation are either interval or ratio, and are linearly related (Oakes, 1998), and this is the case with our linguistic variables.

4.2.2.2 Our Use of Multiple Regression Analysis

As described above, multiple regression is used to estimate which independent variables are most useful in predicting the dependent variable. In the forthcoming analyses, we have utilised multiple regression because of its ability to select variables on the basis of their combined relationship with another variable, rather than necessarily because we want to subscribe to the implied cause and effect relationship. For example, in our analysis we have included personality as the *dependent* variable; this however, does not mean that we believe personality to be *caused* by the linguistic features. Therefore our reasons for performing the analysis in this manner are as follows:

1. Since we are interested in the overall projection and realisation of personality through language, we are interested in which combination of linguistic variables best give a sense of the language produced by an individual along each particular personality dimension. Therefore, it is the combination of these features which overall indicate personality, rather than the contribution of each personality type to the realisation of each individual linguistic variable, which we are most interested in.

2. Assigning personality traits as independent variables (and thus predicting the dependent linguistic variables) would lead to a regression analysis being performed for each of the linguistic variables. This use of multiple measures—as in the case of correlations—leads to an inflation of overall Type I error rate, and is therefore undesirable. Also, although it is possible to compare the relative strength of relationship between the personality and linguistic variables, this approach does not indicate what possible combination of linguistic variables are important in their overall relationship with personality.

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

Therefore the following multiple regression analyses assign the personality trait as the dependent variable, and the linguistic features as the independent variables.

4.2.2.3 Statistical Basis for Selection of Variable

For each personality trait in turn (Extraversion, Neuroticism, and Psychoticism) a stepwise regression was performed on the LIWC variables which reached $p < .1$ significance in the correlations, with the variables entered in order of strength of correlation (as found in Tables 4.3, 4.1, and 4.2). Since multiple regression demands that the independent variables (that is, the LIWC variables in this study) are in fact *independent* of each other, this meant that some additional pre-selection had to be carried out before entering variables from the correlation into the multiple regression equation.

In selecting items to enter into the equation, generally the most specific (and therefore lowest level sub-category) LIWC variable was chosen, although if this variable was not retained in the eventual regression equation, then the analysis was re-run with the superordinate replacing the sub-category(s). If in the final equation, the superordinate category was retained where sub-categories were not, then the results from such superordinate analysis are used.

4.2.2.4 Theoretical Basis for Selection of Variables

In order to ensure the greater comparability of our regression analyses, a number of selection criteria have been applied to the linguistic variables, in addition to the statistical requirements of the tests (independence and relationship with the dependent variable).

Topic independence Perhaps the most obvious limitation upon the generalisability of written texts is that of topic. Indeed, Pennebaker and King (1999) noted this in their factor analysis of LIWC variables, and as a result excluded variables categorised as personal concerns (also known as current concerns), because they viewed them as being more topic relevant than process related. Similarly, in order to abstract away from demands of topic, and to help ensure that our results are not specific to such writings, we have also performed our analyses with personal concern words omitted.

Independence across genre Whilst the exclusion of personal concern words allows the topic of a text to be abstracted away from the analysis, these results only appear representative of e-mail, and possibly other personal written communicative texts. In the current analysis it is desirable to identify linguistic features which characterise personality at a deeper level and which are not constrained by genre, thereby allowing greater generalisability of such findings. Pennebaker and King (1999) addressed this issue of linguistic reliability across genre in their factor analysis by only selecting variables for inclusion which demonstrated consistency of usage in the LIWC validation studies by having an overall Cronbach's α greater than .60. Therefore, we also adopt this metric to ensure linguistic independence across genre.

Independence of language sparsity In order to ensure that maximum independence across texts is achieved for linguistic features identified by these analyses, a minimum frequency of occurrence is also specified. Pennebaker and King (1999) in their analysis, require that linguistic variables occur with a minimum frequency of at least 1%. The merits of this are apparent: by specifying a minimum word category usage, this ensures a more accurate account of the linguistic characterisation across different texts, especially shorter samples, since in theory such resulting features would occur in a text with a minimum length of 100 words. Although in more recent studies (Newman et al., in press), Pennebaker and collaborators have used a lower minimum frequency,

a more stringent measure of sparsity was regarded as more appropriate in the current study.

Conversely, the disadvantage of this approach is that such restrictions deny the possibility of finding hapax legomena, that is a word which occurs once in a text (Hunston, 2002). This may, for example be very characteristic of one personality type in particular, and would thus allow the identification of the author of the text in some way. For example, as a high Psychotic, or as a highly Extraverted individual. However, due to the sparsity of such features, they may not be reliable indicators of particular personality language types.

4.2.2.5 Method

As in the previous correlations, we again use mean LIWC dictionary variable scores for past and next texts, along with personality information for each of the participants. These data are then subject to the following analyses.

Regression including all variables As noted above, there are statistical requirements of the multiple regression analysis, for determining the variables which can be used in the analysis. The result of entering all suitably independent variables is shown in Table 4.4.

Regression calculated for topic independence In order to similarly assess which LIWC variables most characterised personality when topic was discounted, the previous analysis was repeated with Personal Concern variables excluded from the equation. The results can be found in Table 4.5.

Regression showing genre independence To ensure consistency of linguistic features used across genre, Pennebaker and King's metric of linguistic reliability, that is consistency of usage across their validation studies of greater than .60, was used for selection of variables into the multiple regression analysis. The result can be found in Table 4.6 (although the analysis was repeated with the inclusion of Personal Concern categories which did show consistency across validation studies, these did not make it into the final equation).

Regression independent of data sparsity To identify word categories used with con-

Dependent variable	Independent variables	β	*p*	R^2	*p*
E Score	Sports	−.27	.0052		
	Affective Processes	.23	0155	**.11**	**.0032**
N Score	Inclusive	.28	.0036		
	Total First Person	.21	.0302	**.11**	**.0030**
P Score	Motion	−.24	.0069		
	Death	.18	.0369		
	Certainty	.20	.0196		
	First-Person Singular	−.21	.0144		
	Feeling	−.23	.0086		
	Sadness	.18	.0463	**.27**	**.0000**

Table 4.4: LIWC Multiple Regression Analysis with EPQ-R Scores.

sistent frequency across texts, in addition to independence of genre, further regression analysis was conducted in which words with a mean usage of less than 1% were excluded (found in Table 4.7).

4.2.2.6 Results

Regression equation for all variables The result of entering all eligible variables is shown in Table 4.4.

The language use of high Extraverts was found to include less references to Sports (*football, game, play*), whilst using more words denoting Affective processes (*happy, ugly, bitter*). The beta values showed a −.27 and .23 correlation respectively, with this accounting for 11% of variance in Extraversion (R^2 = .11).

Turning to the highly Neurotic individuals, they were found to use a greater number of Inclusive words (*with, and, include*) and make more references to First Person (both themselves alone and in combination with others, for example, *I, me, we*). The β correlations were .28 and .21, with the amount of Neuroticism variance accounted for being 11% (R^2 = .11).

As can be seen from this analysis, highly Psychotic individuals tend to use fewer Motion words (such as *walk, move, go*), Feeling words (*touch, hold, felt*), and refer to

Dependent variable	Independent variables	β	*p*	R^2	*p*
E Score	Numbers	−.21	.0267		
	Word Count	.20	.0345	**.08**	**.0144**
N Score	Inclusive	.28	.0036		
	Total First Person	.21	.0302	**.11**	**.0030**
P Score	Motion	−.25	.0044		
	Certainty	.21	.0178		
	First-Person Singular	−.21	.0161		
	Feeling	−.25	.0054		
	Sadness	.19	.0309	**.26**	**.0000**

Table 4.5: LIWC (Topic Controlled) Multiple Regression Analysis with EPQ-R Scores.

themselves less frequently (i.e., use fewer First Person Singular words, such as *I, my, me*). These relationships showed β values of between −.21 and −.24. In addition, high Psychotics also used more Certainty words (*always, never*), and made more reference to Death (*dead, burial, coffin*) and Sadness words (*grief, dry, sad*). Beta values for these relationships were between .18 and .20, with all these words in this equation accounting for 27% of the Psychoticism variance (R^2 = .27).

Regression equation independent of topic To establish the LIWC variables which most characterised personality when topic was discounted, we conducted the following analysis. The results can be found in Table 4.5.

Since no Personal Concern word categories made it into the previous multiple regression equation for Neuroticism, this remains unchanged in the current analysis. However, in the case of Psychoticism, omitting Personal Concern 'Death' words from the regression analysis leads to the other variables from the previous analysis being retained in the equation with even stronger correlations (Motion, −.25; Feeling, −.25; First-Person Singular, −.21, Certainty, .21; Sadness, .19), and accounting for only slightly less of the Psychoticism variance (26% versus 27% with Death words included).

The exclusion of the Personal Concern category Sports words from the analysis of Extraversion resulted in an entirely different set of variables in the equation. In this

Dependent variable	Independent variables	β	*p*	R^2	*p*
E Score	Numbers	−.21	.0267		
	Word Count	.20	.0345	**.08**	**.0144**
N Score	Inclusive	.28	.0036		
	Total First Person	.21	.0302	**.11**	**.0030**
P Score	First Person Singular	−.30	.0018		
	Anger	.22	.0188		
	Cognitive Mechanisms	.24	.0148	**.16**	**.0005**

Table 4.6: LIWC (Genre Controlled) Multiple Regression Analysis with EPQ-R Scores.

case, the solution to explain the most variance in highly Extraverted texts meant that fewer references to Number words (*one, thirty, million*) were found, whilst the texts were found to be longer (i.e., greater Word Count). The Beta correlations were −.21 and .20 for Numbers and Word Count respectively, which explained 8% variance in P Score (R^2 = .08). Affective processes, although entered into the regression analysis, did not reach the required significance for retention in the final equation.

Regression equation independent of genre In Table 4.6, of the variables which achieved the required consistency, the variability for texts produced by high Neurotics again is best described by higher use of Inclusive words and references to First Person. The Beta correlations are the same as before, .28 and .21 respectively, with the overall variance accounted for being 11%.

Similarly the results for high Extraverts again showed their preference for using fewer Number words combined with producing longer texts (i.e., Word Count). Again the Beta correlations were the same (−.21 and .20) respectively, with this accounting for 8% of Extravert variance. In their factor analysis, Pennebaker and King report that they excluded LIWC linguistic categories which 'did not refer to features or meanings of specific words', citing Words per Sentence, or Word Count as examples. When Word Count is excluded from the regression analysis, Number words are in the equation to explain 4% of the Extraversion variance ($\beta = -.21$; p=.0452; R^2=.04).

In contrast to the analyses of Neuroticism and Extraversion, the removal of insufficiently consistent variables from the analysis of Psychoticism led to several vari-

Dependent variable	Independent variables	β	*p*	R^2	*p*
E Score	*None*				
N Score	Inclusive	.28	.0036		
	Total First Person	.21	.0302	**.11**	**.0030**
P Score	First Person Singular	−.31	.0021		
	Cognitive Mechanisms	.26	.0099	**.11**	**.0020**

Table 4.7: LIWC (Sparsity Controlled) Multiple Regression Analysis with EPQ-R Scores.

ables which had featured in the previous final equation being excluded completely (Motion and Feeling), whilst Certainty was replaced by the superordinate—and more consistent—category, Cognitive Mechanisms. Therefore, when consistency of usage is taken into account, the linguistic characteristics of highly Psychotic language are low self reference (First Person Singular; $\beta = -.30$), increased use of Anger words ($\beta = .22$) and Cognitive Mechanism words, such as *cause, know, ought* ($\beta = .24$). These linguistic features accounted for 16% of the variance in Psychoticism (R^2=.16).

Regression Equation Independent of Data Sparsity To identify words consistent across genre, and which are also used frequently enough to avoid issues associated with data sparseness, further regression analysis was conducted in which words with a mean usage of less than 1% were excluded (found in Table 4.7).

Again, the LIWC linguistic variables which best explain the variance in highly Neurotic individuals are increased Inclusive words and references to first person (βs = .28 and .21; $R^2 = .11$). Number words failed to reach required frequency levels, and therefore were not entered into the analysis, and when multiple regression was again performed, none of the remaining variables reached the required significance level ($p < .05$) for retention in the equation.

In the case of Psychoticism, Anger words were excluded due to their infrequency, and the resulting equation shows a reduced usage of First-Person singular words and an increased use of Cognitive Mechanism words in high Psychotic language. The Beta correlations are −.31 and .26 respectively, with these variables accounting for 11% of Psychoticism variance ($R^2 = .11$).

4.2.2.7 Summary

Since the use of multiple regression was proposed in order to select the most useful linguistic characteristics of personality, it is unsurprising that this has reduced the features greatly in comparison to those found related to personality in the correlation analysis. Indeed, by preselecting variables for the regression analysis based on independence from topic, genre, and data sparsity the number of features resulting from each analysis has fallen in most cases. This however, is not the case for the characteristic features of Neuroticism—increased use of Inclusive words and First Person references—which have remained stable throughout the analysis, since these appear to be relatively stable, central linguistic features of LIWC.

Extraversion also appears to be relatively stably characterised by fewer references to numbers and an increased text length once the topic-specific category of Sports is excluded (which also led to Affective Processes being lost from the equation).

Due to the larger number of variables available to be entered into the regression equation, Psychoticism produces a much more varied pattern of characteristic features across the conditions. The core features which are consistent throughout are fewer first-person singular referents and thought-related words (with Cognitive Mechanism words subsuming the Certainty category). Negative-hostile words are featured throughout, with the exception of the data-sparsity controlled equation. Motion—perhaps a less intuitive feature of Psychoticism—is lost half way through the variable selection due to lack of consistency of usage across all genres.

4.2.2.8 Methods of variable selection

As discussed above there are issues regarding the pre-selection of variables for entry into the regression equation. Both topic and genre independence are desirable for increasing the generalisability of features to different text types. However, the selection of variables on the basis of independence from data sparsity using a relatively stringent occurrence criterion was questioned due to the possibility of characteristic hapax legomena, that is the rare occurrence of words which may 'give away' a certain personality type.

With reference to the above data, by selecting to avoid data-sparsity, we found that

Anger would be lost from the Psychoticism regression equation. This de-selection of Anger from the equation suggests this more stringent approach is correct, since if such a characteristic does not occur consistently across texts, then we would not like to regard it as a consistent feature of Psychoticism. References to Numbers and length of text (Word Count) are also lost from the Extraversion equation at this stage.

4.3 Analysis of Lexical Diversity

Lexical diversity is a term used to describe the informational content of a sample of language, and is usually expressed in terms of measures of repetition. In broad terms this can be used as a description of language complexity (e.g., Trott, 1994). Studies looking at the variation of lexical diversity with accent (Giles et al., 1981; Bradac and Wisegarver, 1984) and socio-economic factors (Bradac et al., 1976) have led Bradac (1990) to claim that it exhibits within group variation, rather than between group variation, which would exist between speakers of different classes or gender.

An early finding related lexical diversity inversely to the anxiety of the communicator, so that as a person's anxiety increased, their lexical diversity would decrease (Howeler, 1972). Furthermore, studies investigating the attitudinal consequences of lexical diversity have indeed found that it is inversely related to perceived anxiety (Bradac et al., 1980), as well as to perceptions of the speaker's competence and message effectiveness (Giles et al., 1981).

Such relationships between lexical diversity and within group variation along with its relationship to perceived and actual measures suggests its appropriateness for study in conjunction with measures of individual difference such as personality. Additionally, since it does not rely upon a pre-defined dictionary database for its analysis[2] it can provide further insight into measures such as the amount of words not covered by a standard dictionary (e.g., tendencies shown by Psychoticism with regard to the LIWC dictionary reported above), and whether this is related to repetition and use of many different words.

[2]Note that since the calculation of TTR here relies upon the pattern matching of word strings, care needs to be taken in cleaning up the corpus for spelling errors. This is described in more detail below, in Section 4.3.2.

4.3.1 Calculation of Lexical Density

The lexical density of a text can be calculated using the number of unique words (Types) divided by the number of words (Tokens), which results in the type-token ratio (or TTR). The TTR can then be multiplied by 100 to give the percentage of unique words in a text. This is the method of TTR calculation utilised by LIWC (Berry et al., 1997; Pennebaker and Francis, 1999; Pennebaker and King, 1999). Whilst this calculation can be performed on a text as a whole, it is subject to a 'length effect', whereby lower TTRs, and thus smaller numbers of unique words, are found—unsurprisingly—in longer texts. This can be particularly problematic if comparison of TTR is to be carried out between texts of different lengths (Gill, 1998). Therefore, in order to avoid this 'length effect', TTR can be calculated by dividing the text into a series of 'bins' of a set word length, with the TTR calculated individually for each of these bins, and then the mean TTR of these bins reported for the text as a whole (cf. Bradac et al., 1977, 1988).

4.3.2 Measurement of Lexical Diversity

In analysing the e-mail data, several measures of lexical density were used: TTR for the whole text, and TTR calculated using the bin technique, with the sizes of bin used being of 50, 25, and 10 word lengths. This was calculated for each text (Past and Next) individually,[3] and in the case of calculating bin TTRs, any words remaining at the end of a text which would not fill the required sized bin were discarded. The spelling corrected versions of the e-mail texts which were used for the LIWC analysis, were used here for calculating TTR. Since the Perl script calculating TTR use a strict pattern-matching technique, different spellings of the same word would have incorrectly been counted as different words.

[3]Each text was first run through a tokeniser script which removed punctuation, before TTR was calculated by a second Perl script.

TTR or related variable	r	p
Total Unique Words	.19	.058
25 Bin Total Unique Words	.18	.060
25 Bin Total Words	.18	.063
Total Words	.18	.066
10 Bin Total Unique Words	.18	.067
10 Bin Total Words	.18	.069
50 Bin Total Unique Words	.17	.077
50 Bin Total Words	.17	.080

Table 4.8: Correlation of EPQ-R Extraversion Scores with TTR and related variables.

TTR or related variable	r	p
10 Word Bin TTR	−.27	.006
50 Word Bin TTR	−.24	.013
25 Word Bin TTR	−.23	.021

Table 4.9: Correlation of EPQ-R Neuroticism Scores with TTR and related variables.

4.3.3 Correlation Analysis

4.3.3.1 Method

Initial exploration of the lexical density data was carried out by calculating Pearson correlations for each TTR and related measure along with the EPQ-R personality scores (see Tables A.1 and A.2).

4.3.3.2 Results

Pearson correlation coefficients for Extraversion, Neuroticism, and Psychoticism, are displayed for TTR or related features which showed a two-tailed significance of $p < .1$ (Tables 4.8, 4.9, and 4.10).

As can be seen from the correlations of personality and TTR, high Psychotics show a preference for using more unique words, with this being most apparent over a greater section of text which indicates that this is a genuine feature, rather than an effect of a smaller bin measurement size (which often exaggerates diversity). This is contrary to the findings for high Neurotics who use a lower lexical diversity more generally. Since

TTR or related variable	r	p
50 Word Bin TTR	.32	.001
25 Word Bin TTR	.23	.019
10 Word Bin TTR	.18	.061

Table 4.10: Correlation of EPQ-R Psychoticism Scores with TTR and related.

this is less distinguishable across bin sizes, but the strongest relationship is found for the smallest bin measurement, this indicates that their language is more repetitious as a whole, rather than as an artifact of length. Whilst the majority of the correlations between Psychoticism and Neuroticism and TTR bin results are approaching strong significance, the correlation results for Extraversion are all features of length, and show a trend towards higher Extraverts producing longer texts.

4.3.4 Multiple Regression Analysis

4.3.4.1 Method

Stepwise multiple regression analysis was performed on the linguistic features which showed at least a correlation of $p < .1$ with the measures of personality. Although all TTR measurements relate to lexical diversity, here we treat them as independent features, since they are calculated in different ways.

4.3.4.2 Results

The results of the stepwise multiple regression analysis are shown in Table 4.11. As expected from the simple correlations, high Psychotics show a preference for greater lexical diversity when measured across the larger text sample of 50 words (50 Word Bin TTR, $\beta = .32$) which accounts for 10% of Psychotic variance (R^2=.10), whereas higher Neurotics demonstrate lower lexical diversity when measured across a smaller text sample (10 Word Bin TTR, $\beta = -.27$) which explains 7% of Neurotic variance (R^2=.07).

Since a significance of $p < .05$ was required for a variable to be retained in the stepwise multiple regression analysis equation, it is of little surprise that none of the

Dependent variable	Independent variables	β	p	R^2	p
E Score	*None from Stepwise analysis*				
N Score	10 Word Bin TTR	−.27	.0057	**.07**	**.0057**
P Score	50 Word Bin TTR	.32	.0010	**.10**	**.0010**

Table 4.11: TTR Multiple Regression Analysis with EPQ-R Scores.

Extraversion features appeared in the final equation (the previous discussion of the correlations, above, noted that none of them achieved a significance of $p < .05$).

The most salient results from the regression analysis relate to the apparent opposition between the diversity characteristics of Psychoticism and Neuroticism. This is realised by the regression beta value being almost equal and opposite in polarity, and also in bin-size—the length of text over which the diversity is measured. In both cases this confirms trends apparent in the previous correlation results: High Psychotic authors show a preference for using many different words in their texts, whereas Highly Neurotic authors use much more repetitious language. The fact that the greatest R^2values for the regression analysis are achieved using different bin sizes relates back to the effects of length—in this case size of bin used for measuring TTR—and for the Psychoticism results a larger bin measurement exaggerates the measurement of linguistic variety whereas for Neuroticism the repetition is increased by using a small bin measurement. This result confirms the need to take length into account in calculating lexical diversity, and the effects that such measures can contribute towards the results.

However, referring back to the correlation of TTR and personality reported above it can be seen that if the middle-sized 25 word bin results are compared (as used by Bradac et al., 1988), they still show equal and opposite significant effects (.23), although this is admittedly slightly weaker.

4.4 Psycholinguistic Properties of the Texts

In addition to the LIWC and lexical diversity analyses, here we describe an additional technique. This uses a novel method which derives content analysis of a text based on

the properties of words as featured in the empirically derived MRC Psycholinguistic Database (Coltheart, 1981; Wilson, 1987, see review above, in section 2.5.3.5). Therefore, like the LIWC analysis, the MRC Psycholinguistic Database analysis provides psychological information about the language, however this relates to psycholinguistic properties rather than psychological properties. The methods used for our implementation, are described in detail below.

4.4.1 MRC Analysis Technique

An initial edit of the machine usable dictionary was carried out using the standard MRC Psycholinguistic Database utility tool 'PsychDict' in order to restrict the database to words for which psycholinguistic information was available. From the total number of 150,837 database words, this figure was reduced to a more manageable 39,300 words.

In order to allow the disambiguation of different word senses and allow the accurate lookup of words from the corpus, the spelling corrected version (as used for LIWC and TTR analyses) of the e-mail corpus was tagged for parts of speech. This was performed using Ratnaparkhi's MXPOST maximum entropy tagger (Ratnaparkhi, 1996),[4] since when evaluated against hand-tagged sample texts this gave the most accurate performance for the current data.[5] The lookup program extracts each word and its respective POS in turn for each individual e-mail text. Since the number and detail of POS classifications used by MXPOST was much greater than the ten used by MRC Database, an initial step was to 'translate' (using a simple algorithm) the tags into a form compatible with the database, with this subsequent part of speech information used to distinguish between different senses of the target word.[6] Each word and associated POS pair were then looked-up in the minimised MRC database.

[4]MXPOST is available from `http://www.cis.upenn.edu/~adwait/statnlp.html`.

[5]Thanks to Tim Willis for conducting this evaluation of taggers.

[6]In the case of several entries for the same word with the same POS, psycholinguistic information from the first entry was used by the look up program.

4.4.2 Calculation of MRC psycholinguistic textual properties

Once each word in the text had been looked up in this way, values for psycholinguistic information for each of these words (if present in the database) were collected for each of the following measures: Number of Letters, Number of Phonemes, Number of Syllables, Kucera and Francis Frequency (including categories and samples measures), Thorndike and Lorge frequency, Brown Verbal Frequency, Familiarity, Concreteness, Imagability, Meaningfulness, Age of Acquisition, and use of word status categories (e.g., Standard, Dialect, Archaic, Obsolete, etc.). For each of these categories, the mean and standard deviation of the entries was calculated.

In addition to this psycholinguistic information, the program also calculated the number of words captured by the database, the total number of strings (one or more characters separated by one or more spaces) in the e-mail text, the percentage of words which were and were not captured by the dictionary and the number and percentage of groups of numbers (0-9) and non-alphanumeric characters (!@#., ?/, etc).

4.4.3 Correlation Analysis

4.4.3.1 Method

Pearson correlation coefficients and two-tailed tests of significance were calculated for each psycholinguistic (and related property) property with the EPQ-R personality scores (Tables A.1 and A.2).

4.4.3.2 Results

The results of simple correlations ($p < .1$) of this psycholinguistic and additional data with the EPQ-R personality scores can be found in Tables 4.12, 4.13, and 4.14.

For Extraversion (Table 4.12), the most strongly significant finding relates to the inverse relationship with use of less concrete language. The majority of the marginally significant findings appear to be related to a positive length effect (i.e., that Extraverts produce longer texts), with the exception of their increase in the use of percentage of dialect words.

MRC variable	r	p
Mean Concreteness	−.21	.028
Number of Words Captured by Dictionary	.18	.061
Total Strings	.18	.069
Count of Dialect Words	.17	.075
Std. Dev. of Kucera & Francis no. of Categories	.17	.079
Count of Standard Words	.17	.087
Percentage of Dialect Words	.17	.089

Table 4.12: Correlation of EPQ-R Extraversion Scores with MRC.

MRC variable	r	p
Mean Concreteness	.27	.005
Std. Dev. of Concreteness	.21	.033
Number of Digits	−.20	.041
Mean Brown Verbal Frequency	.19	.052
Percentage of Digits	−.18	.068

Table 4.13: Correlation of EPQ-R Neuroticism Scores with MRC.

MRC variable	r	p
Percentage of Words not in Dictionary	.27	.005
Mean Number of Phonemes	.20	.039
Mean Number of Syllables	.18	.073
Percentage of Obsolete Words	.17	.083
Std. Dev. of Kucera & Francis Writ. Freq.	.17	.092

Table 4.14: Correlation of EPQ-R Psychoticism Scores with MRC.

Dependent variable	Independent variable	β	p	R^2	p
E Score	Mean Concreteness	−.21	.0278	**.05**	**.0278**
N Score	Mean Concreteness	.33	.0006		
	Mean Brown Verbal Frequency	.27	.0055	**.14**	**.0004**
P Score	Percentage of Words not in Dictionary	.29	.0024		
	Std. Dev. of Kucera & Francis Written Frequency	.19	.0432	**.11**	**.0023**

Table 4.15: MRC Multiple Regression Analysis with EPQ-R Scores.

From the correlation data (Table 4.13) high Neurotics—in contrast to the behaviour of Extraverts—tend to use more concrete language overall. Their additional positive relationship to the standard deviation of concreteness appearing to indicate that the high Neurotics use also use words with a greater variety of concreteness. Neuroticism also shows an inverse relationship with the number of digits used in a text, and also show a marginal significance for language more frequently found in speech.

High Psychotics (Table 4.14), tend to use longer, more unusual words (demonstrated by the positive relationship with percentage of words not in dictionary, number of syllables and phonemes). They also show a tendency to use more outmoded or obsolete words.

4.4.4 Multiple Regression Analysis

4.4.4.1 Method

Psycholinguistic and additional properties showing a significant correlation of $p < .1$ were entered into a stepwise multiple regression analysis for the respective personality traits. As in previous regression analyses, precautions were taken to ensure independence of variables entered into the equation.

4.4.4.2 Results

When these variables showing a significance of $p < .1$ are entered into a stepwise multiple regression analysis for the respective personality traits, the equation results can be found in Table 4.15.

The regressions show high Psychotics use more unusual non-dictionary words, and of the more standard words that they use, they prefer to use a variety of frequent and infrequently occurring words (Percentage of Words not in Dictionary, $\beta = .29$; Std. Dev. of Kucera & Francis Written Frequency $\beta = .19$), explaining 11% of variance in P Score ($R^2 = .11$). High Extraverts show a tendency to use words rated as being less concrete, with a Beta correlation of $-.21$, which explains 5% of variance ($R^2 = .05$). Turning finally to the characteristics of the language of high Neurotics, this shows greater concreteness, and a preference for words which are found to be common in speech (βs=.33 and .27 respectively), accounting for 14% of variance.

The most noticeable feature of the regression analysis of the psycholinguistic data is the lack of features again found for Extraversion, in this case solely a low average level of concreteness in language use marks out a high Extravert. The mean concreteness of a text also identifies the personality type for which the greatest variance can be explained: Neuroticism. Here, conversely high levels of concreteness, combined with the use of more common words found in speech characterise higher Neurotics. Higher Psychoticism also shows a relatively high psycholinguistic characterisation, being demonstrated by the use of more unusual words and a variety of words found frequently and infrequently in writing.

4.5 Discussion

In this chapter we have investigated two issues: firstly, the identification of linguistic features which are most characteristic of particular personality types; secondly we have investigated additional analyses to that provided by LIWC. Using these approaches we have found that overall Psychoticism explains the greatest linguistic variation across all analyses, followed by Neuroticism. Linguistic features of Extraversion appear to be least well identified by these analyses. Overall the psycholinguistic analysis found

the most features, followed by the psychological analyses of the LIWC, with the measurement of lexical diversity and related variables finding the fewest.

The discussion of this chapter is structured as follows: we first evaluate the analysis methods used in this chapter; secondly, we describe the linguistic characteristics of the personality and relate these to our hypotheses.

4.5.1 Approaches to analysis

Using multiple regression analysis we have shown which LIWC and MRC psycholinguistic features best characterise different personality dimensions. This is an advancement on the multi-dimensional approaches used in the previous chapter, because it does not impose a top-down selection of variables and factor structure upon the data. Rather, it enables the identification of specific linguistic variables which may be particularly related to a personality trait. Furthermore, the information provided by the LIWC and MRC analyses of the personality texts provides a valuable insight into the way in which the use of psychologically and psycholinguistically significant words characterises different personality traits.

However, we note that there are limitations associated with the content analysis approach of both LIWC and MRC. The first relates to the analysis at a theoretical level: These content analysis approaches are limited by the items which are defined in the dictionaries, since this is a 'top-down' method of linguistic analysis.[7] This is particularly the case for the LIWC analysis, because its dictionaries have been selected for their psychological relevance; the MRC database is designed to have a much broader coverage, extending across the majority of the vocabulary. In the case of the LIWC, with the exception of the Linguistic Dimensions, these dictionary categories may not necessarily generalise well across genre or topic. Pennebaker and King (1999), as we have previously noted, observed that in some of their validations studies, 'the types of words people used varied tremendously depending on the [...] topic' (p. 1300). Similarly, in the replication of a study which linked therapeutic outcomes with LIWC analysis of diary entries, Stephenson et al. (1997) concludes 'the work of Pennebaker

[7]Note that here 'top-down' is used as a description of content analysis, and does not necessarily mean multi-dimensional analysis. However, as we demonstrated in the previous chapter, a 'top-down' content analysis method can be incorporated into a 'top-down' multi-dimensional analysis method.

and others suggests that at least the *approach* we have adopted will prove fruitful, even if the particular linguistic correlates of progress vary from one treatment setting to another' (p. 409). Indeed, Mehl and Pennebaker (in press) have recently acknowledged that linguistic style as measured by articles, prepositions, first person pronouns, present tense verbs, and positive and negative emotion words showed greater consistency than content. In response to such content-specific limitations of 'top-down' content analysis, Campbell and Pennebaker (2003) adopt latent semantic analysis as a 'bottom-up' alternative method (see e.g., Landauer and Dumais, 1997, for an overview of this approach). We note with relation to this thesis however, that although latent semantic analysis is a data-driven approach, it expresses its findings in terms of vector measures for the texts. This therefore, is even more opaque than, for example multi-dimensional analysis, which does at least allow the examination of linguistic features which compose the factors. Latent semantic analysis has therefore not been adopted as one of the analytic methods of this thesis.

Secondly, there are practical problems related to the LIWC and MRC analyses: Ball (1994) notes that a problem for all top-down approaches is that of 'recall', which relates to the technique's success in identifying and counting features. This is particularly pertinent to LIWC, due to the relative small size of its dictionaries (despite their inclusion of words and word-stems to broaden potential matches, these only total around 2,000 words compared with the 40,000 of the MRC database); furthermore the simple pattern-matching technique used to identify input words with the dictionaries in both techniques depends upon the input texts being 'cleaned' or edited to specific guidelines, with failure to do so resulting in the words not being recognised—or counted. An additional problem is that this precludes the incorporation of systematic non-standard features (e.g., words or spellings) in the analysis. Additional problems relate to the replicability of top-down analysis, since this often uses specific methodologies or categorisation techniques (Tribble, 2000). In the case of LIWC, this is a particular problem, since the dictionaries or algorithms are not published (cf. e.g., Biber, 1988, who publishes these details with his analyses), and to perform comparable analysis requires purchasing the LIWC software. In the case of the psycholinguistic analysis, the MRC Psycholinguistic database is freely accessible, but at present a

look-up program is not generally available.

The analyses of lexical diversity accounted for lesser variance than the two content analysis approaches, however, we maintain that this is still a valuable analysis method. Indeed, since it does not rely upon pre-defined dictionaries, it is unaffected by many of the criticisms of the content analyses methods raised above. Indeed, we note that this analysis method was able to identify differences between Neurotic and Psychotic language use. However, as discussed above, this technique depends upon consistency of spelling in the corpus, in order to identify repeated terms, and also the length of the sample must be considered for the lexical diversity to be more generally comparable.

4.5.2 Summary of findings and review of hypotheses

Using these approaches we have been able to identify a number of linguistic features which are characteristic of the different personality traits. We now relate these back to our original hypotheses (Table 4.16) and discuss the features of Extraversion, Neuroticism and Psychoticism:

We therefore describe the characteristics of Extravert language as a use of words which are related to more abstract concepts, such as *thoughts, flavours, pains*. Despite the hypothesis that Extraverts would show a lower lexical diversity, this was not shown in our analysis.

For high Neurotics, we note a greater use of more concrete language, which may be describes as entities which can be sensed, for example, *table, spoon, girl*, rather than the abstract references preferred by Extraverts. The psycholinguistic analysis also shows a positive relationship between Neuroticism and Brown Verbal Frequency, and therefore high Neurotics would tend to use more words which are found frequently in speech, for example, *I, and, that*, rather than less common words such as *abject, suspicion, tether*. From the LIWC analysis, we find that high Neurotics make a greater number of first-person references (confirming our Theory hypothesis), and use more inclusive words such as *with, and, include*. These features contribute towards a greater level of repetition in language of high Neurotics (confirming our Theory hypothesis).

Psycholinguistic analysis revealed that high Psychotics use a greater number of unusual words, and also a variety of words found frequently and infrequently in writ-

ing (confirming our Theory hypothesis). Also, from the psychological analysis, we note that they avoid references to themselves (first person singular), and use a greater number of words relating to cognitive mechanisms, such as *cause, know, ought*. This results in them using highly diverse language (confirming our Theory hypothesis).

These analyses have further confirmed hypotheses for Psychoticism and Neuroticism, however we note the relatively few features which are present for Extraversion. This is surprising, given that much of the previous work on personality and language has investigated Extraversion and found many significant features. However, with reference to our hypotheses, we note that many of these are not explicitly measured by any of the analysis techniques which we have used so far. For example, we have the Realisation hypothesis that Extraverts will express their loudness through capital letters and exclamation marks, for Fluency, we expect that they will use elliptical dots (...), or hyphens (-), or for Conversational features that they will use fewer hedges.

We note that this limitation in potential objects of study is a characteristic of top-down analysis techniques. Therefore, in order to study features of Extraversion—and the other personality dimensions—more successfully, we suggest the adoption of data-driven techniques.

4.6 Conclusion

In this chapter we have taken our e-mail personality corpus and the previous factor analysis exploration as our starting point. We then extend this in two ways: by using more appropriate statistical analyses, and also by the analysis of different linguistic features. The first half of the chapter investigated statistical and analytic methods which are better suited to providing a more detailed linguistic description of personality. Using a combination of exploratory correlation and multiple regression analyses, we derived a combination of features which best characterise personality. In our selection of LIWC variables, we discuss a number of implications which affect the generalisability of the results.

In the second part of this chapter we investigated the analysis of additional linguistic features, namely lexical density and a novel technique which measured psycholin-

Extraversion											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	$\odot^{t,g}$	o							
Real.	o	o	o								
Fluency	o	o	o								
Gramm.	o	o	o	$\bullet^{r}$	o	o	o	o			
Conv.	$\odot^{t,g}$	o	o	o	o	o	o	o			
Lexis	o	o	o	o	o	o	o	o	o	o	(−Numbers)^{t,g}; (−Sports; +Affect. proc.)^{a}; −Mean conc.^{m}
Percept.	o	o									

Neuroticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	o	o	$\oplus^{s}$						
Conv.	$\oplus^{r}$	o									
Lexis	$\oplus^{s}$	o	o	o	o						+Inclusive words^{s}; +Mean Conc.^{m}; +Mean Brown verb. freq.^{m}
Percept.	o	o									

Psychoticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	o	o	o	o	$\oplus^{m}$	$\oplus^{r}$	o				
Conv.	o										
Lexis	o	o	o	o	o	o	o	o	o	o	+Cog. mech.^{s}; (+Anger)^{g}; (−Motion; +Certainty; −Feeling; +Sad)^{a,t}; (+Death)^{a}; Std. Dev. of Kucera & Francis writ. Freq.^{m}
Percept.	o	o									

Table 4.16: Review of hypotheses

Note. o indicates an hypothesis; ⊕ confirmation of hypothesis; ⊖ inverse of hypothesis; ⊙ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{a} 'all' LIWC multiple regression; ^{t} topic controlled LIWC regression; ^{g} topic and genre controlled LIWC regression; ^{s} topic, genre and sparsity controlled LIWC regression; ^{m} MRC database analysis; ^{r} Type-token ratio analysis. Please refer to Section 2.6 for a full description of the hypotheses.

guistic properties of language. In comparing these analyses, we found that the psycholinguistic analysis explained the most variance in overall language use for Neuroticism, whereas the psychological properties measured by the LIWC analysis measured a similar amount of linguistic variance for Psychoticism and Neuroticism. Linguistic features characteristic of Extraversion appeared to be most difficult to measure, with features only being revealed by the psycholinguistic analysis. Overall lexical diversity accounted for slightly less linguistic variance. We concluded the chapter with a discussion of the analysis methods used, and related the findings to our hypotheses.

In summary, using a variety of analysis techniques we have shown that Neuroticism and Psychoticism are relatively richly encoded linguistically. These analyses have—on the whole—failed to find linguistic characteristics of Extraversion (with the exception of the psycholinguistic analysis). Since Extraversion appears to be such a ubiquitous, and widely recognised personality variable, this is surprising. We propose that this may be in part due to the top-down method of analysis which we have used. In the next chapter, we therefore investigate other methods and approaches to linguistic analysis which use a bottom-up approach which will allow us to derive characteristic linguistic features directly from the text.

Chapter 5

Data-driven Methods

In the previous chapters we have examined the linguistic features which are characteristic of personality, mainly using top-down analysis measures. These methods, however, may not necessarily provide the most appropriate measures for identifying the linguistic characteristics which are relevant to personality. Indeed, most of these measures were designed with a different purpose in mind.

Instead, in this chapter we focus upon the use of empirical data-driven methods which we use to derive linguistic characteristics of personality from our data. We firstly compare high and low scores on the personality dimensions adopting a method previously used to identify multi-word features from different types of texts; In the second half of the chapter we extend this analysis by adopting additional techniques from computational corpus linguistics which we use to identify a number of different features. We also explore analysis methods which allow us to better trace linguistic behaviour along different groups of personality dimension scorers and we further refine the annotation of our corpus. We conclude the chapter with a summary, and discuss our results in light of our hypotheses.[1]

[1] Work from this chapter is partially reported in Gill and Oberlander (2002), Gill and Oberlander (2003a) and Oberlander and Gill (2004).

5.1 Introduction to N-gram Analysis

At the end of each of the previous two chapters which have examined top-down analysis approaches, we summarised the success of these methods, with particular reference to our personality corpus. Whilst we noted that the content based psychological and psycholinguistic analyses—and also our measure of lexical diversity—were able to relatively successfully identify linguistic features which were characteristic of Neuroticism and Psychoticism, they were unsuccessful for Extraversion (with the exception of the psycholinguistic measures which found a single feature, concreteness). In Chapter 3, which used multi-dimensional analysis methods on the psychological data, the lack of significant findings for Extraversion was similarly apparent.

We therefore observed that one of the main criticisms of content analysis methods is that of topic-specificity, since they are often designed—or at least developed—for a specific purpose, noting the psychological and psycholinguistic bases of the LIWC and MRC psycholinguistic database methods. A further criticism of content analysis methods is one relating to their limited ability to measure the phenomena which they claim to measure, namely their 'recall'. We noted that this was particularly relevant to LIWC given that its much smaller dictionaries lead to a potentially more limited coverage of the lexicon. With particular reference to our hypotheses for the linguistic characteristics of Extraversion, we noted that several of these were likely not to be measured by traditional measures of content. Pennebaker and King (1999) acknowledge a further limitation of content analysis techniques, such as LIWC, which is that they are able to identify *which* words are used, but not *how* they are used, for example, the 'context, irony, sarcasm, or [...] multiple meanings of words' (p. 1297). Although the disambiguation of word senses is less of a problem for the psycholinguistic analysis, since this uses part-of-speech information, contextual information has still been ignored in these analyses.

Therefore, in this chapter instead of top-down approaches, like Tribble (2000) we adopt data-driven techniques from computational corpus linguistics; specifically the analysis of n-grams. This has previously been put a variety of uses (see our review of the literature, in Section 2.5.5.2), however firstly we refer particularly to their use in identify characteristic multi-word terms which distinguish specific types of texts

(Damerau, 1993). Since n-gram analysis is a data-driven approach, it will allow us to identify key features which are characteristic of different personality groups. Additionally, this technique which calculates the probability of groups of adjacent terms, or n-grams, occurring together in a text, enables us to view the probabilistic space in which language occurs. Therefore the n-gram analysis method—unlike the content analysis approaches—retains some of the contextual information of language use and provides us with a greater insight into differences in language structuring and the use of formulaic language, which is related to a number of important functions (e.g., Wray and Perkins, 2000).

We now describe the simplest n-gram analysis, which looks solely at two-word sequences of language. Later in the chapter we explore a variety of more sophisticated techniques.

5.1.1 Method

5.1.1.1 Procedure

In order to study the characteristics of personality language at the extreme high and low ends of the dimensions, the original e-mail corpus of texts was divided into sub-corpora. High and Low personality group samples were achieved by splitting them at greater than 1 standard deviation above and below the mean EPQ-R score for each dimension. The resulting sizes of the sub-corpora are around 12,000 and 8,000 words for the high and low Extraversion groups (gathered from 21 and 17 participants respectively); 12,000 for both the high and low Neuroticism groups (20 and 21 participants); and 8,500 and 10,000 for the high and low Psychoticism groups (18 and 21 participants). (Further information about participants and the subcorpora to which they contribute can be found in Tables A.1 and A.2.)

5.1.1.2 Analysis

Each corpus was tokenised using a simple white space technique, with punctuation retained but separated from adjacent words with 'space' character.

N-gram profiles were generated for each corpus using standard n-gram software.[2] These were then ranked using the log-likelihood statistic (G^2), since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993, see also discussion in Section 2.5.5.2). Rankings for each group are based on the top 50 bigrams with frequency of $N \geq 5$ for these bigrams, and a significance of $p < .001$.

Additionally, in order to examine the relationship between features shared between the high and low groups, relative frequency ratios were also calculated for items common to both sub-corpora. Like Damerau (1993) we calculate the relative-frequency ratio as a measure of feature usage across corpora, which can be interpreted as a likelihood ratio (Manning and Schütze, 1999). We also calculate log-likelihood to allow the comparison of measures (cf. Rayson, 2003), but we do not refer directly to this in our results.

Notice, that here the n-gram analysis and relative frequency ratios are used here to slightly different ends, compared with, for example Damerau (1993), who uses them to distinguish texts on the basis of technical expressions. Rather, it is more similar to that of Milton (1998) who uses these techniques to identify characteristic phrases 'over-used' by English learners when compared to native speakers. Therefore, we do not filter out function words or rarer collocations.

5.1.2 Results

The results are presented as follows: We present the bigram analyses for Extraversion, Neuroticism, and Psychoticism (Tables 5.1, 5.2, and 5.3). In each table we display results (separated by a horizontal rule) for features shared by high and low personality groups, features unique to the high group and then the low group. For the shared results, features are ordered by their relative-frequency ratio, and for the unique results this is by relative frequency.

[2]Ted Pedersen's n-gram and associated statistical software is available from: `http://www.d.umn.edu/~tpederse/code.html`.

Feature	High Freq.	High R.Freq.	Low Freq.	Low R.Freq.	G^2	R.Freq. Ratio
it was	46	0.0034	22	0.0025	1.64	1.39
next week	24	0.0018	12	0.0013	0.66	1.33
a bit	29	0.0022	15	0.0017	0.62	1.28
up with	19	0.0014	10	0.0011	0.36	1.26
! !	45	0.0033	24	0.0027	0.76	1.24
i was	33	0.0025	18	0.0020	0.45	1.22
will be	24	0.0018	13	0.0015	0.35	1.22
to see	32	0.0024	19	0.0021	0.15	1.12
at the	27	0.0020	16	0.0018	0.13	1.12
which is	15	0.0011	9	0.0010	0.06	1.11
for a	34	0.0025	21	0.0024	0.07	1.07
i have	44	0.0033	29	0.0032	0.00	1.01
to get	34	0.0025	23	0.0026	0.01	0.98
. i	99	0.0074	69	0.0077	0.10	0.95
on friday	11	0.0008	8	0.0009	0.04	0.91
, and	48	0.0036	36	0.0040	0.31	0.88
in the	41	0.0030	34	0.0038	0.92	0.80
and then	23	0.0017	19	0.0021	0.50	0.80
apart from	6	0.0004	5	0.0006	0.14	0.80
i am	33	0.0025	28	0.0031	0.91	0.78
i think	16	0.0012	14	0.0016	0.57	0.76
, but	35	0.0026	31	0.0035	1.37	0.75
a lot	10	0.0007	9	0.0010	0.44	0.74
going to	36	0.0027	33	0.0037	1.79	0.72
a few	12	0.0009	11	0.0012	0.60	0.72
to do	23	0.0017	23	0.0026	1.93	0.66
i've been	9	0.0007	12	0.0013	2.54	0.50
. .	152	0.0113	0	0	154.62	-
of the	40	0.0030	0	0	40.69	-
, which	25	0.0019	0	0	25.43	-
had a	22	0.0016	0	0	22.38	-
which was	19	0.0014	0	0	19.33	-
got a	17	0.0013	0	0	17.29	-
new year	18	0.0013	0	0	18.31	-
a good	16	0.0012	0	0	16.28	-
forward to	15	0.0011	0	0	15.26	-
looking forward	15	0.0011	0	0	15.26	-
need to	15	0.0011	0	0	15.26	-
i'll be	14	0.0010	0	0	14.24	-
on saturday	13	0.0010	0	0	13.22	-
as well	11	0.0008	0	0	11.19	-
we went	11	0.0008	0	0	11.19	-
<END> hi	9	0.0007	0	0	9.16	-
able to	9	0.0007	0	0	9.16	-
couple of	10	0.0007	0	0	10.17	-
the moment	10	0.0007	0	0	10.17	-
want to	10	0.0007	0	0	10.17	-
take care	8	0.0006	0	0	8.14	-
catch up	7	0.0005	0	0	7.12	-
other than	6	0.0004	0	0	6.10	-
. <END>	0	0	20	0.0022	36.78	-
i don't	0	0	18	0.0020	33.11	-
went to	0	0	15	0.0017	27.59	-
to go	0	0	14	0.0016	25.75	-
all the	0	0	12	0.0013	22.07	-
i went	0	0	12	0.0013	22.07	-
, because	0	0	11	0.0012	20.23	-
one of	0	0	11	0.0012	20.23	-
i can	0	0	10	0.0011	18.39	-
i'm going	0	0	10	0.0011	18.39	-
trying to	0	0	10	0.0011	18.39	-
don't know	0	0	9	0.0010	16.55	-
i've got	0	0	9	0.0010	16.55	-
on thursday	0	0	9	0.0010	16.55	-
anyway ,	0	0	8	0.0009	14.71	-
lots of	0	0	8	0.0009	14.71	-
this week	0	0	8	0.0009	14.71	-
should be	0	0	7	0.0008	12.87	-
on monday	0	0	6	0.0007	11.04	-
on sunday	0	0	6	0.0007	11.04	-
the pub	0	0	6	0.0007	11.04	-
the same	0	0	6	0.0007	11.04	-
loads of	0	0	5	0.0006	9.20	-
two weeks	0	0	5	0.0006	9.20	-

Table 5.1: Bigram analysis of High and Low Extraverts.

Feature	High Freq.	High R.Freq.	Low Freq.	Low R.Freq.	G^2	R.Freq. Ratio
! !	53	0.0040	23	0.0017	12.00	2.29
i am	39	0.0029	20	0.0015	6.12	1.94
i think	32	0.0024	20	0.0015	2.73	1.59
in the	44	0.0033	28	0.0021	3.50	1.56
, but	59	0.0044	38	0.0028	4.46	1.54
next week	23	0.0017	15	0.0011	1.65	1.52
i have	49	0.0037	33	0.0025	3.05	1.48
for a	37	0.0028	26	0.0019	1.87	1.42
i don't	27	0.0020	20	0.0015	1.01	1.34
a bit	36	0.0027	27	0.0020	1.24	1.33
going to	44	0.0033	34	0.0025	1.23	1.29
want to	15	0.0011	12	0.0009	0.32	1.24
i've got	17	0.0013	14	0.0010	0.27	1.21
to get	30	0.0022	26	0.0019	0.26	1.15
i'm going	14	0.0010	13	0.0010	0.03	1.07
. i	111	0.0083	107	0.0080	0.05	1.03
looking forward	8	0.0006	8	0.0006	0.00	0.99
a few	16	0.0012	17	0.0013	0.04	0.94
to do	28	0.0021	31	0.0023	0.17	0.90
i'll be	10	0.0007	11	0.0008	0.05	0.90
to see	25	0.0019	28	0.0021	0.19	0.89
to be	23	0.0017	27	0.0020	0.34	0.85
of the	32	0.0024	40	0.0030	0.94	0.80
it was	28	0.0021	36	0.0027	1.05	0.77
<END> hi	6	0.0004	8	0.0006	0.30	0.75
at the	24	0.0018	36	0.0027	2.48	0.66
this week	12	0.0009	18	0.0013	1.24	0.66
on saturday	8	0.0006	13	0.0010	1.23	0.61
will be	15	0.0011	28	0.0021	4.07	0.53
. .	204	0.0152	0	0	281.65	-
, and	83	0.0062	0	0	114.59	-
have to	38	0.0028	0	0	52.46	-
to go	35	0.0026	0	0	48.32	-
. <END>	28	0.0021	0	0	38.66	-
and then	17	0.0013	0	0	23.47	-
back to	18	0.0013	0	0	24.85	-
i went	17	0.0013	0	0	23.47	-
i can	15	0.0011	0	0	20.71	-
up with	13	0.0010	0	0	17.95	-
which is	13	0.0010	0	0	17.95	-
i need	12	0.0009	0	0	16.57	-
don't know	11	0.0008	0	0	15.19	-
i've been	11	0.0008	0	0	15.19	-
has been	9	0.0007	0	0	12.43	-
lots of	10	0.0007	0	0	13.81	-
on sunday	9	0.0007	0	0	12.43	-
the end	9	0.0007	0	0	12.43	-
end of	8	0.0006	0	0	11.04	-
my own	7	0.0005	0	0	9.66	-
catch up	6	0.0004	0	0	8.28	-
anyway ,	0	0	22	0.0016	30.62	-
had a	0	0	19	0.0014	26.45	-
a good	0	0	16	0.0012	22.27	-
need to	0	0	16	0.0012	22.27	-
the moment	0	0	16	0.0012	22.27	-
one of	0	0	15	0.0011	20.88	-
a lot	0	0	14	0.0010	19.49	-
couple of	0	0	14	0.0010	19.49	-
which was	0	0	14	0.0010	19.49	-
i haven't	0	0	12	0.0009	16.70	-
to make	0	0	12	0.0009	16.70	-
a couple	0	0	11	0.0008	15.31	-
a while	0	0	11	0.0008	15.31	-
last week	0	0	10	0.0008	13.92	-
new year	0	0	10	0.0008	13.92	-
on friday	0	0	10	0.0008	13.92	-
should be	0	0	11	0.0008	15.31	-
would be	0	0	9	0.0007	12.53	-
apart from	0	0	8	0.0006	11.14	-
as usual	0	0	8	0.0006	11.14	-
at least	0	0	7	0.0005	9.74	-
take care	0	0	5	0.0004	6.96	-

Table 5.2: Bigram analysis of High and Low Neurotics.

Feature	High Freq.	High R.Freq.	Low Freq.	Low R.Freq.	G^2	R.Freq. Ratio
want to	13	0.0014	10	0.0009	1.05	1.53
at the	28	0.0030	22	0.0020	2.06	1.50
a few	13	0.0014	11	0.0010	0.66	1.39
, but	26	0.0028	24	0.0022	0.75	1.28
a bit	25	0.0027	23	0.0021	0.74	1.28
should be	10	0.0011	10	0.0009	0.14	1.18
will be	13	0.0014	14	0.0013	0.06	1.10
<END> hi	8	0.0009	9	0.0008	0.01	1.05
to do	21	0.0022	25	0.0023	0.00	0.99
in the	31	0.0033	39	0.0035	0.07	0.94
. <END>	16	0.0017	21	0.0019	0.10	0.90
this week	9	0.0010	12	0.0011	0.08	0.88
to see	16	0.0017	22	0.0020	0.22	0.86
i have	23	0.0025	35	0.0032	0.91	0.78
to go	15	0.0016	23	0.0021	0.63	0.77
. i	57	0.0061	95	0.0086	4.35	0.71
it was	18	0.0019	30	0.0027	1.37	0.71
to get	17	0.0018	30	0.0027	1.81	0.67
next week	8	0.0009	17	0.0015	2.00	0.56
i think	10	0.0011	22	0.0020	2.85	0.54
for a	17	0.0018	38	0.0034	5.13	0.53
going to	16	0.0017	40	0.0036	7.04	0.47
. .	86	0.0092	0	0	134.04	-
i was	22	0.0024	0	0	34.29	-
for the	21	0.0022	0	0	32.73	-
on the	20	0.0021	0	0	31.17	-
had a	17	0.0018	0	0	26.50	-
bit of	14	0.0015	0	0	21.82	-
i should	13	0.0014	0	0	20.26	-
i can	12	0.0013	0	0	18.70	-
a good	11	0.0012	0	0	17.14	-
anyway ,	11	0.0012	0	0	17.14	-
i'm not	11	0.0012	0	0	17.14	-
up with	11	0.0012	0	0	17.14	-
the moment	10	0.0011	0	0	15.59	-
which was	10	0.0011	0	0	15.59	-
some work	9	0.0010	0	0	14.03	-
apart from	8	0.0009	0	0	12.47	-
catch up	8	0.0009	0	0	12.47	-
know what	8	0.0009	0	0	12.47	-
lots of	8	0.0009	0	0	12.47	-
since i	8	0.0009	0	0	12.47	-
trying to	8	0.0009	0	0	12.47	-
don't know	7	0.0007	0	0	10.91	-
has been	7	0.0007	0	0	10.91	-
i suppose	7	0.0007	0	0	10.91	-
ended up	6	0.0006	0	0	9.35	-
be able	5	0.0005	0	0	7.79	-
new year	5	0.0005	0	0	7.79	-
thought i'd	5	0.0005	0	0	7.79	-
, and	0	0	59	0.0053	72.43	-
! !	0	0	45	0.0041	55.25	-
i am	0	0	29	0.0026	35.60	-
to be	0	0	25	0.0023	30.69	-
on saturday	0	0	24	0.0022	29.46	-
and then	0	0	21	0.0019	25.78	-
back to	0	0	21	0.0019	25.78	-
i had	0	0	21	0.0019	25.78	-
i've got	0	0	20	0.0018	24.55	-
i went	0	0	18	0.0016	22.10	-
went to	0	0	18	0.0016	22.10	-
i don't	0	0	17	0.0015	20.87	-
have been	0	0	15	0.0014	18.42	-
i haven't	0	0	15	0.0014	18.42	-
i've been	0	0	15	0.0014	18.42	-
i'm going	0	0	13	0.0012	15.96	-
one of	0	0	13	0.0012	15.96	-
as well	0	0	11	0.0010	13.50	-
how are	0	0	11	0.0010	13.50	-
managed to	0	0	11	0.0010	13.50	-
on sunday	0	0	11	0.0010	13.50	-
a lot	0	0	10	0.0009	12.28	-
end of	0	0	10	0.0009	12.28	-
couple of	0	0	9	0.0008	11.05	-
looking forward	0	0	7	0.0006	8.59	-
<END> hello	0	0	6	0.0005	7.37	-
new year's	0	0	5	0.0005	6.14	-
take care	0	0	5	0.0005	6.14	-

Table 5.3: Bigram analysis of High and Low Psychotics..

5.1.2.1 Categorisation of Linguistic Features

On the basis of the linguistic features resulting from the bigram analysis, we have divided them into categories to aid comprehension. This is in contrast to the groupings of top-down analysis techniques, since we have grouped features into categories on the basis of the data. We outline these categories below. Following these descriptions, we then describe how these features relate to the personality dimensions of Extraversion, Neuroticism and Psychoticism.

Surface Realisation Features These gross features are perhaps the most intuitive in their representation of authorial personality. Indeed, it is possible that these provide the primary information in such impression formation. For example, [<END> *hi*], the <END> (end-of-message marker) followed by *hi* indicates message-initial *hi* (since the <END> marker separates concatenated files in the corpus). Conversely there is the message-initial 'hello' ([<END> *hello*]) which indicates greater formality. The use of [. <END>] indicates that the author has ended their message with a full stop, again a possible mark of greater formality. Additionally [. *i*] indicates the extraposition of the first-person singular pronoun to the start of a sentence, or to a position after the elliptical (...) which may occur mid-sentence. Use of punctuation in non-standard ways may also be an indicator of informality, and represent a more colloquial 'e-mail style', e.g., multiple exclamation marks [*! !*], or multiple full stops [. .] as in the elliptical (...) (Baron, 1998; Colley and Todd, 2002).

Quantification Patterns involving quantification are also apparent from the bigrams. In some cases this refers to large amounts in an exaggerated manner suggestive of hyperbole, for example, [*a lot*], [*lots of*], [*loads of*] and [*all the*]. This can be contrasted with very specific and precise expressions such as [*one of*], and more restrained, and perhaps slightly understated expressions, like [*a bit*], [*bit of*], [*a few*], [*few drinks*], [*some of*] and [*couple of*].

Social Devices The use of formulaic or stylistic expressions as social devices here generally indicate a relaxed and informal style, and their omission may point to a more careful or reserved author (cf. Wray and Perkins, 2000). For example, the use of [*catch up*] indicates a desire for potential future interaction and [*take care*] implies

a concern for the other person's well-being. Additionally, other expressions such as [*other than*] and [*apart from*] appear to be used as way of summarising the message and bringing it to a conclusion.

Self/Other Reference References to self along with others may be demonstrated through the use of the first person plural *we* ([*we went*]), or through first person singular pronoun *I*. The latter singular form, as discussed in Surface Realisations, may involve the positioning of the first-person pronoun in a sentence, as in [. *i*]. This first person singular form also shows greater occurrence in the e-mail data, referring to events or states in the past ([*i was*], [*i had*], [*i went*], [*i've done*], [*i've been*]) or present ([*i am*], [*i've got*], [*i don't*], [*i'm not*], [*i think*], [*i have*], [*i haven't*]), and abilities or desires for the future ([*i'll be*], [*i'm going*], [*i can*], [*i should*], [*i will*]). There is also evidence of *I* being used in relation to expressions functioning as Social Devices ([*i suppose*], [*thought i'd*]), which brings the author and their opinion to the fore. Additional references to others may take the form of explicit reference, or through other means such as [*up with*] which in many cases indicates a shared experience (prompting the question *with whom?*).

Valence Evidence of negative valence is found in e-mails through the use of negation, such as [*i don't*], [*don't know*], [*i'm not*] and [*i haven't*], whereas use of [*a good*] is suggestive of positive affect. Also, bigrams such as [*looking forward*] and [*forward to*] (presumably as in *looking forward to*) which we have also regarded as a Social Device and Temporal References are also suggestive of optimism.

Ability Personal views on capability are suggested by the different collocations (or *colligations;* Hunston, 2002) with infinitival *to*.[3] Emphasis of one's ability to do something, should they choose, can be confidently and assertively relayed using *want–* and *able– (to)*. By contrast, a more timid and tentative disposition of intent to perform something can be expressed as [*trying to*] do something. Use of the bigram [*have to*] suggests an external, rather than internal, locus of obligation which would more likely be suggested by [*need to*]. If the intention is to undertake some activity, this can be expressed as [*going to*] do it, and if this has been successfully accomplished—possibly

[3]This confirms the usefulness (for current purposes) of retaining functors usually filtered out by a stop list, since it allows us to study colligation or 'upward' collocations (Sinclair, 1991).

with a bit of effort—they have [*managed to*] do it.

Modality Similarly, collocations or colligations with the verb *be* show a distinction in use of modal auxiliaries which has an effect on the projection of certainty. For example, the weaker and more tentative *should be* is contrast with the more strongly predictive [*will be*] and its contracted form [*i'll be*] (*i will be*) (Coates, 1983). This contrast is also present between, for example, the bigrams [*i should*] and [*i can*].

Message Planning/Expression Indicators of grammatical construction can be found in the use of connectives: for example, co-ordinating conjunctions such as [, *and*] and [, *but*], or the use of the subordinating [, *which*]. The use of conjunctive adverb [, *anyway*] should also be noted, although this may be used in a similar way to the Social Devices [*other than*] and [*apart from*] which function to bring the message to a conclusion.

Temporal References The bigram analysis reveals temporal expressions, in the reference to specific periods, for example days ([*on saturday*], [*on friday*], [*on sunday*]), or larger time measurements ([*last week*], [*this week*], [*next week*], [*new year*], and [*new year's*] presumably as in *new year's eve*). More general and vague references to time are found in bigrams such as [*the moment*], [*a while*], and [*a bit*]. Additional expressions which encode the author's feelings toward the period are found in phrases such as [*as usual*], [*looking forward*].

5.1.2.2 Extraversion

The distribution of bigrams relevant to these features in relation to Extraversion is presented in Table 5.4. The Surface Realisation Features reveal that Introverts and Extraverts differ at a gross level, with the latter group using message-initial *hi* exclusiviely. Similarly, use of punctuation also differs between the two groups, with Extraverts preferring multiple exclamation marks [*! !*], and solely using multiple full stops [. .] suggestive of the elliptical (...), again a feature of informal style, and 'looser' use of language.

In terms of quantification, Introverts generally tend to show a preference for a greater use of quantifiers, particularly those suggesting exaggeration ([*all the*], [*lots of*]

Category	High E	←	→	Low E
Surface Realisation	<END> hi . .	! !		. <END>
Quantification	couple of	a bit	a lot a few	all the one of lots of loads of
Social Devices	looking forward catch up take care			
Self/other Reference	i'll be we went up with	i was i will	i think i am . i	i don't i went i'm going i can i've got
Valence	a good			i don't don't know
Ability	want to need to able to		going to	trying to
Modality	will be i'll be			should be
Message Planning	other than , which		apart from , and , but	
Temporal Reference	on saturday new year looking forward forward to the moment	next week	on friday	this week two weeks on monday on sunday on thursday

Table 5.4: High–Low Extravert Bigram Features.

Category	High N	←	→	Low N
Surface Features	. . . <END>	! ! . i	<END>hi	
Quantification	lots of		a few	one of couple of a lot a couple a while at least
Social Devices	catch up		looking forward	take care
Self/other Reference	i went i can i need i've been my own up with	i am i think i have i don't i've got i'm going . i	i'll be	i haven't
Valence		i don't		a good i haven't
Ability	have to to go	going to want to	to do to see to be	need to to make
Modality			i'll be will be	should be
Message Planning	, and	, but		anyway ,
Temporal Reference	on sunday	next week a bit	this week looking forward on saturday	last week on friday new year as usual the moment a while

Table 5.5: High–Low Neuroticism Bigram Features.

Category	High P	←	→	Low P
Surface Features	. . 	 <END> hi	. <END> 	! ! <END> hello
Quantification	bit of lots of	a few a bit		one of a lot couple of
Social Devices	catch up			looking forward take care
Self/other Reference	i was i can i suppose i should i'm not thought i'd up with		i have i think . i	i am i had i've got i don't i'm going i went i've been i haven't
Valence	i'm not don't know			i haven't i don't
Ability	trying to	want to	to do to see to get to go going to	to be i've got managed to
Modality	i should i can has been be able	should be will be		
Message Planning	anyway ,	, but		, and
Temporal Reference	new year		next week	on saturday on sunday looking forward new year's

Table 5.6: High–Low Psychoticism Bigram Features.

and [*loads of*]) or specificity [*one of*], whereas Extraverts show a preference for [*a bit*] and uniquely use [*couple of*].

The Extravert use of the Social Devices [*catch up*] and [*take care*] indicate a relaxed and informal style; their omission points to a more socially restrained Introvert. The Extravert use of [*other than*] perhaps as a method of summary is in contrast with the Introvert [*apart from*].

References to self in the texts demonstrate differences between Extraverts and Introverts: Introverts make extensive use of the first person singular pronoun ([*i don't*], [*i went*], [*i'm going*], [*i can*], [*i've got*] are all unique to the Introvert text), and also show preference for the following shared bigrams: [*i've done*], [*i think*], [*i am*]. For Extraverts, the only unique first person bigram is [*i'll be*], and they also show greater use of [*i was*] and [*i will*], although relatively less preferred than Introvert forms. This underscores the increased Introvert tendency to focus on self, possibly also by use of extraposition ([. *i*]), whereas Extraverts suggest interaction with others: the unique use of a bigram containing a first person plural ([*we went*]), along with [*up with*] indicative of a shared experience. These results apparently contradict Furnham (1990) on pronouns, but given that the vast majority of pronouns here are first-person singular, and thus focusing on self, this is unsurprising.

Bigrams containing negations were used mainly by Introverts, as in [*i don't*] and [*don't know*] (indeed [*i don't*] is the bigram with most frequent use of *i*), whilst Extraverts used the bigram [*a good*] which is suggestive of positive affect.[4] Similarly, the Extravert preference for [*looking forward*] and [*forward to*] (presumably as in *looking forward to*) also suggests a more positive disposition.

Confident Extravert views on personal ability are suggested by their unique use of [*want to*] and [*able to*], whereas Introverts use the more tentative forms [*trying to*] or [*going to*]. Such patterns are also reflected in use of modal auxiliaries with Extraverts preferring the stronger forms [*will be*], and are unique in their use of the contracted form [*i'll be*] (*i will be*), and the Introvert use of the weaker *should be*.

[4]Further investigation shows that *good* is not directly negated (as in [*not good*]). Compare the Introvert [*i can*], which was generally followed by *not*. Although the effect of negation was not viewed as important by Pennebaker and collaborators in the functioning of LIWC (Berry et al., 1997; Pennebaker and Francis, 1999), it certainly has implications for models of syntax and semantics.

Differences can also be found in Message Planning, with Introverts showing preference for the co-ordinating conjunctions [, *and*] and [, *but*], whilst Extraverts uniquely use the subordinating [, *which*], usually in an evaluative sense. Evidence of the looser Extravert style is also demonstrated through their use of Temporal Referents, preferring the less specific expressions, such as [*next week*], [*new year*], or [*the moment*] to more specific references to days or periods of time ([*on monday*], [*on thursday*], [*two weeks*], [*this week*]).

5.1.2.3 Neuroticism

The distribution of features and their bigrams relevant to Neuroticism is presented in Table 5.5. High Neuroticism is demonstrated through the unique use of Surface Realisations such as multiple full stops [. .], and ending the message with a full stop [. <END>]. There is also a High Neurotic tendency to use multiple exclamation marks [*! !*], and to start sentences with the self referent "I" [. *i*]. Low Neurotics, on the other hand showed a preference for starting their texts with the informal greeting "hi" [<END> *hi*]. Here we also note the reference to specific topics, and we note that High Neurotics make reference to [*exam results*] exclusively.

Although High Neurotic texts demonstrate a unique use of the feature suggestive of exaggeration [*lots of*], Low Neurotic texts apparently show greater use of many different quantifying references, preferring [*a few*], and uniquely using [*one of*], [*couple of*], [*a couple*], [*a lot*] and [*at least*] (however, this latter feature may be used as a Social Device).

In the use of social devices, High Neurotics focus more on the potential future interaction, uniquely using [*catch up*], however Low Neurotics are more likely to use the form [*looking forward*] which expresses positive affect with regard to the future meeting, and also solely use [*take care*], implying a concern for the other person's well-being.

Use of terms of self reference in the texts point to differences between the language use of High and Low Neurotics: High Neurotics show a much greater concern for self, and thus increased use of bigrams which include the first-person singular "I" as in, [*i went*], [*i can*], [*i need*], and [*i've been*] which are used uniquely, and a preference for

[*i am*], [*i think*], [*i have*], [*i don't*], [*i've got*], all of which suggest an interest in current or past events, with the exception of [*i'm going*]. Their preference for [. *i*] suggests that High Neurotics promote their main focus of interest—themselves—to the start of the sentence for emphasis.

Although the bigrams of both High and Low Neurotics indicate use of contracted negation (High N preferring [*i don't*], Low N using [*i haven't*] uniquely), the Low Neurotic texts also show the unique use of the positive evaluative term [*a good*], which along with the preference for [*looking forward*] is suggestive of greater positive affect.

Collocations with the infinitival *to* give an indication of the High and Low Neurotics' perceptions of personal ability and obligation: Whilst [*have to*] is unique to High Neurotics, and is suggestive of an external locus of obligation, [*need to*] is only found in Low Neurotic texts which is much more indicative of an internal desire to do something. The shared bigrams [*going to*] and [*want to*] are both preferred by Higher Neurotics.

Indications of the projection of certainty are shown in expressions of Modality. Low Neurotics show a preference for the stronger predictive *will* ([*will be*], [*i'll be*]), although they show a unique use of the weaker and more tentative *should be*. The preference for the use of modals appears to be the preserve of the Low Neurotics, since High Neurotics do not use any uniquely, with those which are shared, strongly preferred by the Low Neurotics.

In terms of Message Planning and Expression, High Neurotics show preference for the co-ordinating conjunctions [, *and*] and [, *but*], whilst Low Neurotics uniquely use the conjunctive adverb [, *anyway*].

A wide variety of Temporal References show much greater use by Low Neurotics, for example, uniquely using [*on saturday*], [*on friday*], [*last week*], [*new year*], [*the moment*], [*a while*], and [*as usual*], and showing a preference for [*this week*] and [*looking forward*]. By contrast, High Neurotics only used [*on sunday*] uniquely, and showed preference for [*next week*] and [*a bit*].

5.1.2.4 Psychoticism

The distribution of these features and bigrams in relation to Psychoticism is presented in Table 5.6. Degree of Psychoticism of texts is distinguished by Surface Realisation Features, for example, High Psychotics show a unique use of multiple full stops [. .], whilst Low Psychotics show a preference start sentences with "I" [. *i*], to end their message with a full stop [. <END>], and use multiple exclamation marks [*! !*] uniquely. When starting their texts, High Psychotics showed a preference for the greeting "hi" [<END> *hi*], with Low Psychotics uniquely using the more formal "hello" [<END> *hello*]. However, the degree of Psychoticism in texts shows relatively little relation to the use of quantifiers, although High Psychoticism appears to be related to the bigrams [*a few*] and [*few drinks*], apparently pointing to the phrase *a few drinks*.

In their use of Social Devices, High Psychotics focus more on the potential future interaction, using the term [*catch up*] uniquely, whilst Low Psychotics express an attitude towards the future ([*looking forward*]) and tend to show more of a concern for the other person ([*take care*]). When the High and Low Psychoticism texts are compared for social references, Low Psychotics show a greater overall preference for use of first-person pronoun, for example [*i have*], [*i think*], and [. *i*], and also in the unique use of [*i am*], [*i had*], [*i've got*], [*i don't*], [*i'm going*], [*i went*], [*i've been*], and [*i haven't*]. High Psychotics, did however still make substantial social references, uniquely using: [*i was*], [*i can*], [*i suppose*], [*i should*], [*i'm not*], [*thought i'd*], and also the allusion to other social actors using [*up with*].

Both High and Low Psychotic texts show contracted negations which are used uniquely, with the former using [*i'm not*] and [*don't know*], and the latter [*i haven't*] and [*i don't*]. Bigrams containing positive affect words did not appear in the analysis for either group.

Ability represented through collocations with infinitival *to* demonstrate a desire or attempt to achieve something on the part of the High Psychotics in the respective bigrams [*want to*] and [*trying to*] (the latter used exclusively). Low Psychotics show a preference for the more certain [*going to*], and unique use of [*managed to*] which indicates successful accomplishment. Similarly, the High Psychotics also show preference for the modal auxiliaries [*should be*] and [*will be*] and also the exclusive use of

[*i should*] and [*i can*].

High Psychotics use the conjunctive adverb [, *anyway*] uniquely, and also show preference for the co-ordinating conjunction [, *but*], whilst Low Psychotics uniquely use the co-ordinating conjunction [, *and*].

References to time appear a predominant feature of Low Psychotic texts, showing a tendency for using the term [*next week*], and the unique use of [*on saturday*], [*on sunday*], [*looking forward*], and [*new year's*] (presumably as in *new year's eve*). In contrast the bigram [*new year*] was solely used exclusively by High Psychotics.

5.1.3 Summary of Bigram Findings

On the basis of our findings for the bigram analysis, we propose the following features to be indicative of personality: For Extraversion, we propose that higher Extraverts use more informal Surface Realisations and Social Devices, more positive Valence words, and refer to their ability and the future with more confidence and certainty. They are also more likely to refer to other people. Introverts make extensive reference to themselves, especially using extraposition—often at the expense of the topic of the communication, use more quantification, especially for the purpose of exaggeration or specificity, and use more negations. Introverts ended to use more coordination features than the Extraverts.

In the case of Neuroticism we propose that High Neurotics make distinctive use of punctuation in their Surface Realisations, use Social Devices to focus on potential interaction rather than for positive affect or concern for others. In terms of Valence, positive affect is avoided, and more negations are used. Extraposition and extensive use of self-referents focuses the interaction on themselves, whilst Quantifiers are used for exaggeration, and obligation is emphasised rather than Ability. Low Neurotics use more informal Surface Realisation, use a greater variety of Quantifiers, show greater use of Modals, especially indicating stronger prediction, and also greater use of Temporal Referents. Here we find that High Neurotics use more co-ordination features.

For Psychoticism, High and Low groups are distinguished by their use of punctuation and formality of Surface Realisation features. High Psychotics focus on future interactions rather than positive affect or concern for others, and demonstrate a greater

expression of Modality. Low Psychotics show a greater use of self reference, express their Ability in terms of intention or accomplishment, rather than striving to achieve something, and make greater use of Temporal reference. The two groups demonstrated significant use of negation and co-ordination, with the High Psychotics showing preference for contrasting co-ordination.

5.2 3D Distribution of Personality Bigram Features

Whilst so far the analysis has addressed each personality trait in turn, it is apparent that many of these features are not exclusive to one particular personality dimension (compare Tables 5.4, 5.5). and 5.5.

We therefore present the previous bigram findings for Extraversion, Neuroticism and Psychoticism together tabularly so as to represent the three-dimensional personality space. In each case the horizontal x-axis represents the Neuroticism scale, and the vertical y-axis denotes varying degrees of Psychoticism. Separate tables are used for variation of the Extraversion scale: Table 5.7 displays low Extravert features, Table 5.8 the neutral Extravert features, and Table 5.9 the high Extravert features.

Each personality dimension has been divided into 'high', 'neutral', and 'low' features which represents a simplification of the bigram and relative frequency ratio analyses: In these representations, a feature is regarded as being characteristic of the high end of the trait, if it either occurs uniquely in the sub-corpora of that particular trait, or if it is shared, then it shows greater usage by the high personality trait group. Conversely, bigrams are regarded as features of a low personality trait group if they either occur uniquely to that sub-corpus, or if shared by high and low groups are used more by the low trait group. Bigram features are regarded as 'neutral' if they do not feature in our analysis for that dimension.

In discussing these results we will first examine what features are preserved as characteristic solely of the High–Low language of a particular dimension, whilst remaining neutral on the rest. We will then discuss what patterns emerge from the combination of different dimensions.

	Low N	Neutral N	High N
High P	should be apart from anyway , a few	trying to	lots of i can don't know , but
Neutral P	on friday	two weeks the same the pub on thursday on monday loads of all the , because	
Low P	to do this week one of a lot	went to	to go to get on sunday in the i've got i've been i'm going i went i think i don't i am going to and then . i . <END> , and

Table 5.7: 3D Personality Bigrams, Low Extraversion.

	Low N	Neutral N	High N
High P		thought i'd some work since i on the know what i'm not i suppose i should for the ended up bit of be able	has been
Neutral P	would be to make last week at least as usual a while a couple		the end my own i need have to
Low P	to be i haven't	new year's managed to i had how are have been <END> hello	end of back to

Table 5.8: 3D Personality Bigrams, Neutral Extraversion.

	Low N	Neutral N	High N
High P	will be	i was	want to
	which was		up with
	the moment		catch up
	new year		a bit
	had a		. .
	at the		
	a good		
	<END> hi		
Neutral P	of the	we went	which is
	need to	other than	
	i'll be	got a	
		forward to	
		able to	
		, which	
Low P	to see	as well	next week
	take care		i have
	on saturday		for a
	looking forward		! !
	it was		
	couple of		

Table 5.9: 3D Personality Bigrams, High Extraversion.

5.2.1 Features of the Personality Dimensions

From the bigram analysis, we find that Introverts (Table 5.7) are characterised by Temporal references ([*two weeks*], [*on thursday*]) and exaggerating Quantification ([*loads of*], [*all the*]), whereas the Extraverts (Table 5.9) make reference to others using the first-person plural pronoun *we*, assert their confidence and ability using [*able to*], use the Expression and Social Device features [, *which*] and [*other than*], and suggest positive affect with [*forward to*] (as in *looking forward to*).

The results for Neuroticism bigrams, independent of other dimensions can be found in Table 5.8. Here, the relatively few features point to self reference ([*my own*], [*i need*]), obligated ability ([*have to*]), and very ominously a reference to [*the end*]) for high scorers on this trait. The Low Neurotic features reveal tentative modality ([*would be*]), Temporal Reference ([*last week*], [*a while*], [*as usual*]), and loose Quantification ([*at least*], [*a couple*]).

The High Psychotic bigram features (independent of other dimensions; Table 5.8) show some self-reference, which also includes negation and tentative modality ([*i'm not*], [*i should*]), and vague Quantification references ([*few drinks*], [*bit of*])

The Low Psychotics show self—and explicit other—reference ([*that i have*], [*i had*], [*<END> hi there*], also showing interest in the other person [*how are you*]), and Temporal Reference ([*new year's*][*a while .*]). Striving references to ability are used ([*managed to*]), with also evidence for distinctive use of message final punctuation ([*! ! <END>*]).

5.2.2 Interaction between Personality Dimensions

We now turn to examine the effects of interaction between different personality dimensions upon language use. For example, here we can see that for High or Low Extraversion the most characteristic features are shown when interaction occurs with the extremes of Psychoticism or Neuroticism (here it should again be noted that here 'neutral' means that the feature did not appear in the data for the dimension).

Low Extravert and Low Psychotics (Tables 5.7), especially in combination with High Neuroticism appear to use a features containing a great deal of self-reference, in

combination with expression of obligation or negation ([*have to*], [*i don't*]) (although expressions of ability, [*going to*], do not seem stable across the trait of Neuroticism). The Surface Realisation features ([. <*END*>]) and ([. *i*]) both appear to be stable features of High Neuroticism, but occur with different locations on the other scales.

Similarly, other Surface Realisation and Social Device features are High Extravert and highly Neurotic (Table 5.9), but are distinguished as Low Psychotic ([*! !*] or High Psychotic ([. .], [*catch up*]). For expressions of greeting, such as ([<*END*> *hi*] vs [<*END*> *hello*]) it is the degree of Psychoticism and Extraversion which are most important: the former is high on both dimensions, the latter is low on both.

For expressions relating to positive affect ([*looking forward*]), or expressing interest in the other person's well being ([*take care*]), High Extraversion and Low Psychoticism appear important.

5.2.3 Summary of High–Low corpus comparison

In the bigram analysis we divided the corpus into the groups of extreme personality, by each dimension in turn. In this way we treated each dimension independently, since they are viewed as being located orthogonally in personality space. The analysis methods which we then applied to these high and low groups were derived from Damerau (1993) who had previously used relative-frequency ratio to find key phrases related to specialist text-types. We finally combined the results for Extraversion, Neuroticism and Psychoticism tabularly to provide a more coherent view of the distribution of features across the different dimensions. However, we note that our separation of the personality dimensions by high and low groups does not allow comparison with individuals in the mid-section of the dimension. Additionally, these dimensions may not be entirely independent.

We also note that our purpose is slightly different to that of Damerau (1993), and it may be more appropriate to analyse a greater number and variety of n-gram features, rather that restricting this to a few of the most highly collocating features. Additionally, in such analysis, the calculation of G^2 would be a more appropriate measure for distinguishing feature usage, rather than the use of relative-frequency ratio.

5.3 Stratified Corpus Comparison

In the previous discussion we raised several methodological points. Therefore, to address the first issue of feature and personality independence, we utilise techniques from comparative corpus linguistics, and define a 'reference corpus' from authors with a personality profile which is not extreme on any of the measured dimensions. We can then compare authors from each of the extreme personality groups with this 'neutral' (here termed 'mid') group. Furthermore, to control for individuals who may be extreme on more than one dimension, we also ensure that authors representative of the extreme groups are measured as being 'neutral' on the other dimensions.

However, unlike other corpus comparison studies, this gives us a three-way corpus comparison, so we are able to trace the behaviour of linguistic features over the breadth of a personality dimension (divided into high–mid–low categories). Generally other studies have divided the data using binary categories, such as native/non-native or young/old language users, or those of higher/lower socio-economic class (Milton, 1998; Granger and Rayson, 1998; Aarts and Granger, 1998; Rayson and Hodges, 1997).

Like other corpus comparison studies, here we also measure unigram features in addition to the bigrams measured on the previous analysis, and 3–5-grams. Like these studies, we also calculate the statistical significance of the use of these features across corpora.

5.3.1 Method

5.3.1.1 Procedure

Similarly to the the previous n-gram analysis, the original e-mail corpus of texts was divided into subcorpora. High and Low personality group samples were again achieved by splitting them at greater than 1 standard deviation above and below the mean EPQ-R score for each dimension. However, in this case the additional requirement was made that authors had to be *within* 1 standard deviation on the dimensions other than the one for which they were extremely high or low.

Additionally, all texts which were within 1 standard deviation across *all* personality

dimensions were assigned to the personality 'neutral' Mid subcorpus.

The resulting sizes of the subcorpora are around 6,000 words for the high Extraversion, and over 2,000 words for the low Extraversion groups (produced by 11 and 4 authors, respectively); Just over 3,000 words for the high Neurotic and around 6,000 words for the low Neurotic groups (6 and 9 authors); high and low Psychotics were both around 4,000 words (9 authors each); the Neutral group was around 5,000 words (23 authors). These word counts for the respective constrained subcorpora are therefore smaller than for the unconstrained analyses which were in the region of 8,500 to 12,000 words. (Further information about participants and the subcorpora to which they contribute can be found in Tables A.1 and A.2.)

5.3.1.2 Analysis

In order to address some of the issues which were raised regarding the previous analysis, and also to take account of the differently prepared corpus, there are a number of differences in the analysis procedure which we use. These are: First, in this analysis we use a version of the corpus which has been tokenised using the CLAWS tagger (available via Wmatrix tool; Rayson, 2003). This is an advancement over the simplistic white-space metric which was used in the previous analysis: here, additionally, the tokenisation splits multi-word units, into their constituent parts, for example *can't* will be divided into *ca* and *n't*, and also provides some basic annotation, for example marking clause boundaries (represented here as <NC>, and which is generally equivalent to the start or end of a sentence), and ellipsis (<E>); Secondly we calculate 1–5 word n-grams, but do not use a rank or frequency cut-off during calculation (previously, this had been limited to the top 50 ranked features with a frequency ≥ 5; here however, we present all features achieving a frequency of $\geq$5). This gives a broader picture of the features which are used overall, and enables a more accurate log-likelihood statistic of their occurrence between groups to be calculated; Third, we use a lower significance level in the measurement of collocation where it is available. Using Pedersen's n-gram software (Banerjee and Pedersen, 2003), here we select a log-likelihood significance value of $p \leq 0.01$, rather than $p \leq 0.001$, for the bigrams and trigrams (since this suite does not calculate significance for 4- and 5-gram collocations, we include all of these

features in our analysis, but approach interpretation with greater caution); Fourth, we now need to make a three-way comparison of the linguistic features across the high-mid-low corpora for each group. To do this, we again use a program to sort the two data files and then output features which are unique to each of these data files, and those which are common to both. For each personality dimension, the initial stage is to compare the High and Low corpus groups, however this approach differs from the previous analysis, because a second stage of analysis is then performed. This then compares, in turn, the features which are unique to and shared between these High and Low groups with the Mid group. From these analyses, we are then able to calculate the relationships between the three groups, and for each feature in each corpus we identify its frequency and relative frequency, and then where relevant, its relative-frequency ratio and log-likelihood between High-Low, High-Mid and Low-Mid groups. This allows us to compare the relative usage and statistical significance of the difference in the use of features between groups.

5.3.2 Results

Here we present the results from the three-way analysis tabularly, specifying that the feature should exhibit a frequency in one of the three groups of at least 5 occurrences, and these are ordered by log-likelihood (G^2) value. Rayson (2003), in his evaluation of the G^2 test, regards the 15.13 critical value as equivalent to $p \leq 0.01$ significance (and thus 10.83 equivalent to $p \leq 0.01$, etc.) when carrying out multiple comparisons of language. This is instead of the critical value of 6.64 normally used to indicate $p \leq 0.01$ for the χ^2 distribution. As we note in our discussion of these measures (Section 2.5.5.1), Rayson raises the critical value to account for lowering the Cochran rule which enables the comparison of expected frequencies of 1 or more (Cochran, 1954). Since we only examine expected frequencies of 5 or more—which compare more reliably with the χ^2 distribution—we regard Rayson's suggestion of the 15.13 critical value for $p \leq 0.01$ to be particularly conservative. Therefore in our presentation of results we display features with lower critical values, in this case 10.83 or greater. In reporting these—and subsequent findings—we annotate the levels of significance generally associated with these critical values, for example, $p \leq 0.05$ with 3.84, $p \leq$

0.01 with 6.64, etc., and separate these levels with a rule. Note that if a feature is overused by the Mid group, we do not report the G^2 for this, and in cases where the relative-frequency ratio or G^2 is not available, we replace this by '-'.

The results for Extraversion, Neuroticism, and Psychoticism are found in Tables 5.10, 5.11, and 5.12. In the subsequent presentation of the results, we will draw attention to features which are characteristic of the High or Low groups, compared with the usage of the feature more generally. In doing this we distinguish whether a feature is under- or over-used by one extreme group in particular relative to the two other groups, or whether the extreme groups both differ with respect to the Mid group. This will be done for each personality group in turn, with these results then discussed in context with the features of the other groups.

Turning first to the results for Extraversion (Table 5.10). These show that the most significant features for High Extraverts are an overuse of the features [*was a*], [*it .*], [*get to*], [*i really*], and [*was*], and an under-use of ellipsis (<E>); for the Low Extraverts (Introverts) the most significant features are an overuse of [*played*], [*i played*], and [*bread*], along with the under-use of [*, and*], [*! !*], [*i 'll*], and [*, i*]; High and Low groups both differed from the Mid in their use of fewer [*!* <NC>], and—particularly in the case of the High Extraverts—increased use of [*got a*], and this pattern is also found in the following features with a lower significance: [*it 'll*], [*then i*], [*year .*], [*) and*], and [*it 'll be*]. Patterns for lower significance High Extravert features are reduced use of [*fairly*], [*, although*], hyphen or dash [-], and [*which was*], and increased use of [*'ll have*], [*from the*], [*of it*], [*what i*], and newclause (<NC>); Introvert features are reduced use of [*i have*], and increased use of [*supposed to be*] (which is also reflected in its constituent elements), and the name [*jim*].[5]

The most significant Neuroticism results (Table 5.11) for the High group show an under-use of [*and*], and sentence-initial *it* ([<NC> *it*]), and a large overuse of punctuation features ellipses (<E>), especially in combination with structural or contextual information, ([<E> <NC>], [<E> <NC> *i*]), and multiple exclamation marks (between two and five repetitions, [*! ! ! ! !*]), and also [*film*], [*the film*], and [*well i*]. The Low Neuroticism group show an overuse of the name [*dave*], as well as [*, as*],

[5]Note that names reported in the data have been anonymised.

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
<E>	1	45	0.0063	18	0.0016	0	0	4.04	-	-	28.49****	-	28.52****	+		
played	2	0	0	2	0.0002	10	0.0037	-	21.67	-	-	23.50****	26.07****			+
i played	3	0	0	0	0	7	0.0026	-	-	-	-	23.44****	18.25****			+
got a	4	11	0.0015	0	0	3	0.0011	-	-	1.37	21.15****	10.05**	0.24		−	
was a	4	11	0.0015	0	0	0	0	-	-	-	21.15****	-	6.97**	+		
it .	5	10	0.0014	0	0	0	0	-	-	-	19.23****	-	6.34*	+		
, and	6	29	0.0040	39	0.0034	0	0	1.20	-	-	0.55	-	18.37****			−
! !	7	28	0.0039	14	0.0012	0	0	3.23	-	-	13.87***	-	17.74****	+		
! <NC>	8	28	0.0039	103	0.0089	13	0.0049	0.44	0.55	0.80	17.15****	4.90*	0.42		+	
bread	9	0	0	3	0.0003	6	0.0022	-	8.67	-	-	9.88**	15.64****			+
get to	10	8	0.0011	0	0	0	0	-	-	-	15.38****	-	5.07*	+		
i really	10	8	0.0011	0	0	0	0	-	-	-	15.38****	-	5.07*	+		
i 'll	11	24	0.0034	20	0.0017	0	0	1.94	-	-	4.79*	-	15.21****	+		
, i	11	24	0.0034	24	0.0021	0	0	1.62	-	-	2.74	-	15.21****			−
was	12	91	0.0127	81	0.0070	20	0.0075	1.81	1.07	1.70	15.18****	0.07	5.10*	+		
i have	13	22	0.0031	69	0.0060	0	0	0.52	-	-	8.13**	-	13.94***			−
it 'll	14	7	0.0010	0	0	3	0.0011	-	-	0.87	13.46***	10.05**	0.04		−	
then i	14	7	0.0010	0	0	3	0.0011	-	-	0.87	13.46***	10.05**	0.04		−	
year .	14	7	0.0010	0	0	0	0	-	-	-	13.46***	-	4.44*	+		
supposed	15	0	0	1	0.0001	5	0.0019	-	21.67	-	-	11.75***	13.04***			+
supposed to	15	0	0	1	0.0001	5	0.0019	-	21.67	-	-	11.75***	13.04***			+
supposed to be	15	0	0	1	0.0001	5	0.0019	-	21.68	-	-	11.75***	13.04***			+
jim	15	0	0	2	0.0002	5	0.0019	-	10.83	-	-	9.19**	13.04***			+
fairly	15	0	0	4	0.0003	5	0.0019	-	5.42	-	-	6.04*	13.04***			+
, although	15	0	0	5	0.0004	5	0.0019	-	4.33	-	-	4.96*	13.04***			+
-	16	44	0.0061	76	0.0066	37	0.0139	0.94	2.11	0.44	0.13	12.53***	12.66***			+
, which was	17	9	0.0013	1	0.0001	1	0.0004	14.54	4.34	3.35	11.77***	0.99	1.81	+		
) and	18	6	0.0008	0	0	2	0.0007	-	-	1.12	11.54***	6.70**	0.02		−	
it 'll be	18	6	0.0008	0	0	2	0.0008	-	-	1.12	11.54***	6.70**	0.02		−	
'll have	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.80	+		
from the	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.80	+		
of it	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.80	+		
what i	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.80	+		
<NC>	19	291	0.0406	595	0.0514	119	0.0446	0.79	0.87	0.91	11.12***	2.10	0.71		+	

Table 5.10: Tokenised n-gram analysis, Extraversion

Note. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
<E>	1	67	0.0164	18	0.0016	1	0.0001	10.55	0.09	117.24	103.21****	11.40***	126.03****	+		
<E> <NC>	2	20	0.0049	10	0.0009	0	0	5.67	-	-	21.62****	-	40.46****	+		
! ! ! !	3	17	0.0042	1	0.0001	1	0.0001	48.21	1.62	29.76	38.59****	0.12	27.58****	+		
! ! ! ! !	4	14	0.0034	0	0	0	0	-	-	-	37.65****	-	28.33****	+		
! ! !	5	21	0.0051	4	0.0003	2	0.0003	14.89	0.81	18.38	36.89****	0.06	30.71****	+		
! !	6	27	0.0066	14	0.0012	4	0.0006	5.47	0.46	11.81	28.40****	2.10	34.40****	+		
have to	7	16	0.0039	24	0.0021	0	0	1.89	-	-	3.67	-	32.37****			−
<NC> well	8	11	0.0027	0	0	6	0.0008	-	-	3.21	29.57****	11.56***	5.60*	+		
think i	9	10	0.0024	0	0	5	0.0007	-	-	3.50	26.88****	9.63**	5.66*	+		
<E> <NC> i	10	13	0.0032	3	0.0003	0	0	12.29	-	-	21.32****	-	26.30****	+		
<NC> it	11	0	0	29	0.0025	24	0.0034	-	1.34	-	-	1.11	21.69****	−		
it .	12	0	0	0	0	11	0.0015	-	-	-	-	21.19****	9.94**			+
. <NC> well	13	9	0.0022	1	0.0001	5	0.0007	25.52	8.10	3.15	18.30****	5.19*	4.48*	+		
film	14	9	0.0022	2	0.0002	0	0	12.75	-	-	14.97***	-	18.21****	+		
all the	14	9	0.0022	9	0.0008	0	0	2.83	-	-	4.68*	-	18.21****	+		
, as	15	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.14**			+
dave	15	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.14**			+
, and	16	0	0	39	0.0034	19	0.0027	-	0.79	-	-	0.73	17.17****	−		
the film	17	6	0.0015	0	0	0	0	-	-	-	16.13****	-	12.14***	+		
well i	17	6	0.0015	0	0	0	0	-	-	-	16.13****	-	12.14***	+		
! <NC>	18	13	0.0032	103	0.0089	33	0.0046	0.36	0.52	0.69	15.80****	11.89***	1.36		+	
rowing	19	0	0	0	0	8	0.0011	-	-	-	-	15.41****	7.23**			+
to the	20	0	0	24	0.0021	16	0.0022	-	1.08	-	-	0.06	14.46***	−		
its	21	7	0.0017	6	0.0005	0	0	3.31	-	-	4.50*	-	14.16***	+		
i will	21	7	0.0017	13	0.0011	0	0	1.53	-	-	0.78	-	14.16***			−
we	22	14	0.0034	98	0.0085	29	0.0041	0.40	0.48	0.84	12.46***	13.63***	0.27		+	
still	23	0	0	30	0.0026	15	0.0021	-	0.81	-	-	0.45	13.56***	−		
about it	24	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
there 's	24	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
she 's	24	2	0.0005	0	0	7	0.0010	-	-	0.50	5.38*	13.48***	0.84		−	
<E> well	25	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.12**	+		
experiment	25	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.12**	+		
was a	25	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.12**	+		
!	26	46	0.0113	129	0.0112	43	0.0060	1.01	0.54	1.87	0.00	13.43***	8.66**			−
,	27	91	0.0223	294	0.0254	242	0.0339	0.88	1.33	0.66	1.22	10.84***	12.28***			+
i would	28	6	0.0015	3	0.0003	0	0	5.67	-	-	6.49*	-	12.14***	+		
i ca	28	6	0.0015	10	0.0009	0	0	1.70	-	-	1.00	-	12.14***			−
, so	29	0	0	24	0.0021	13	0.0018	-	0.88	-	-	0.15	11.75***	−		
of time	30	3	0.0007	0	0	6	0.0008	-	-	0.88	8.06**	11.56***	0.04		−	
<NC> also	30	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.42*			+
going on	30	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.42*			+
night .	30	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.42*			+
thesis	31	6	0.0015	1	0.0001	3	0.0004	17.01	4.86	3.50	10.99***	2.24	3.39	+		
stuff	31	3	0.0007	3	0.0003	12	0.0017	2.83	6.48	0.44	1.56	10.99***	1.91			+

Table 5.11: Tokenised n-gram analysis, Neuroticism

Note. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
!	1	10	0.0022	129	0.0112	96	0.0186	0.19	1.67	0.12	39.76****	13.91***	71.35****	−		
! <NC>	2	9	0.0020	103	0.0089	63	0.0122	0.22	1.37	0.16	29.06****	3.78	39.70****	−		
<E>	3	34	0.0074	18	0.0016	6	0.0012	4.74	0.75	6.34	30.35****	0.40	24.92****	+		
! !	4	0	0	14	0.0012	20	0.0039	-	3.20	-	-	11.30***	25.54****			+
! <NC> i	5	0	0	27	0.0023	19	0.0037	-	1.58	-	-	2.24	24.26****	−		
, and	6	0	0	39	0.0034	18	0.0035	-	1.03	-	-	0.01	22.98****	−		
to the	7	0	0	24	0.0021	17	0.0033	-	1.59	-	-	2.06	21.71****	−		
of the	7	0	0	25	0.0022	17	0.0033	-	1.52	-	-	1.74	21.71****	−		
on saturday	8	0	0	8	0.0007	16	0.0031	-	4.48	-	-	12.98***	20.43****			+
on the	9	12	0.0026	24	0.0021	0	0	1.25	-	-	0.40	-	18.03****			−
and i	10	0	0	41	0.0035	14	0.0027	-	0.77	-	-	0.78	17.87****	−		
<NC> it	10	0	0	29	0.0025	14	0.0027	-	1.08	-	-	0.06	17.87****	−		
have n't	10	0	0	29	0.0025	14	0.0027	-	1.08	-	-	0.06	17.87****	−		
i have	11	7	0.0015	69	0.0060	29	0.0056	0.25	0.94	0.27	17.14****	0.07	12.07***	−		
<E> <NC>	12	11	0.0024	10	0.0009	0	0	2.76	-	-	5.26*	-	16.52****	+		
was a	13	6	0.0013	0	0	6	0.0012	-	-	1.12	15.06***	14.11***	0.04		−	
<NC> have	13	6	0.0013	0	0	0	0	-	-	-	15.06***	-	9.01**	+		
night .	14	0	0	0	0	6	0.0012	-	-	-	-	14.11***	7.66**			+
i really	14	3	0.0007	0	0	6	0.0012	-	-	0.56	7.53**	14.11***	0.71		−	
, as	15	5	0.0011	0	0	0	0	-	-	-	12.55***	-	7.51**	+		
, we	15	5	0.0011	0	0	0	0	-	-	-	12.55***	-	7.51**	+		
it all	15	5	0.0011	0	0	0	0	-	-	-	12.55***	-	7.51**	+		
of it	15	5	0.0011	0	0	0	0	-	-	-	12.55***	-	7.51**	+		
here	16	8	0.0017	14	0.0012	0	0	1.43	-	-	0.64	-	12.02***			−
out of	17	3	0.0007	0	0	5	0.0010	-	-	0.67	7.53**	11.76***	0.31		−	
for my	17	0	0	0	0	5	0.0010	-	-	-	-	11.76***	6.38*			+
will probably	17	0	0	0	0	5	0.0010	-	-	-	-	11.76***	6.38*			+
working	18	3	0.0006	9	0.0008	16	0.0031	0.84	3.98	0.21	0.07	11.60***	8.36**			+
i did	19	0	0	4	0.0003	9	0.0017	-	5.04	-	-	8.07**	11.49***			+
went to	19	0	0	20	0.0017	9	0.0017	-	1.01	-	-	0.00	11.49***	−		
i had	20	4	0.0009	8	0.0007	15	0.0029	1.25	4.20	0.30	0.13	11.46***	5.60*			+
saturday	21	5	0.0011	14	0.0012	20	0.0039	0.90	3.20	0.28	0.05	11.30***	8.03**			+
have	22	28	0.0061	134	0.0116	58	0.0112	0.52	0.97	0.54	11.05***	0.04	7.58**	−		
! ! <NC>	23	1	0.0002	8	0.0007	13	0.0025	0.31	3.64	0.09	1.60	8.57**	10.90***			+

Table 5.12: Tokenised n-gram analysis, Psychoticism

Note. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

[*it* .], and references to [*rowing*], and conversely show an under-use of [*have to*], and [*all the*]; High and Low groups are distinguished from the Mid group by both over-using [*think i*], [<NC> *well*], [. <NC> *well*], and under-using [*!*<NC>]. Other, less significant, patterns in the data for the High Neurotics are the under-use of [*its*], [*still*], and [, *so*], along with the overuse of [<E> *well*], [*experiment*], [*was a*], and [*thesis*]; for the Low group, there is tendency towards an under-use of [*its*], [*i will*], [*!*], [.], [*i would*], [*i ca*] (as in *i ca n't*), and an overuse of [*about it*], [*there 's*], [*of time*], [<NC> *also*], [*going on*], [*night* .], and [*stuff*]; The High and Low group both use fewer instances of [*we*] and [*of time*] than the Mid group, but use more [*she 's*].

The Psychoticism data (Table 5.12) also show distinctive use of punctuation in combination with structural information, and the High group show an overuse of ellipsis (<E>) under-use of [*!*], [*!!*], [*!* <NC>], [<NC> *it*], [, *and*], [*and i*], [*i have*], [*have n't*], [*to the*], [*of the*], and [*on saturday*]; By contrast, the Low Psychotic group show an underuse of [*on the*], and [<E> <NC>].

Features which show a lower level of significance for the High group, are an overuse of [<NC> *have*], [, *as*], [, *we*], [*it all*], [*of it*], along with an under-use of [*went to*], and [*have*]; Low Psychotics overuse [*night*], [*for my*], [*will probably*], [*working*], [*i did*], [*i had*], [*saturday*], and [*!!* <NC>]; Both High and Low groups overused [*was a*], [*i really*], and [*out of*] compared with the Mid group.

5.3.3 Summary

Considering the results for Extraversion, Neuroticism, and Psychoticism overall, we can see that even at the most conservative 15.13 critical level, there is still a reasonable number of significant features. For Extraversion and Psychoticism, the top 12 ranked features achieve this significance, whereas for Neuroticism this is 19. Although in the case of the latter, this appears to be inflated due, in part, to the repetition of multiple exclamation mark results. This brings us to another issue, that of features being shared across the three personality dimensions, and this is most noticeable with regard to the punctuation/contextual markers. Here we find that the unigram feature [<E>], ellipsis alone, is strongly characteristic of High Extraverts and High Neurotics, and to a slightly lesser extent Low Psychotics. However, by examining the contextual in-

formation for this feature available via the n-grams, it is possible to see how these features are used slightly differently across the personality dimensions: ellipsis alone is strongly characteristic of High Extraverts, since they do not regularly combine ellipsis with other features or constructions; For High Neurotics this is similarly strongly shown to be followed by a newclause, or by a newclause starting with *I* (they also show a trend towards following an ellipsis with *well*); Although the High Psychotics overuse this feature, it is slightly less strongly differentiated across the groups, and there is also evidence that it is under-used by Low Psychotics when preceding a newclause. Similarly the different behaviour of (multiple) exclamation marks across the different personality dimensions can be discerned through the n-gram data. For example, Low Extraverts avoid [*! !*], and both High and Low Extraverts avoid a single exclamation mark before a newclause (<NC>), indicating a sentence-final position; whereas High Psychotics avoid (multiple) exclamation marks in all positions; High Neurotics on the other hand overuse multiple exclamation marks, but (along with, to a lesser extent, Low Neurotics) avoid using single occurrences before a newclause.

These comparisons also reveal the value of using data which has been tokenised in a more sophisticated manner which identifies clause boundaries, and also allows us to more easily distinguish multiple full stops or ellipsis, as opposed to, for example, sentence final punctuation.

Looking at the personality features more generally, here we can see that Low Extraverts appear to talk more about (playing) activities, and seem to make fewer references to themselves. High Extraverts tend to talk more about events in the past (was) and in the future (get to). In terms of less strongly significant features, Low Extraverts appear to talk about obligation regarding the future (supposed) and hedge (fairly), whereas High Extraverts use the more definite (contracted) modal form (will), and show evidence of evaluating past events (which was).

Punctuation (ellipsis, exclamation marks) and discourse markers relative to clause structure (well, i) are important to High Neurotics, and they also appear to talk about films, whereas Low Neurotics are characterised by a lack of external obligation (have to), quantification (all the) and conjunction (, as). Less significant patterns for this dimension appear to be that High Neurotics talk about what is perhaps concerning

them at the that moment (experiment, thesis), and are more likely to mention what they cannot do (i ca [n't]); Low Neurotics talk less about what they will do (i will) and also talk more vaguely about things (stuff).

High Psychotics are less likely to consistently use (patterns of) punctuation such as exclamation or ellipsis, use conjunction (and), or talk about things relative to themselves (i have, haven't, i). Conversely less significant patterns show that Low Psychotics are more likely to refer to themselves (i, my) rather than with others (we), and also refer to weekends (saturday).

5.4 Lemmatised corpus analysis

In the previous section, we noted that the utilisation of a more sophisticated tokenisation technique, allowed us greater confidence in identifying features and patterns and relating them to the context or structure of their use. However, it is also apparent that some of the features found in this analysis might show greater significance or generalisability if the analysis was performed on a lemmatised—or stemmed—form of the corpus. In such a processed corpus words such as *play*, *plays*, *played*, or *playing*, are all realised in the base form of the verb, that is as *play*. More importantly, in our data there are instances of proper nouns being used, for example, names of places (*Edinburgh*), days of the week (*Saturday*), or names of people (*Dave*), with these providing too much specificity to allow broader patterns of language usage to emerge, or for the results to be easily generalised.

We therefore perform an additional 3-way stratified corpus comparison in the same way as before, but process the personality corpora, so as to lemmatise it and replace proper nouns with their equivalent part-of-speech (POS) tags.

5.4.1 Method

Using the same subcorpora as the previous stratified corpus comparison, we additionally pre-processed these using the CLAWS tagger (Rayson, 2003) to give vertical-output lemmatised words and POS tags. Additional scripts were then used to convert this into the form of lemmas, and in the case of the features being a proper noun, this

was replaced by the POS tag. Additionally all punctuation (including the newclause marker) was replaced by 'punctuation tag'.

Each of these resulting lemmatised/proper noun tagged subcorpora were then analysed in the same way as before, to give frequency and relative frequency information for High, Low and Mid groups, with additional relative-frequency ratio and log-likelihood information for High-Low, High-Mid, and Low-Mid comparisons.

5.4.2 Results

Again the results are ordered by significance of the feature's log-likelihood (G^2) value, with the conservative 15.13 critical value for most significant features (indicated by a horizontal rule), and we also include features to the 10.83 value. These are displayed for Extraversion, Neuroticism, and Psychoticism, in the following tables (Tables 5.13, 5.14, and 5.15).

As in the overview of the previous results, here we will discuss the features which are characteristic of the High or Low groups, compared with the usage of the feature more generally. Again, we will take the particularly conservative critical value of 15.13 to indicate a high level of significance, with the other features below this level displaying less significant examples of the particular behaviour. In the results reported here the following CLAWS POS tags are used: NP1 (singular proper noun), and NPD1 (singular weekday noun).

For the High Extraverts (Table 5.13) we find that the lemmatisation results in changes in the most significant features, with them overusing [*be so*], [*year <P>*], [*<P> take*], [*with i*], [*NP1 for*], in addition to [*i really*] (which was found in the unlemmatised data), and under-using [*week <P>*]; the Low Extraverts still overuse references to play ([*play*], [*i play*]), [*that be*], and [*bread*], with the additional overuse of 'supposed' showing greater significance ([*be supposed*]) and the under-use of [*i will*]; Additionally both High and Low Extraverts show an overuse of [*get a*], and [*christmas <P>*] compared with Mid.

Additionally, new features arising out of this analysis with a lower level of significance are for Extraverts, an overuse of [*day <P>*], [*will have*], [*cool <P>*], [*today <P>*], and an under-use of [*that i*]; for Low Extraverts, we note an under-use of

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
play	1	3	0.0004	2	0.0002	14	0.0052	2.42	30.31	0.08	0.97	35.63****	22.53****			+
get a	2	15	0.0021	0	0	5	0.0019	-	-	1.12	28.86****	16.74****	0.05		−	
be so	3	14	0.0020	0	0	0	0	-	-	-	26.93****	-	8.88**	+		
i play	4	0	0	0	0	7	0.0026	-	-	-	-	23.43****	18.24****			+
christmas <P>	5	10	0.0014	0	0	3	0.0011	-	-	1.24	19.24****	10.04**	0.11		−	
year <P>	5	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.34*	+		
week <P>	6	0	0	42	0.0036	7	0.0026	-	0.72	-	-	0.69	18.24****	−		
i will	7	28	0.0039	35	0.0030	0	0	1.29	-	-	1.02	-	17.76****			−
<P> take	8	9	0.0013	0	0	0	0	-	-	-	17.31****	-	5.71*	+		
with i	8	9	0.0013	0	0	0	0	-	-	-	17.31****	-	5.71*	+		
be supposed	9	0	0	0	0	5	0.0019	-	-	-	-	16.74****	13.03***			+
that be	9	0	0	0	0	5	0.0019	-	-	-	-	16.74****	13.03***			+
bread	10	0	0	3	0.0003	6	0.0022	-	8.66	-	-	9.87**	15.63****			+
NP1 for	11	8	0.0011	0	0	0	0	-	-	-	15.39****	-	5.07*	+		
i really	11	8	0.0011	0	0	0	0	-	-	-	15.39****	-	5.07*	+		
then i	12	7	0.0010	0	0	3	0.0011	-	-	0.87	13.47***	10.04**	0.04		−	
day <P>	12	7	0.0010	0	0	0	0	-	-	-	13.47***	-	4.44*	+		
will have	12	7	0.0010	0	0	0	0	-	-	-	13.47***	-	4.44*	+		
NP1 and	13	21	0.0029	22	0.0019	0	0	1.54	-	-	2.00	-	13.32***			−
be supposed to	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.75***	13.03***			+
be supposed to be	14	0	0	1	0.0001	5	0.0019	-	21.67	-	-	11.75***	13.03***			+
supposed to be	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.75***	13.03***			+
supposed	14	0	0	1	0.0001	5	0.0019	-	21.65	-	-	11.74***	13.03***			+
supposed to	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.74***	13.03***			+
fairly	14	0	0	4	0.0003	5	0.0019	-	5.41	-	-	6.03*	13.03***			+
<P> although	14	0	0	7	0.0006	5	0.0019	-	3.09	-	-	3.34	13.03***	−		
that i	14	0	0	27	0.0023	5	0.0019	-	0.80	-	-	0.22	13.03***	−		
and i	15	20	0.0028	44	0.0038	0	0	0.73	-	-	1.35	-	12.69***			−
and NP1	16	19	0.0027	13	0.0011	0	0	2.36	-	-	5.84*	-	12.05***	+		
take	17	25	0.0035	13	0.0011	7	0.0026	3.11	2.33	1.33	11.79***	2.93	0.48	+		
cool <P>	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
from the	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
of it	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
today <P>	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
what i	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		

Table 5.13: Lemmatised n-gram analysis, Extraversion.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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> it	1	0	0	45	0.0039	38	0.0053	-	1.37	-	-	2.00	34.37****	−		
<P> <P> <P> <P> <P>	2	16	0.0039	2	0.0002	0	0	22.67	-	-	31.66****	-	32.37****	+		
NP1 <P>	3	16	0.0039	57	0.0049	0	0	0.80	-	-	0.68	-	32.36****			−
<P> <P> <P> <P>	4	21	0.0051	10	0.0009	2	0.0003	5.95	0.32	18.37	23.50****	2.65	30.70****	+		
<P> as	5	0	0	0	0	14	0.0020	-	-	-	-	26.97****	12.66***			+
<P> <P> <P>	6	43	0.0105	56	0.0048	23	0.0032	2.18	0.67	3.27	13.89***	2.85	22.43****	+		
film	7	11	0.0027	2	0.0002	0	0	15.58	-	-	19.60****	-	22.25****	+		
i go	8	8	0.0020	0	0	8	0.0011	-	-	1.75	21.50****	15.41****	1.23		−	
that be	9	7	0.0017	0	0	11	0.0015	-	-	1.11	18.81****	21.19****	0.05		−	
<P> he	9	0	0	0	0	11	0.0015	-	-	-	-	21.19****	9.95**			+
<P> well	10	21	0.0051	12	0.0010	10	0.0014	4.96	1.35	3.67	20.43****	0.48	12.53***	+		
will be	11	0	0	39	0.0034	21	0.0029	-	0.87	-	-	0.26	18.99****	−		
have be	11	0	0	37	0.0032	21	0.0029	-	0.92	-	-	0.10	18.99****	−		
all the	12	9	0.0022	9	0.0008	0	0	2.83	-	-	4.67*	-	18.20****	+		
<P> so	13	0	0	34	0.0029	20	0.0028	-	0.95	-	-	0.03	18.09****	−		
be in	14	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.14**			+
<P> <P> well	15	12	0.0029	4	0.0003	6	0.0008	8.50	2.43	3.50	16.67****	1.94	6.78**	+		
film be	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
the film	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
well i	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
year <P>	17	0	0	0	0	8	0.0011	-	-	-	-	15.41****	7.24**			+
to do	18	0	0	11	0.0009	17	0.0024	-	2.50	-	-	5.80*	15.37****			+
to the	19	0	0	24	0.0021	16	0.0022	-	1.08	-	-	0.06	14.47***	−		
though <P>	20	7	0.0017	12	0.0010	0	0	1.65	-	-	1.06	-	14.16***			−
to NP1	20	7	0.0017	25	0.0022	0	0	0.79	-	-	0.31	-	14.16***			−
we	21	18	0.0044	119	0.0103	47	0.0066	0.43	0.64	0.67	13.74***	7.13**	2.21		+	
still	22	0	0	30	0.0026	15	0.0021	-	0.81	-	-	0.45	13.57***	−		
about it	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
it do	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
rowing	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
and she	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
the film be	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
the time	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
experiment	24	5	0.0012	0	0	2	0.0003	-	-	4.37	13.44***	3.85*	3.54		−	
<P> which	25	0	0	15	0.0013	14	0.0020	-	1.51	-	-	1.22	12.66***	−		
have not	25	0	0	32	0.0028	14	0.0020	-	0.71	-	-	1.20	12.66***	−		
NP1 and	25	0	0	22	0.0019	14	0.0020	-	1.03	-	-	0.01	12.66***	−		
stuff	26	3	0.0007	3	0.0003	13	0.0018	2.83	7.02	0.40	1.56	12.48***	2.38			+
<P> <P> we	27	4	0.0010	34	0.0029	5	0.0007	0.33	0.24	1.40	5.73*	12.45***	0.25		+	
i ca	28	6	0.0015	10	0.0009	0	0	1.70	-	-	1.00	-	12.13***			−
of time	29	3	0.0007	0	0	6	0.0008	-	-	0.87	8.06**	11.56***	0.04		−	
get a	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
go on	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
party <P>	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
stuff <P>	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
have to	30	21	0.0051	30	0.0026	11	0.0015	1.98	0.59	3.34	5.47*	2.35	11.24***	+		
thesis	31	6	0.0015	1	0.0001	3	0.0004	16.99	4.86	3.50	10.99***	2.24	3.39	+		
<P> the	32	0	0	16	0.0014	12	0.0017	-	1.21	-	-	0.26	10.85***	−		
he be	32	0	0	22	0.0019	12	0.0017	-	0.88	-	-	0.12	10.85***	−		
well <P>	32	0	0	18	0.0016	12	0.0017	-	1.08	-	-	0.04	10.85***	−		

Table 5.14: Lemmatised n-gram analysis, Neuroticism.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
have be	1	0	0	37	0.0032	21	0.0041	-	1.27	-	-	0.76	26.83****	–		
<P> it	1	0	0	45	0.0039	21	0.0041	-	1.05	-	-	0.03	26.83****	–		
i go	2	0	0	0	0	11	0.0021	-	-	-	-	25.87****	14.06***			+
<P> NPD1 <P>	3	10	0.0022	0	0	2	0.0004	-	-	5.59	25.09****	4.70*	6.75**	+		
<P> <P> NPD1 <P>	3	10	0.0022	0	0	2	0.0004	-	-	5.59	25.09****	4.70*	6.75**	+		
be work	4	0	0	0	0	10	0.0019	-	-	-	-	23.52****	12.78***			+
<P> he	5	9	0.0019	0	0	0	0	-	-	-	22.58****	-	13.51***	+		
to the	6	0	0	24	0.0021	17	0.0033	-	1.59	-	-	2.05	21.72****	–		
of the	6	0	0	25	0.0022	17	0.0033	-	1.52	-	-	1.74	21.72****	–		
on the	7	12	0.0026	24	0.0021	0	0	1.25	-	-	0.40	-	18.01****			–
be off	8	7	0.0015	0	0	4	0.0008	-	-	1.96	17.56****	9.41**	1.20		–	
that be	9	0	0	0	0	7	0.0014	-	-	-	-	16.46****	8.94**			+
you be	10	6	0.0013	0	0	0	0	-	-	-	15.05***	-	9.01**	+		
he be	11	10	0.0022	22	0.0019	0	0	1.14	-	-	0.11	-	15.01***			–
all i	12	0	0	0	0	6	0.0012	-	-	-	-	14.11***	7.67**			+
i really	12	0	0	0	0	6	0.0012	-	-	-	-	14.11***	7.67**			+
too <P>	12	0	0	0	0	6	0.0012	-	-	-	-	14.11***	7.67**			+
night <P>	13	0	0	19	0.0016	11	0.0021	-	1.30	-	-	0.46	14.06***	–		
have	14	67	0.0145	256	0.0221	129	0.0250	0.66	1.13	0.58	10.24**	1.25	13.64***	–		
<P> anyway	15	9	0.0019	14	0.0012	0	0	1.61	-	-	1.19	-	13.51***			–
anyway <P>	15	9	0.0019	16	0.0014	0	0	1.41	-	-	0.65	-	13.51***			–
<P> <P> <P> <P>	16	1	0.0002	10	0.0009	15	0.0029	0.25	3.36	0.07	2.52	9.01**	13.19***			+
work	17	18	0.0039	44	0.0038	43	0.0083	1.02	2.19	0.47	0.01	12.99***	7.95**			+
on NPD1	18	10	0.0022	50	0.0043	36	0.0070	0.50	1.61	0.31	4.60*	4.63*	12.84***			+
you <P> <P>	19	0	0	8	0.0007	10	0.0019	-	2.80	-	-	4.69*	12.78***			+
how be	19	0	0	12	0.0010	10	0.0019	-	1.87	-	-	2.06	12.78***	–		
<P> so	19	0	0	34	0.0029	10	0.0019	-	0.66	-	-	1.44	12.78***	–		
get a	20	5	0.0011	0	0	5	0.0010	-	-	1.12	12.54***	11.76***	0.03		–	
<P> perhaps	20	5	0.0011	0	0	0	0	-	-	-	12.54***	-	7.51**	+		
day <P>	20	5	0.0011	0	0	0	0	-	-	-	12.54***	-	7.51**	+		
it all	20	5	0.0011	0	0	0	0	-	-	-	12.54***	-	7.51**	+		
of it	20	5	0.0011	0	0	0	0	-	-	-	12.54***	-	7.51**	+		
this be	20	5	0.0011	0	0	0	0	-	-	-	12.54***	-	7.51**	+		
end up	20	5	0.0011	0	0	2	0.0004	-	-	2.80	12.54***	4.70*	1.69		–	
i have	21	24	0.0052	107	0.0092	60	0.0116	0.56	1.26	0.45	7.31**	1.96	12.18***	–		
<P> anyway <P>	22	8	0.0017	12	0.0010	0	0	1.67	-	-	1.21	-	12.01***			–
here	22	8	0.0017	14	0.0012	0	0	1.43	-	-	0.63	-	12.01***			–
out of	23	3	0.0006	0	0	5	0.0010	-	-	0.67	7.53**	11.76***	0.31		–	
say <P>	23	0	0	0	0	5	0.0010	-	-	-	-	11.76***	6.39*			+
they have	23	0	0	0	0	5	0.0010	-	-	-	-	11.76***	6.39*			+
<P> how	24	0	0	15	0.0013	9	0.0017	-	1.34	-	-	0.48	11.50***	–		
we be	24	0	0	16	0.0014	9	0.0017	-	1.26	-	-	0.30	11.50***	–		

Table 5.15: Lemmatised n-gram analysis, Psychoticism.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

singular proper noun [*NP1 and*], and [*and NP1*].

For High Neurotics (Table 5.14) their previous overuse of multiple exclamation marks is here represented more generally as an overuse of multiple Punctuation ([<P> <P> <P>]), with a similar process relating to the reduction of full stop and newclause ([<P> *well*], [<P> <P> *well*]), whereas overuse of [*film*], [*the film*], and [*well i*] is retained. The reduction of punctuation results in the most significant High Neurotic feature being the under-use of [<P> *it*], with other under-used features being [*will be*], [*have be*], [<P> *so*], along with [*to do*]; Low Neurotics overuse [<P> *he*], [*be in*], and [*year* <P>], and under-use [*NP1* <P>]; High and Low groups also show an overuse of [*i go*] and [*that be*].

Additional less strongly significant features identified in this analysis are, for High Neurotics the under-use of [<P> *which*], and [*well* <P>]; and for Low Neurotics the overuse of [*party* <P>], [*stuff* <P>], and [<P> *the*], and the under-use of [*though* <P>], [*to NP1*], [<P> <P> *we*] and [*i ca*] (as in *i ca n't*).

For the High Psychotic group (Table 5.15) the most significant features which result from this analysis are the overuse of [<P> <P> *NP1* <P>], [<P> *NP1* <P>], and [<P> *he*], and the under-use of [*have be*], [<P> *it*], [*to the*], and [*of the*]; Low Psychotics are characterised by an overuse of [*i go*], [*be work*], [*that be*], and an under-use of [*on the*]; High and Low Psychotics both overuse [*be off*].

Other features resulting from this analysis, but showing lower significance, are High Psychotics overuse of [*you be*], [<P> *perhaps*], [*day* <P>], [*this be*], [*here*], and an under-use of [*night* <P>], [*have*], [<P> *so*], [<P> *how*], and [*we be*]; Low Psychotics overuse [*all i*], [*i really*],[6] [*too* <P>], [*work*], [*on NPD1*], [*you* <P> <P>], [*how be*], [*say* <P>], and [*they have*], and underuse [*he be*], [<P> *anyway*], [*anyway* <P>], and [<P> *anyway* <P>]; Terms shared between High and Low groups [*get a*], [*end up*], and [*out of*] which are over-used, and [*i have*] which is under-used.

[6]Note that, unlike the previous analysis, the bigram [*i really*] is not found in the lemmatised High P data because its items do not collocate significantly together at the $p \leq 0.01$ level.

5.5 Discussion

5.5.1 Data-driven analysis

Firstly we have investigated the behaviour of simple bigram collocation measures across the high and low subcorpora of our personality dimensions. This allowed us to derive groups of linguistic features on the basis of our data (i.e., data-driven), rather than by imposing top-down categories on our data, as for example, by multi-dimensional analysis. However, as a result of this analysis, several issues were raised: First, we have assumed that the personality dimensions are orthogonal and have treated the linguistic behaviour for each dimension independently. However, given the relatively modest size of our corpus, it may be the case that a small number of individuals who are extreme on more than one personality dimension skew the results if they demonstrate uncharacteristic linguistic behaviour; Secondly this analysis compares the linguistic behaviour of extreme personality groups - it does not consider the distribution of language features across authors with an 'average' or 'neutral' personality type; Thirdly, we discuss the value of measuring collocations of n-grams with different lengths.

As an advancement upon the previous analysis, we apply corpus-comparison techniques to the study of our personality corpus, by adopting a 'Mid' reference corpus. We additionally compared the High and Low groups together to allow the novel three-way High-Mid-Low stratified corpus comparison. Given that personality score—as we have measured it here, using the EPQ-R—is a continuous, rather than discrete variable (Butler, 1985)—such as gender, as analysed by (Rayson and Hodges, 1997)—then this analysis provides a way in which language behaviour can be investigated more closely along the range of the personality dimension (i.e., high–mid–low). The disadvantage of this method is that we are forced to be even more selective in allocating texts to corpus groups, which reduces the size of these subcorpora even further. However, even with these reduced subcorpora sizes, this analysis found features which significantly varied across each dimension, even using a minimum frequency of 5, and the conservative 15.13 critical level recommended by Rayson (2003).

Finally, we extend this analysis further by using more sophisticated annotation

Extraversion											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	$\ominus^{k,l}$	$\ominus^{k,l}$	$\oplus^{k,l}$	$\oplus^{k,l}$							−Quantification
Real.	○	$\oplus^{k}$	○								
Fluency	○	$\oplus^{k,l}$	$\ominus^{k}$								
Gramm.	$\odot^{k,l}$	$\odot^{k,l}$	○	○	○	○	○	$\oplus^{k}$			
Conv.	$\oplus^{k,l}$	$\oplus^{k}$	$\oplus^{k}$	$\oplus^{k,l}$	○	$\oplus^{k,l}$	$\odot^{k,l}$	$\odot^{k,l}$			
Lexis	$\oplus^{k}$	○	$\oplus^{k}$	○	$\oplus^{k}$	○	○	○	○	○	
Percept.	○	○									

Neuroticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	○	○	○	○	$\oplus^{k,l}$						−Temporal; −Quantification; −Modality
Conv.	○	○									+Surface features
Lexis	$\oplus^{k,l}$	○	○	○	○						
Percept.	○	○									

Psychoticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	$\oplus^{k,l}$	○	○	○	○	○	○				+Temporal
Conv.	$\odot^{l}$										
Lexis	$\oplus^{k,l}$	○	○	○	○	$\ominus^{k}$	○	○	○	○	+Modality
Percept.	○	○									

Table 5.16: Review of hypotheses

Note. ○ indicates an hypothesis; $\oplus$ confirmation of hypothesis; $\ominus$ inverse of hypothesis; $\odot$ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{k} tokenised corpus comparison; ^{l} lemmatised corpus comparison. Please refer to Section 2.6 for a full description of the hypotheses.

techniques, namely lemmatisation and part-of-speech, which allowed us to derive results which were more easily generalisable.

5.5.2 Summary of findings and review of hypotheses

Using data-driven techniques we have been able to investigate linguistic features which characterise the expression of personality in e-mail communication, without being restricted by predefined analysis methods. We relate these findings back to the hypotheses in Table 5.16, and summarise them as follows:

Extraversion is characterised in High Extraverts by the use of more informal surface realisations (confirming our Grammatical hypothesis), and non-standard multiple punctuation, for example, exclamations (confirming our Realisation and Conversa-

tional hypotheses), and also ellipses and dashes (confirming our Fluency hypothesis). They also use fewer negatively valanced expressions (confirming our Theory hypothesis), and more strongly predictive modality confirming our Conversational hypothesis).

Neuroticism is characterised by increased reference to self in High Neurotic individuals (confirming our Theory hypothesis), however, we also note their use of long chains of punctuation features (exclamation and ellipsis), use of negation and negative valence expressions.

Psychoticism is characterised by High Psychotics avoiding references to themselves, however they do make increased references to others (this partially confirms our Theory hypothesis which predicted that they would make fewer references to themselves or others). We also note that High Psychotics avoid non-standard punctuation features and references to work.

5.6 Conclusion

In this chapter we have investigated the use of empirical corpus linguistic techniques which have incorporated multi-word units in the investigation of personality language characteristics. In the adoption of the stratified corpus comparison, we have been able to chart the use of linguistic features along the cross section of personality, and although this has resulted in smaller subcorpora groups, it has still shown significant results. Furthermore, by using corpora which have been lemmatised and proper nouns replaced by POS tags, we are now able to examine broader language features, with more generalisable results, which reduces the risk of over-fitting to our particular language data, or of data sparsity.

To extend this generality of findings, and to find larger linguistic patterns, reducing the data even further may be advantageous. In the next chapter we explore empirical approaches to higher-level analysis of syntactic and semantic information.

Chapter 6

Data-driven Syntactic and Semantic corpus comparison

At the end of the last chapter, we introduced the basic annotation of our linguistic data using simple syntactic information: namely using the lemmatised form of words, whereby they are reduced to their 'stem' or most basic form, and using part of speech tags to replace proper nouns. In this chapter we retain the stratified corpus comparison technique which we developed previously and extend the annotation of data further, to allow the analysis of higher-level syntactic and semantic features of personality language. Such techniques enable us to determine more generalisable features, and also reduce the effects of data sparsity.[1]

[1]This work is partially reported in Oberlander and Gill (to appear).

6.1 Introduction

In this chapter we explore the use of higher-level syntactic and semantic categories to annotate our personality corpus. This allows us to abstract away from content-specific features, and enables us to derive results which will generalise better to different forms of language. Indeed, Milic (1966) notes that 'Lexical choices are conscious and context-bound', and therefore reasons that the 'grammatical or syntactic component of writing [is] the best source of information about a writer's style' (p. 83). In the last chapter we also noted the relatively modest size of our subcorpus personality groups. An additional benefit of our adoption of higher-level annotations is that it reduces the possible effects of data-sparsity.

We structure the chapter as follows; Firstly we describe the syntactic analyses. We outline the different levels of annotation used, and in order to provide results which are comparable with previous findings, we present the unigrams both separately, and also as part of the combined 1–5-gram analysis. The second part of the chapter investigates our semantic analysis. We describe the different annotation used and present the unigrams analyses. We also additionally present our novel n-gram analysis of the semantic categories. We conclude the chapter with a discussion of the methods used, and compare our findings to those proposed in the hypotheses.

6.2 Syntactic Analysis of the Corpus

6.2.1 Method

The Penn part-of-speech tagged (using the MXPOST tagger; Ratnaparkhi, 1996) version of the personality corpus, which we have used previously for the MRC Psycholinguistic Database analysis (described in Section 4.4.1), was processed in order to remove the original words, but to leave their associated POS tags. A further level of processing was additionally carried out to reduce the POS tags from the detailed Penn tagset to more general syntactic categories. The 45 Penn tags (a key to the major tags, excluding punctuation, is included in Table B.1; see Marcus et al., 1994, for more details) were converted to the 10 broader categories of Noun (NN), Adjective (ADJ), Verb

(VBN), Adverb (ADV), Preposition (PRP), Conjunction (CONJ), Pronoun (PRN), Interjection (INT), Past Participle (VPP), and Other (O), as used in the electronic version of the Shorter Oxford English Dictionary and which is incorporated into the MRC Psycholinguistic Database. These categories were converted using the same algorithm as was used in the lookup program of our MRC Database analysis (Section 4.4.1). In addition to these categories, we also make use of <P> indicating punctuation, and 'NA', which indicates that a feature does not belong to any of the above categories and generally represents the <END>, end of text marker.

The resulting two versions of the corpora with Penn POS and general syntactic categories were then divided into the High-Mid-Low stratified corpus groups, as in the previous chapter. The same analysis was then carried out for both tagged corpus versions to give frequency and relative frequency information for High, Low and Mid groups, with additional relative-frequency ratio and log-likelihood information for High-Low, High-Mid, and Low-Mid comparisons. In order to observe relationships between personality groups and their use of broad grammatical categories we firstly display the results of the unigram analysis separately for each analysis. As in previous n-gram analyses, we also display the results of the overall n-gram analyses (1–5 item sequences) together.

6.2.2 Unigram Syntactic Analysis Results

The results of the unigram analysis of the most detailed Penn POS tag categories can be found for Extraversion, Neuroticism, and Psychoticism in Tables 6.1, 6.2, and 6.3. Results of the unigram analysis which utilised the reduced set of syntactic tags can be found in tables 6.4, 6.5, and 6.6. For completeness we display the results for all tags present in our data, however to aid interpretation we identify the groupings of features achieving 10.83 and 6.63 critical values with a rule (the latter value is normally regarded as indicating the $p \leq 0.01$ level for the χ^2 distribution; see Sections 2.5.5.1 and 5.3.2 for further discussion).[2]

Across the three tables of results for the Penn POS tag analysis, no features made

[2]Note that no features reached the 15.13 level for either analysis, and that for the reduced POS categories, no features reached the 10.83 critical level.

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
VBD	1	282	0.0413	333	0.0305	94	0.0370	1.36	1.21	1.12	13.93***	2.66	0.87	+		
PDT	2	2	0.0003	19	0.0017	1	0.0004	0.17	0.23	0.74	9.06**	3.34	0.06		+	
VBZ	3	112	0.0164	211	0.0193	65	0.0256	0.85	1.33	0.64	1.98	3.74	7.77**			+
VBN	4	118	0.0173	202	0.0185	66	0.0260	0.93	1.41	0.67	0.34	5.43*	6.73**			+
:	5	74	0.0108	80	0.0073	32	0.0126	1.48	1.72	0.86	5.85*	6.18*	0.50		−	
CC	6	258	0.0378	338	0.0310	88	0.0347	1.22	1.12	1.09	5.80*	0.88	0.50	+		
VBP	7	224	0.0328	434	0.0398	95	0.0374	0.83	0.94	0.88	5.51*	0.29	1.12		+	
.	8	335	0.0491	622	0.0570	120	0.0473	0.86	0.83	1.04	4.88*	3.65	0.13		+	
VBG	9	208	0.0305	272	0.0249	74	0.0292	1.22	1.17	1.05	4.75*	1.39	0.11	+		
RRB	10	27	0.0040	24	0.0022	5	0.0020	1.80	0.90	2.01	4.38*	0.05	2.39	+		
RB	11	558	0.0818	951	0.0871	234	0.0922	0.94	1.06	0.89	1.40	0.60	2.32			
DT	11	499	0.0731	754	0.0691	162	0.0638	1.06	0.92	1.15	0.98	0.84	2.32			
NNP	12	266	0.0390	452	0.0414	89	0.0351	0.94	0.85	1.11	0.61	2.13	0.76			
IN	13	679	0.0995	1100	0.1008	231	0.0910	0.99	0.90	1.09	0.06	2.02	1.40			
RBR	14	4	0.0006	11	0.0010	4	0.0016	0.58	1.56	0.37	0.93	0.54	1.88			
LRB	15	23	0.0034	25	0.0023	5	0.0020	1.47	0.86	1.71	1.77	0.10	1.32			
,	16	208	0.0305	297	0.0272	66	0.0260	1.12	0.96	1.17	1.58	0.11	1.30			
JJ	17	386	0.0566	583	0.0534	127	0.0500	1.06	0.94	1.13	0.77	0.45	1.47			
WRB	17	21	0.0031	46	0.0042	10	0.0039	0.73	0.94	0.78	1.47	0.04	0.40			
CD	18	40	0.0059	72	0.0066	20	0.0079	0.89	1.19	0.74	0.36	0.48	1.12			
MD	19	114	0.0167	191	0.0175	37	0.0146	0.96	0.83	1.15	0.15	1.07	0.53			
FW	20	5	0.0007	6	0.0006	3	0.0012	1.33	2.15	0.62	0.22	1.06	0.41			
NNPS	21	4	0.0006	3	0.0003	1	0.0004	2.13	1.43	1.49	1.00	0.09	0.14			
NNS	22	182	0.0267	275	0.0252	73	0.0288	1.06	1.14	0.93	0.36	0.99	0.29			
NN	23	725	0.1063	1215	0.1113	279	0.1099	0.95	0.99	0.97	0.97	0.03	0.23			
TO	24	228	0.0334	369	0.0338	76	0.0299	0.99	0.89	1.12	0.02	0.95	0.70			
<ENDS>	24	23	0.0034	47	0.0043	9	0.0035	0.78	0.82	0.95	0.95	0.30	0.02			
PRP	25	696	0.1020	1118	0.1024	277	0.1091	1.00	1.07	0.93	0.01	0.89	0.89			
JJS	26	8	0.0012	15	0.0014	5	0.0020	0.85	1.43	0.60	0.13	0.46	0.79			
WP	27	30	0.0044	40	0.0037	8	0.0032	1.20	0.86	1.40	0.57	0.16	0.75			
VB	28	330	0.0484	554	0.0507	121	0.0477	0.95	0.94	1.01	0.48	0.39	0.02			
EX	29	12	0.0018	19	0.0017	6	0.0024	1.01	1.36	0.74	0.00	0.40	0.34			
WDT	30	41	0.0060	66	0.0060	13	0.0051	0.99	0.85	1.17	0.00	0.31	0.26			
UH	31	11	0.0016	21	0.0019	5	0.0020	0.84	1.02	0.82	0.23	0.00	0.13			
RP	32	55	0.0081	93	0.0085	23	0.0091	0.95	1.06	0.89	0.11	0.07	0.22			
JJR	33	10	0.0015	19	0.0017	4	0.0016	0.84	0.91	0.93	0.20	0.03	0.01			
POS	34	24	0.0035	38	0.0035	9	0.0035	1.01	1.02	0.99	0.00	0.00	0.00			

Table 6.1: Penn syntactic tag unigram analysis, Extraversion.

Note. See Table B.1 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
,	1	93	0.0242	297	0.0272	243	0.0359	0.89	1.32	0.67	1.02	10.09**	11.08***			+
NNP	2	114	0.0296	452	0.0414	248	0.0366	0.72	0.88	0.81	10.90***	2.46	3.59		+	
JJ	3	179	0.0465	583	0.0534	419	0.0618	0.87	1.16	0.75	2.68	5.20*	10.56**			+
:	4	50	0.0130	80	0.0073	50	0.0074	1.77	1.01	1.76	9.54**	0.00	7.89**	+		
CC	5	155	0.0403	338	0.0310	210	0.0310	1.30	1.00	1.30	7.09**	0.00	6.01*	+		
VBP	6	183	0.0475	434	0.0398	250	0.0369	1.20	0.93	1.29	4.02*	0.89	6.68**	+		
PRP	7	424	0.1102	1118	0.1024	648	0.0956	1.08	0.93	1.15	1.62	1.93	5.06*	+		
.	8	213	0.0553	622	0.0570	332	0.0490	0.97	0.86	1.13	0.14	4.99*	1.90		+	
VBZ	9	83	0.0216	211	0.0193	163	0.0241	1.12	1.24	0.90	0.70	4.36*	0.67			+
RP	10	46	0.0120	93	0.0085	54	0.0080	1.40	0.94	1.50	3.38	0.15	4.01*	+		
PDT	11	11	0.0029	19	0.0017	8	0.0012	1.64	0.68	2.42	1.63	0.89	3.67			
UH	12	9	0.0023	21	0.0019	6	0.0009	1.22	0.46	2.64	0.23	3.19	3.48			
NNS	12	95	0.0247	275	0.0252	203	0.0300	0.98	1.19	0.82	0.03	3.48	2.49			
VBN	13	63	0.0164	202	0.0185	146	0.0215	0.88	1.16	0.76	0.74	1.95	3.44			
RBR	14	9	0.0023	11	0.0010	11	0.0016	2.32	1.61	1.44	3.32	1.24	0.65			
<ENDS>	15	13	0.0034	47	0.0043	19	0.0028	0.78	0.65	1.20	0.63	2.63	0.26			
IN	16	352	0.0915	1100	0.1008	650	0.0959	0.91	0.95	0.95	2.55	0.99	0.53			
WDT	17	22	0.0057	66	0.0060	54	0.0080	0.95	1.32	0.72	0.05	2.24	1.80			
VB	18	212	0.0551	554	0.0507	327	0.0483	1.09	0.95	1.14	1.02	0.52	2.22			
WRB	19	19	0.0049	46	0.0042	21	0.0031	1.17	0.74	1.59	0.33	1.41	2.12			
VBG	20	102	0.0265	272	0.0249	193	0.0285	1.06	1.14	0.93	0.28	2.01	0.35			
RB	21	308	0.0800	951	0.0871	582	0.0859	0.92	0.99	0.93	1.70	0.07	1.02			
WP	22	11	0.0029	40	0.0037	30	0.0044	0.78	1.21	0.65	0.56	0.61	1.64			
RRB	23	13	0.0034	24	0.0022	16	0.0024	1.54	1.07	1.43	1.48	0.05	0.90			
POS	24	9	0.0023	38	0.0035	25	0.0037	0.67	1.06	0.63	1.25	0.05	1.47			
LRB	25	13	0.0034	25	0.0023	16	0.0024	1.47	1.03	1.43	1.23	0.01	0.90			
JJS	26	6	0.0016	15	0.0014	14	0.0021	1.13	1.50	0.75	0.07	1.19	0.35			
NN	27	416	0.1081	1215	0.1113	779	0.1150	0.97	1.03	0.94	0.27	0.50	1.05			
MD	28	64	0.0166	191	0.0175	105	0.0155	0.95	0.89	1.07	0.13	1.01	0.20			
RBS	29	1	0.0003	1	0.0001	2	0.0003	2.84	3.22	0.88	0.52	0.99	0.01			
TO	30	137	0.0356	369	0.0338	218	0.0322	1.05	0.95	1.11	0.26	0.33	0.85			
FW	31	2	0.0005	6	0.0006	2	0.0003	0.95	0.54	1.76	0.00	0.64	0.32			
VBD	32	108	0.0281	333	0.0305	199	0.0294	0.92	0.96	0.96	0.58	0.18	0.15			
DT	33	279	0.0725	754	0.0691	469	0.0692	1.05	1.00	1.05	0.47	0.00	0.37			
CD	33	22	0.0057	72	0.0066	39	0.0058	0.87	0.87	0.99	0.35	0.47	0.00			
EX	34	5	0.0013	19	0.0017	10	0.0015	0.75	0.85	0.88	0.36	0.18	0.06			
JJR	35	8	0.0021	19	0.0017	14	0.0021	1.19	1.19	1.01	0.17	0.24	0.00			

Table 6.2: Penn syntactic tag unigram analysis, Neuroticism.

Note. See Table B.1 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
:	1	59	0.0135	80	0.0073	36	0.0073	1.84	1.00	1.84	12.14***	0.00	8.63**	+		
VBD	2	134	0.0307	333	0.0305	199	0.0405	1.01	1.33	0.76	0.00	9.82**	6.32*			+
LRB	3	23	0.0053	25	0.0023	13	0.0026	2.30	1.16	1.99	7.99**	0.18	4.10*	+		
VBN	4	64	0.0147	202	0.0185	111	0.0226	0.79	1.22	0.65	2.75	2.83	7.89**			+
RRB	5	22	0.0050	24	0.0022	12	0.0024	2.29	1.11	2.06	7.58**	0.09	4.27*	+		
FW	6	0	0	6	0.0006	5	0.0010	-	1.85	-	-	1.00	6.37*	–		
RP	7	52	0.0119	93	0.0085	35	0.0071	1.40	0.84	1.67	3.61	0.83	5.62*	+		
UH	8	2	0.0005	21	0.0019	6	0.0012	0.24	0.64	0.37	5.56*	1.04	1.65		+	
JJ	9	239	0.0547	583	0.0534	220	0.0448	1.02	0.84	1.22	0.10	5.02*	4.56*			–
NNP	10	183	0.0419	452	0.0414	240	0.0489	1.01	1.18	0.86	0.02	4.26*	2.50			+
PRP	11	401	0.0918	1118	0.1024	516	0.1051	0.90	1.03	0.87	3.59	0.24	4.17*			+
VBP	12	144	0.0330	434	0.0398	183	0.0373	0.83	0.94	0.88	3.91*	0.53	1.23		+	
VBZ	13	84	0.0192	211	0.0193	73	0.0149	0.99	0.77	1.29	0.00	3.88*	2.59		+	
DT	14	342	0.0783	754	0.0691	334	0.0681	1.13	0.99	1.15	3.64	0.05	3.32			
RBR	15	7	0.0016	11	0.0010	2	0.0004	1.59	0.40	3.93	0.88	1.69	3.56			
JJR	16	3	0.0007	19	0.0017	10	0.0020	0.39	1.17	0.34	2.78	0.16	3.20			
VB	17	203	0.0465	554	0.0507	268	0.0546	0.92	1.08	0.85	1.16	0.96	3.02			
WDT	18	19	0.0043	66	0.0060	20	0.0041	0.72	0.67	1.07	1.69	2.55	0.04			
NN	19	506	0.1158	1215	0.1113	519	0.1057	1.04	0.95	1.10	0.57	0.96	2.13			
VBG	20	102	0.0234	272	0.0249	106	0.0216	0.94	0.87	1.08	0.31	1.59	0.32			
EX	21	12	0.0027	19	0.0017	8	0.0016	1.58	0.94	1.69	1.47	0.02	1.34			
TO	22	140	0.0321	369	0.0338	180	0.0367	0.95	1.09	0.87	0.29	0.80	1.44			
WRB	23	16	0.0037	46	0.0042	26	0.0053	0.87	1.26	0.69	0.24	0.85	1.38			
MD	24	87	0.0199	191	0.0175	82	0.0167	1.14	0.95	1.19	0.99	0.12	1.30			
.	24	228	0.0522	622	0.0570	269	0.0548	0.92	0.96	0.95	1.30	0.28	0.29			
RBS	25	0	0	1	0.0001	1	0.0002	-	2.22	-	-	0.31	1.27			
POS	26	12	0.0027	38	0.0035	20	0.0041	0.79	1.17	0.67	0.53	0.32	1.20			
CD	27	36	0.0082	72	0.0066	36	0.0073	1.25	1.11	1.12	1.16	0.27	0.24			
RB	28	357	0.0817	951	0.0871	414	0.0844	0.94	0.97	0.97	1.07	0.30	0.19			
JJS	29	9	0.0021	15	0.0014	7	0.0014	1.50	1.04	1.44	0.89	0.01	0.54			
WP	30	20	0.0046	40	0.0037	21	0.0043	1.25	1.17	1.07	0.65	0.33	0.05			
IN	31	455	0.1042	1100	0.1008	486	0.0990	1.03	0.98	1.05	0.35	0.10	0.60			
PDT	31	7	0.0016	19	0.0017	6	0.0012	0.92	0.70	1.31	0.04	0.60	0.24			
,	32	128	0.0293	297	0.0272	143	0.0291	1.08	1.07	1.01	0.49	0.45	0.00			
NNPS	33	1	0.0002	3	0.0003	2	0.0004	0.83	1.48	0.56	0.03	0.18	0.23			
<ENDS>	34	19	0.0043	47	0.0043	19	0.0039	1.01	0.90	1.12	0.00	0.16	0.13			
NNS	35	114	0.0261	275	0.0252	124	0.0253	1.04	1.00	1.03	0.10	0.00	0.06			
CC	36	138	0.0316	338	0.0310	156	0.0318	1.02	1.03	0.99	0.04	0.07	0.00			

Table 6.3: Penn syntactic tag unigram analysis, Psychoticism.

Note. See Table B.1 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

the conservative 15.13 critical level, however one or two (in the case of Neuroticism) made it to the lower 10.83 critical value. The most significant findings for Extraversion (Table 6.1) are the overuse of past tense verbs (VBD; critical value $\geq$ 10.83), overuse of pre-determiners (PDT) by the Mid group, and the overuse of third-person present tense verbs (VBZ), and past participle verbs (VBN) by the Low Extraverts.

For Neuroticism (Table 6.2), the comma (,) is the most significant feature and is over-used by the Low Neurotics, whilst the High Neurotics show an under-use of singular proper nouns (NNP; both critical value $\geq$ 10.83) and adjectives (JJ), and an overuse of colons or ellipses (:), coordinating conjunctions (CC), and third-person present tense verbs (VBZ).

For Psychoticism (Table 6.3), we find that High Psychotics show an overuse of colons or ellipses (:)(critical value $\geq$ 10.83), and parentheses (LRB, RRB), whilst the Low Psychotics show a tendency towards the overuse of past tense verbs (VBD) and past participle verbs (VBN).

Turning now to the analysis using the more general syntactic category tags (Tables 6.4, 6.5, and 6.6), and we can see that the critical values reached indicating significance are reduced yet further, with these barely reaching the 6.63 level. Again, Neuroticism showed slightly greater significance, with two features—rather than one—reaching this level.

We therefore briefly summarise these findings for general syntactic categories as follows: High Extraverts show a tendency to overuse interjections; Low Extraverts (Introverts) prefer to use more past participle verbs, adverbs, verbs, and pronouns, and use fewer prepositions, adjectives, punctuation markers and 'other' word categories; whilst the Mid group showed an under-use of conjunction, and an overuse of nouns and end of text markers (NA).

For Neuroticism, High Neurotics overuse conjunction, verbs, 'other' categories and punctuation, and under-use nouns, and adverbs; The Low Neurotics show an overuse of adjectives, past participle verbs, and an under-use of pronouns and interjection; The Mid group overuse prepositions and the end of text marker (NA).

For Psychoticism, we find that High Psychotics overuse adjectives, 'other' categories, punctuation, and under-use past participle verbs, interjections, verbs and pro-

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
VPP	1	118	0.0173	202	0.0185	66	0.0260	0.93	1.41	0.67	0.34	5.43*	6.73**			+
CONJ	2	258	0.0378	338	0.0310	88	0.0347	1.22	1.12	1.09	5.80*	0.88	0.50	+		
ADV	3	562	0.0824	963	0.0882	238	0.0938	0.93	1.06	0.88	1.67	0.71	2.76			
PRP	4	679	0.0995	1100	0.1008	231	0.0910	0.99	0.90	1.09	0.06	2.02	1.40			
O	5	1071	0.1570	1714	0.1570	369	0.1454	1.00	0.93	1.08	0.00	1.82	1.64			
VBN	6	1156	0.1695	1804	0.1652	449	0.1769	1.03	1.07	0.96	0.44	1.65	0.60			
<P>	7	667	0.0978	1048	0.0960	228	0.0898	1.02	0.94	1.09	0.14	0.84	1.23			
ADJ	8	404	0.0592	617	0.0565	136	0.0536	1.05	0.95	1.11	0.53	0.32	1.03			
NA	9	23	0.0034	47	0.0043	9	0.0035	0.78	0.82	0.95	0.95	0.30	0.02			
PRN	10	696	0.1020	1118	0.1024	277	0.1091	1.00	1.07	0.93	0.01	0.89	0.89			
NN	11	1177	0.1725	1945	0.1782	442	0.1742	0.97	0.98	0.99	0.76	0.19	0.03			
INT	12	11	0.0016	21	0.0019	5	0.0020	0.84	1.02	0.82	0.23	0.00	0.13			

Table 6.4: Reduced syntactic tag unigram analysis, Extraversion.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
ADJ	1	193	0.0501	617	0.0565	447	0.0660	0.89	1.17	0.76	2.15	6.15*	10.50**			+
CONJ	2	155	0.0403	338	0.0310	210	0.0310	1.30	1.00	1.30	7.09**	0.00	6.01*	+		
NN	3	625	0.1624	1945	0.1782	1230	0.1815	0.91	1.02	0.89	4.13*	0.27	5.22*	−		
PRN	4	424	0.1102	1118	0.1024	648	0.0956	1.08	0.93	1.15	1.62	1.93	5.06*	+		
INT	5	9	0.0023	21	0.0019	6	0.0009	1.22	0.46	2.64	0.23	3.19	3.48			
VPP	6	63	0.0164	202	0.0185	146	0.0215	0.88	1.16	0.76	0.74	1.95	3.44			
VBN	7	688	0.1787	1804	0.1652	1132	0.1671	1.08	1.01	1.07	3.04	0.09	1.94			
NA	8	13	0.0034	47	0.0043	19	0.0028	0.78	0.65	1.20	0.63	2.63	0.26			
PRP	9	352	0.0915	1100	0.1008	650	0.0959	0.91	0.95	0.95	2.55	0.99	0.53			
O	10	627	0.1629	1714	0.1570	1035	0.1528	1.04	0.97	1.07	0.62	0.48	1.60			
ADV	11	318	0.0826	963	0.0882	595	0.0878	0.94	1.00	0.94	1.04	0.01	0.78			
<P>	12	382	0.0992	1048	0.0960	657	0.0970	1.03	1.01	1.02	0.31	0.04	0.13			

Table 6.5: Reduced syntactic tag unigram analysis, Neuroticism.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
VPP	1	64	0.0147	202	0.0185	111	0.0226	0.79	1.22	0.65	2.75	2.83	7.89**			+
INT	2	2	0.0005	21	0.0019	6	0.0012	0.24	0.64	0.37	5.56*	1.04	1.65		+	
ADJ	3	251	0.0575	617	0.0565	237	0.0483	1.02	0.85	1.19	0.05	4.34*	3.69	+		
PRN	4	401	0.0918	1118	0.1024	516	0.1051	0.90	1.03	0.87	3.59	0.24	4.17*			+
VBN	5	667	0.1527	1804	0.1652	829	0.1689	0.92	1.02	0.90	3.07	0.27	3.77			
O	6	743	0.1701	1714	0.1570	773	0.1575	1.08	1.00	1.08	3.29	0.01	2.24			
<P>	7	460	0.1053	1048	0.0960	473	0.0964	1.10	1.00	1.09	2.71	0.00	1.83			
ADV	8	364	0.0833	963	0.0882	417	0.0850	0.94	0.96	0.98	0.86	0.41	0.07			
PRP	9	455	0.1042	1100	0.1008	486	0.0990	1.03	0.98	1.05	0.35	0.10	0.60			
NN	9	804	0.1841	1945	0.1782	885	0.1803	1.03	1.01	1.02	0.60	0.09	0.18			
NA	10	19	0.0043	47	0.0043	19	0.0039	1.01	0.90	1.12	0.00	0.16	0.13			
CONJ	11	138	0.0316	338	0.0310	156	0.0318	1.02	1.03	0.99	0.04	0.07	0.00			

Table 6.6: Reduced syntactic tag unigram analysis, Psychoticism.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

nouns; Low Psychotics overuse prepositions, and under-use the end of text marker (NA); The Mid group overuse adverbs and under-use nouns and conjunctions.

For these POS tag unigram results, we note the generally modest levels of significant differences we found between personality groups. We may take this to indicate that these groups generally use similar proportions of these respective parts of speech. (We do however acknowledge the existence of some differences in behaviour which show relatively greater levels of significance for the finer-grained Penn tags, namely, the High Extravert overuse of past tense verbs and the High Neurotic underuse of singular proper nouns; also we additionally note the overuse of commas by Low Neurotics and overuse of colon or ellipsis by High Psychotics.) The reduction of POS tags to broader grammatical categories decreases the significance further.

Even though the parts of speech may not be used differently in terms of proportions, they may occur in different contexts or sequences, thus indicating differences in they way they are used. We therefore turn to the results for the use of n-gram analysis of the syntactic tag data.

6.2.3 N-gram Syntactic Analysis Results

The results showing the combination of 1–5 n-gram features are shown in the following tables. The analysis using the Penn POS tags for Extraversion, Neuroticism, and Psychoticism is presented in Tables 6.7, 6.8, and 6.9, and the results using the reduced syntactic category tags are found in tables 6.10, 6.11 , and 6.12. Due to these features achieving greater levels of significance, here we only display those which reach the critical value of 15.13.

Examining first the n-gram features derived from the Penn POS tag analysis (tables 6.7, 6.8, and 6.9), and here we can see that not only is there a greater number of features which show significance of greater than the 15.13 significance value (greatest in the case of Neuroticism, and least from Psychoticism), but that these features are predominantly bigrams (the exception being the longer n-grams for punctuation found for Neuroticism). In interpreting this data, we look for parts of speech which pattern in the most distinctive way for the different personality groups.

For Extraversion, the most prominent feature is the patterning of adverbs: although

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
CC DT	1	32	0.0047	0	0	0	0	-	-	-	61.16****	-	20.24****	+		
RB TO	1	32	0.0047	0	0	0	0	-	-	-	61.16****	-	20.24****	+		
TO DT	2	29	0.0043	0	0	15	0.0059	-	-	0.72	55.43****	50.05****	1.03		−	
, IN	3	0	0	0	0	15	0.0059	-	-	-	-	50.05****	39.16****			+
CC NNP	4	26	0.0038	0	0	0	0	-	-	-	49.69****	-	16.44****	+		
PRP RB	5	0	0	0	0	14	0.0055	-	-	-	-	46.71****	36.55****			+
CC VBD	6	0	0	0	0	12	0.0047	-	-	-	-	40.04****	31.33****			+
, RB	7	0	0	44	0.0040	13	0.0051	-	1.27	-	-	0.55	33.94****	−		
NNS ,	8	16	0.0023	0	0	0	0	-	-	-	30.58****	-	10.12**	+		
: PRP	9	0	0	0	0	9	0.0035	-	-	-	-	30.03****	23.49****			+
VBD VBN	9	0	0	0	0	9	0.0035	-	-	-	-	30.03****	23.49****			+
NN NN	10	45	0.0066	93	0.0085	0	0	0.77	-	-	2.04	-	28.46****			−
VB DT	11	44	0.0065	81	0.0074	0	0	0.87	-	-	0.57	-	27.83****			−
CD IN	12	0	0	0	0	8	0.0032	-	-	-	-	26.69****	20.88****			+
JJ TO	13	0	0	36	0.0033	10	0.0039	-	1.20	-	-	0.24	26.10****	−		
CC RB	14	41	0.0060	49	0.0045	0	0	1.34	-	-	1.88	-	25.93****			−
NN RB	14	41	0.0060	63	0.0058	0	0	1.04	-	-	0.04	-	25.93****			−
NNP CC	15	40	0.0059	45	0.0041	0	0	1.42	-	-	2.60	-	25.30****			−
CC NN	16	12	0.0018	0	0	0	0	-	-	-	22.94****	-	7.59**	+		
VBG IN	17	36	0.0053	55	0.0050	0	0	1.05	-	-	0.05	-	22.77****			−
, PRP	18	34	0.0050	52	0.0048	0	0	1.05	-	-	0.04	-	21.50****			−
WP PRP	19	11	0.0016	0	0	5	0.0020	-	-	0.82	21.02****	16.68****	0.13		−	
VBN IN	20	33	0.0048	35	0.0032	0	0	1.51	-	-	2.84	-	20.87****			−
VB TO	21	32	0.0047	43	0.0039	0	0	1.19	-	-	0.55	-	20.24****			−
JJ .	21	32	0.0047	57	0.0052	0	0	0.90	-	-	0.24	-	20.24****			−
NNP ,	22	31	0.0045	43	0.0039	0	0	1.15	-	-	0.37	-	19.61****			−
NN VBG	23	10	0.0015	0	0	0	0	-	-	-	19.11****	-	6.32*	+		
. .	24	30	0.0044	17	0.0016	0	0	2.82	-	-	12.33***	-	18.97****	+		
VBD DT	25	61	0.0089	41	0.0038	17	0.0067	2.38	1.78	1.33	18.94****	3.69	1.17	+		
MD RB	26	29	0.0043	57	0.0052	0	0	0.81	-	-	0.83	-	18.34****			−
VBD TO	27	28	0.0041	39	0.0036	0	0	1.15	-	-	0.31	-	17.71****			−
: VBD	28	9	0.0013	0	0	0	0	-	-	-	17.20****	-	5.69*	+		
NNP RB	28	9	0.0013	0	0	0	0	-	-	-	17.20****	-	5.69*	+		
VBN RP	29	0	0	11	0.0010	6	0.0024	-	2.35	-	-	2.54	15.66****	−		
NN RRB	30	8	0.0012	0	0	0	0	-	-	-	15.29****	-	5.06*	+		

Table 6.7: Penn syntactic tag n-gram analysis, Extraversion.

Note. See Table B.1 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
VB IN	1	0	0	0	0	47	0.0069	-	-	-	-	90.23****	42.28****			+
, IN	2	0	0	0	0	36	0.0053	-	-	-	-	69.11****	32.39****			+
PRP RB	3	0	0	0	0	34	0.0050	-	-	-	-	65.27****	30.59****			+
RB RB	4	0	0	105	0.0096	69	0.0102	-	1.06	-	-	0.14	62.08****	−		
VBG RB	5	0	0	0	0	29	0.0043	-	-	-	-	55.67****	26.09****			+
VBN RB	6	0	0	0	0	27	0.0040	-	-	-	-	51.84****	24.29****			+
VB DT	7	0	0	81	0.0074	48	0.0071	-	0.95	-	-	0.06	43.18****	−		
VB PRP	8	0	0	79	0.0072	46	0.0068	-	0.94	-	-	0.12	41.38****	−		
NN VBZ	9	20	0.0052	46	0.0042	0	0	1.23	-	-	0.60	-	40.61****			−
NNS ,	10	0	0	0	0	21	0.0031	-	-	-	-	40.32****	18.89****			+
. . .	11	22	0.0057	4	0.0004	3	0.0004	15.61	1.21	12.91	39.26****	0.06	29.03****	+		
RB .	12	0	0	91	0.0083	43	0.0063	-	0.76	-	-	2.23	38.69****	−		
. . . .	13	17	0.0044	1	0.0001	1	0.0001	48.24	1.61	29.93	38.61****	0.11	27.70****	+		
CC VB	14	19	0.0049	35	0.0032	0	0	1.54	-	-	2.19	-	38.58****			−
. . . . .	15	14	0.0036	0	0	0	0	-	-	-	37.66****	-	28.44****	+		
: :	16	14	0.0036	0	0	0	0	-	-	-	37.65****	-	28.43****	+		
, RB	17	0	0	44	0.0040	41	0.0061	-	1.50	-	-	3.47	36.89****	−		
NN RB	18	0	0	63	0.0058	40	0.0059	-	1.02	-	-	0.01	35.99****	−		
JJ .	19	0	0	57	0.0052	37	0.0055	-	1.05	-	-	0.05	33.29****	−		
RB ,	20	0	0	67	0.0061	36	0.0053	-	0.87	-	-	0.49	32.39****	−		
: : :	21	12	0.0031	0	0	0	0	-	-	-	32.27****	-	24.37****	+		
. .	22	29	0.0075	17	0.0016	6	0.0009	4.84	0.57	8.51	27.65****	1.53	32.22****	+		
RB PRP	23	0	0	45	0.0041	35	0.0052	-	1.25	-	-	0.99	31.49****	−		
CC RB	23	0	0	49	0.0045	35	0.0052	-	1.15	-	-	0.40	31.49****	−		
RB VBN	23	0	0	51	0.0047	35	0.0052	-	1.11	-	-	0.21	31.49****	−		
IN NNS	24	0	0	47	0.0043	31	0.0046	-	1.06	-	-	0.07	27.89****	−		
: : : :	25	10	0.0026	0	0	0	0	-	-	-	26.90****	-	20.31****	+		
CC VBD	26	0	0	0	0	14	0.0021	-	-	-	-	26.88****	12.60***			+
: PRP	27	0	0	0	0	13	0.0019	-	-	-	-	24.96****	11.70***			+
. DT	28	0	0	0	0	12	0.0018	-	-	-	-	23.04****	10.80**			+
VBN IN	29	0	0	35	0.0032	25	0.0037	-	1.15	-	-	0.29	22.49****	−		
: : : : :	30	8	0.0021	0	0	0	0	-	-	-	21.52****	-	16.25****	+		
CD IN	31	0	0	0	0	11	0.0016	-	-	-	-	21.12****	9.90**			+
VBZ VBG	32	0	0	31	0.0028	23	0.0034	-	1.20	-	-	0.42	20.69****	−		
VBZ DT	33	0	0	30	0.0027	22	0.0032	-	1.18	-	-	0.35	19.79****	−		
NNP RB	34	0	0	0	0	10	0.0015	-	-	-	-	19.20****	9.00**			+
WP PRP	34	0	0	0	0	10	0.0015	-	-	-	-	19.20****	9.00**			+
NNP CC	35	0	0	45	0.0041	20	0.0030	-	0.72	-	-	1.60	17.99****	−		
NNS .	35	0	0	37	0.0034	20	0.0030	-	0.87	-	-	0.25	17.99****	−		
NNS :	36	0	0	0	0	9	0.0013	-	-	-	-	17.28****	8.10**			+
NN CC DT	37	8	0.0021	7	0.0006	0	0	3.24	-	-	5.02*	-	16.25****	+		
PRP .	37	8	0.0021	41	0.0038	0	0	0.55	-	-	2.66	-	16.25****			−
NN RRB	38	6	0.0016	0	0	6	0.0009	-	-	1.76	16.14****	11.52***	0.95		−	
NNP VBZ	39	0	0	18	0.0016	17	0.0025	-	1.52	-	-	1.52	15.29****	−		

Table 6.8: Penn syntactic tag n-gram analysis, Neuroticism.

Note. See Table B.1 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

Table 6.9: Penn syntactic tag n-gram analysis, Psychoticism.

Note. See Table B.1 for a description of the tags. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
VB IN	1	33	0.0076	0	0	0	0	-	-	-	82.68****	-	49.71****	+		
NN NN	2	31	0.0071	93	0.0085	0	0	0.83	-	-	0.80	-	46.69****			−
VBD VBN	3	10	0.0023	0	0	19	0.0039	-	-	0.59	25.05****	44.49****	1.89		−	
RB .	4	24	0.0055	67	0.0061	0	0	0.90	-	-	0.22	-	36.15****			−
VBD TO	5	0	0	39	0.0036	28	0.0057	-	1.60	-	-	3.45	35.65****	−		
NNS ,	6	0	0	0	0	15	0.0031	-	-	-	-	35.12****	19.10****			+
NN :	7	23	0.0053	21	0.0019	0	0	2.74	-	-	10.85***	-	34.64****	+		−
, RB	8	22	0.0050	44	0.0040	0	0	1.25	-	-	0.71	-	33.14****			
RB PRP	9	0	0	45	0.0041	24	0.0049	-	1.19	-	-	0.45	30.55****	−		
CC RB	9	0	0	49	0.0045	24	0.0049	-	1.09	-	-	0.12	30.55****	−		
VBN RB	10	12	0.0027	0	0	0	0	-	-	-	30.06****	-	18.08****	+		
. RB	11	0	0	0	0	12	0.0024	-	-	-	-	28.10****	15.28****			+

both use the same syntactic item, it is put to different use. Here we see an Introvert under use ([*PRP RB*], [*CC RB*], [*NN RB*], [*MD RB*]), combined with an Extravert overuse ([*RB TO*], [*NNP RB*]). The sole underuse by Extraverts is [. *RB*] which contrasts nicely with the Introvert avoidance of [*CC RB*], suggesting that the two personality groups have different strategies (co-ordination vs. new sentence) when providing adverbial information.

Secondly for Extraversion, we note the behaviour of nouns: the Extravert overuse of [*CC NNP*], [*NNS ,*], [*CC NN*], [*NN VBG*], [*NNP RB*], [*NN RRB*], and the Introvert under-use of [*NN NN*], [*NN RB*], [*NNP CC*]. This apparently demonstrates a significant and prominent usage of patterns composed of nouns by Extraverts. However, this does not necessarily contradict our Grammatical hypothesis, since as the unigram analysis has shown, there is little difference in relative frequency in the use of nouns. Rather this shows that they are used with a particular pattern by the different groups.

Finally, displaying a similar pattern is the use of coordinating conjunction, with Extraverts over-using this feature ([*CC DT*], [*CC NNP*], [*CC NN*], additionally Introverts show underuse of [*NNP CC*]. However, Introverts display an overuse of coordination followed by (past tense) verb ([*CC VBD*]).

In each of these cases the contrast in usage is very distinct with under-usage of a feature meaning that whilst other groups used the feature extensively, the under-using group did not use the feature at all, and in the case of overuse, the group in question used the feature exclusively.

For Neuroticism, as in the analysis of other data, we find that overuse of punctuation colons and ellipses, and sentence terminators (.?) is a feature of language of individuals on the high end of the scale. In terms of syntactic features for this dimension, as with Extraversion, we find that the use of adverbs is important: Here we find that Low Neurotics show an overuse ([*RB RB*], [*VBG RB*], [*VBN RB*], [*NNP RB*]), combined with a High Neurotic under-use ([*PRP RB*], [*RB .*], [, *RB*], [*NN RB*], [*RB ,*], [*RB PRP*], [*CC RB*], [*RB VBN*]). This strongly indicates the role of adverbial formation, rather than frequency, as a feature of Low Neurotic language.

Similarly to the Extravert findings is the salience of the usage of nouns in distinguishing between High and Low Neurotics. However, here we find a difference in the

types of nouns found across the groups: common plural nouns (NNP) and proper singular nouns (NNS) show an overuse by Low Neurotics ([*NNS ,*], [*NNP RB*], [*NNS :*]) and an under-use by High Neurotics ([*IN NNS*], [*NNP CC*], [*NNS .*], [*NNP VBZ*]); In contrast the use of common, singular or mass nouns (NN) is not as clearly divided, with some being avoided by Low Neurotics ([*NN VBZ*], [*NN CC DT*]), whilst others are avoided by the High Neurotics ([*NN RB*]), while the use of such a noun as the last item within parentheses is used by both the High and Low Neuroticism groups, but avoided by the Mid.

The patterning of verbs within collocations is apparently less frequent within the language of High Neurotics ([*VB DT*], [*VB PRP*], [*RB VBN*], [*VBN IN*], [*VBZ VBG*], [*VBZ DT*], [*NNP VBZ*]), and more frequent in that of the Low Neurotics ([*VB IN*], [*VGB RB*], [*VBN RB*], [*CC VBD*]). Note also the neat contrast between the final feature found more frequently in the language of Low Neurotics—conjunction followed by past tense verb—and that which is avoided by Low Neurotics, conjunction followed by a verb in the base form ([*CC VB*]). This group also appears to avoid common or mass noun followed by third-person singular present tense verb ([*NN VBZ*]).

Turning to the features characteristic of Psychoticism, and we find that although there are fewer which make it past the critical value of 15.13, again the collocation of adverbs, verbs and noun are important, along with the Low Psychotics' overuse of sentence terminator bigrams. Here, the use of adverbs is more complicated in their distribution: High Psychotics overuse the pattern [*VBN RB*], whilst under-using [*RB PRP*], and [*CC RB*]; Low Psychotics under-use [*RB ,*], and [*, RB*], whilst showing contrast by over-using [*. RB*] and [*: RB*] (following an ellipsis).

Again, this more complicated patterning of Psychoticism features is reflected in their use of nouns: High Psychotics overuse noun as the final element of a parenthesised section ([*NN RRB*]); Low Psychotics show a reduced usage of compounded nouns ([*NN NN*]), and other noun combinations ([*NN :*], [*NNS .*], [*NNS VBG*], [*VB NN*]), but an overuse of other noun patterns ([*NNS ,*], [*NN WDT*]).

For verbs we find that High Psychotics overuse [*VB IN*], [*VBN RB*], [*VBD VBG*] patterns, but under-use [*VBD TO*], [*VBG IN*], and [*VBZ DT*]; Low Psychotics under-use [*VBZ VBG*] and [*VB NN*], with both High and Low Psychotics over-using [*VBD*

VBN] relative to the Mid group. Unlike the Extravert use of Nouns, and the Low Neurotic use of Adverbs, for Psychoticism there does not appear to be a strong use of one category over another. Instead this personality dimension appears to be displayed in the specific ways that these grammatical categories collocate.

Turning now to the n-gram analysis using the reduced syntactic category tags (Tables 6.10, 6.11, and 6.12), and here we again see the High Extravert preference for collocations containing adverbs, nouns and conjunctions; for Neuroticism, the Low Neurotics are distinguished by their avoidance of multiple punctuation devices, and overuse of adverbs in collocation contexts. Although this analysis of reduced categories fails to show the distinctive patterning between the High and Low groups in their use of different noun forms, it does however appear to display a High Neurotic preference for collocations with the general category of 'verbs' and the Low Neurotic tendency to use 'past participle verb' forms; High Psychotics are again distinguished by their avoidance of long sequences of punctuation features, with their general overuse of verb collocations only reflected in their use of 'past participle verb' followed by 'adverb', with additionally other overuse of adverb collocations only showing significance for Low Psychotics. Perhaps unexpectedly, High Psychotics in this reduced category analysis generally show an overuse of noun collocations.

6.3 Semantic Analysis of the Corpus

6.3.1 Method

Using the same subcorpora as before, we pre-processed this using the Wmatrix implementation of the CLAWS tagger (Rayson, 2001, 2003), which provided output of the words, along with associated POS and semantic tags (Wilson and Rayson, 1993; Garside and Rayson, 1997; Piao et al., 2003, see Tables B.2 – B.4 for a description of the tags, and Section 2.5.3.3 for further information about this annotation method).

Additional scripts were then used to extract the semantic tags (and punctuation/ structural information) alone, and in the case of multiple tag probabilities, 'slash tags' (where a word can belong to more than one semantic classification), or multi-word units (where a group of words have additionally been tagged as an [idiomatic] phrase),

Table 6.10: Reduced syntactic tag n-gram analysis, Extraversion.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001$. $df = 1$.

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> ADV	1	76	0.0111	0	0	36	0.0142	-	-	0.79	145.26****	120.11****	1.39		−	
<P> NN	2	68	0.0100	0	0	25	0.0099	-	-	1.01	129.97****	83.41****	0.00		−	
CONJ VBN	3	60	0.0088	0	0	0	0	-	-	-	114.68****	-	37.95****	+		
ADV PRP	4	0	0	0	0	34	0.0134	-	-	-	-	113.44****	88.76****			+
NN NN	5	116	0.0170	220	0.0202	0	0	0.84	-	-	2.23	-	73.36****			−
ADV <P>	6	89	0.0130	180	0.0165	0	0	0.79	-	-	3.34	-	56.29****			−
PRN NN	7	75	0.0110	91	0.0083	0	0	1.32	-	-	3.11	-	47.43****			−
PRN ADV	8	0	0	0	0	14	0.0055	-	-	-	-	46.71****	36.55****			+
<P> O	9	71	0.0104	116	0.0106	0	0	0.98	-	-	0.02	-	44.90****			−
ADV O	10	65	0.0095	101	0.0093	0	0	1.03	-	-	0.03	-	41.11****			−
ADJ <P>	11	56	0.0082	88	0.0081	0	0	1.02	-	-	0.01	-	35.42****			−
NN ADV	12	55	0.0081	109	0.0100	0	0	0.81	-	-	1.71	-	34.78****			−
CONJ ADV	13	41	0.0060	49	0.0045	0	0	1.34	-	-	1.88	-	25.93****			−
VBN PRN O	14	0	0	36	0.0033	8	0.0032	-	0.96	-	-	0.01	20.89****	−		
VPP PRP	15	33	0.0048	35	0.0032	0	0	1.51	-	-	2.84	-	20.87****			−
ADJ O	15	33	0.0048	58	0.0053	0	0	0.91	-	-	0.19	-	20.87****			−
<P> ADJ	16	25	0.0037	23	0.0021	0	0	1.74	-	-	3.65	-	15.81****			−
PRN O ADV	17	20	0.0029	51	0.0047	1	0.0004	0.63	0.08	7.44	3.31	14.76***	7.22**			−
VBN O NN <P>	18	25	0.0037	11	0.0010	3	0.0012	3.64	1.17	3.10	14.15***	0.06	4.57*	+		
PRN O ADV VBN	19	19	0.0028	47	0.0043	1	0.0004	0.65	0.09	7.06	2.71	13.25***	6.68**			−
VBN PRN ADV	20	0	0	18	0.0016	5	0.0020	-	1.20	-	-	0.12	13.05***	−		
CONJ VBN PRN	20	0	0	19	0.0017	5	0.0020	-	1.13	-	-	0.06	13.05***	−		
VBN <P> PRN	21	2	0.0003	17	0.0016	8	0.0032	0.19	2.03	0.09	7.54**	2.46	12.14***	−		
<P> O VBN ADJ <P>	22	6	0.0009	0	0	0	0	-	-	-	11.47***	-	3.79	+		
<P> <P> <P>	23	18	0.0026	12	0.0011	0	0	2.40	-	-	5.67*	-	11.38***	+		

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
VBN PRP	1	84	0.0218	0	0	0	0	-	-	-	225.90****	-	170.58****	+		
<P> ADV	2	0	0	0	0	82	0.0121	-	-	-	-	157.42****	73.77****			+
<P> O	3	38	0.0099	116	0.0106	0	0	0.93	-	-	0.16	-	77.17****			−
PRN ADV	4	0	0	0	0	35	0.0052	-	-	-	-	67.19****	31.49****			+
ADV ADV	5	0	0	110	0.0101	73	0.0108	-	1.07	-	-	0.20	65.68****	−		
<P> <P> <P> <P> <P>	6	24	0.0062	0	0	0	0	-	-	-	64.56****	-	48.75****	+		
<P> <P> <P> <P>	7	29	0.0075	2	0.0002	1	0.0001	41.15	0.81	51.06	64.38****	0.03	51.03****	+		
ADJ <P>	8	0	0	88	0.0081	67	0.0099	-	1.23	-	-	1.57	60.28****	−		
ADV O	9	0	0	101	0.0093	63	0.0093	-	1.01	-	-	0.00	56.68****	−		
<P> <P>	10	59	0.0153	48	0.0044	17	0.0025	3.49	0.57	6.11	40.46****	4.28*	54.32****	+		
VPP ADV	11	0	0	0	0	27	0.0040	-	-	-	-	51.84****	24.29****			+
O ADV	12	0	0	89	0.0082	57	0.0084	-	1.03	-	-	0.03	51.28****	−		
<P> <P> <P>	13	36	0.0094	12	0.0011	4	0.0006	8.51	0.54	15.85	50.08****	1.27	50.70****	+		
VBN PRN O	14	24	0.0062	36	0.0033	0	0	1.89	-	-	5.53*	-	48.74****	+		
ADV PRN	15	0	0	45	0.0041	35	0.0052	-	1.25	-	-	0.99	31.49****	−		
CONJ ADV	15	0	0	49	0.0045	35	0.0052	-	1.15	-	-	0.40	31.49****	−		
ADV VPP	15	0	0	52	0.0048	35	0.0052	-	1.08	-	-	0.14	31.49****	−		
PRN ADJ	16	0	0	35	0.0032	28	0.0041	-	1.29	-	-	0.99	25.19****	−		
VPP PRP	17	0	0	35	0.0032	25	0.0037	-	1.15	-	-	0.29	22.49****	−		
ADJ PRN VBN	18	10	0.0026	2	0.0002	5	0.0007	14.19	4.03	3.52	17.29****	3.15	5.71*	+		
PRP ADJ	19	8	0.0021	39	0.0036	0	0	0.58	-	-	2.18	-	16.25****			−
VBN O VBN ADV	20	5	0.0013	23	0.0021	1	0.0001	0.62	0.07	8.80	1.06	15.81****	5.65*			−
PRN <P> ADV	21	2	0.0005	0	0	8	0.0012	-	-	0.44	5.38*	15.36****	1.25		−	
PRN VBN PRN O ADV	22	7	0.0018	3	0.0003	0	0	6.62	-	-	8.42**	-	14.22***	+		
VBN ADJ CONJ	22	7	0.0018	7	0.0006	0	0	2.84	-	-	3.65	-	14.22***			−
ADJ PRN	23	11	0.0029	5	0.0005	7	0.0010	6.24	2.26	2.77	12.73***	1.97	4.58*	+		
VBN PRN O ADV VBN	24	8	0.0021	2	0.0002	1	0.0001	11.35	0.81	14.09	12.72***	0.03	10.87***	+		
VBN PRN O ADV	25	9	0.0023	3	0.0003	2	0.0003	8.51	1.07	7.92	12.52***	0.01	9.65**	+		
NN VBN O ADJ	26	3	0.0008	0	0	6	0.0009	-	-	0.88	8.07**	11.52***	0.03		−	
NN VBN O ADJ NN	26	2	0.0005	0	0	6	0.0009	-	-	0.59	5.38*	11.52***	0.46		−	
PRN O VBN <P>	26	2	0.0005	0	0	6	0.0009	-	-	0.59	5.38*	11.52***	0.46		−	
ADV PRN VBN PRN	27	6	0.0016	1	0.0001	1	0.0001	17.03	1.61	10.56	11.00***	0.11	7.34**	+		

Table 6.11: Reduced syntactic tag n-gram analysis, Neuroticism.
Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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> PRP	1	0	0	0	0	28	0.0057	-	-	-	-	65.56****	35.65****			+
ADV PRN	2	0	0	45	0.0041	24	0.0049	-	1.19	-	-	0.45	30.55****	−		
CONJ ADV	2	0	0	49	0.0045	24	0.0049	-	1.09	-	-	0.12	30.55****	−		
<P> <P>	3	20	0.0046	48	0.0044	0	0	1.04	-	-	0.02	-	30.13****			−
VPP ADV	4	12	0.0027	0	0	0	0	-	-	-	30.06****	-	18.08****	+		
ADJ O	5	17	0.0039	58	0.0053	0	0	0.73	-	-	1.34	-	25.61****			−
NN <P> NN <P>	6	21	0.0048	11	0.0010	13	0.0027	4.77	2.63	1.82	18.84****	5.51*	2.95		−	
<P> NN <P>	7	30	0.0069	23	0.0021	17	0.0035	3.26	1.64	1.98	18.10****	2.34	5.32*	+		
O NN <P> CONJ	8	10	0.0023	11	0.0010	0	0	2.27	-	-	3.39	-	15.06***			−
<P> PRP PRN ADV VBN	9	6	0.0014	0	0	0	0	-	-	-	15.04***	-	9.04**	+		
<P> PRP PRN ADV	10	6	0.0014	0	0	0	0	-	-	-	15.03***	-	9.04**	+		
PRP O NN <P>	11	46	0.0105	85	0.0078	19	0.0039	1.35	0.50	2.72	2.64	8.70**	14.93***			−
PRP PRN ADV VBN	12	12	0.0027	4	0.0004	1	0.0002	7.50	0.56	13.48	14.77***	0.31	12.30***	+		
VPP PRP PRN NN	13	0	0	0	0	6	0.0012	-	-	-	-	14.05***	7.64**			+
VPP PRP PRN NN <P>	13	0	0	0	0	6	0.0012	-	-	-	-	14.05***	7.64**			+
PRP PRN NN	14	6	0.0014	41	0.0038	28	0.0057	0.37	1.52	0.24	6.72**	2.82	13.00***	−		
PRN ADJ	15	0	0	35	0.0032	10	0.0020	-	0.64	-	-	1.73	12.73***	−		
<P> NN <P> NN <P>	16	5	0.0011	0	0	4	0.0008	-	-	1.40	12.53***	9.37**	0.26		−	
PRP ADV PRN	16	5	0.0011	0	0	2	0.0004	-	-	2.81	12.53***	4.68*	1.70		−	
<P> <P> <P>	17	1	0.0002	12	0.0011	14	0.0029	0.21	2.60	0.08	3.53	5.80*	11.98***			+
<P> PRN VBN VPP ADV	18	2	0.0005	0	0	5	0.0010	-	-	0.45	5.01*	11.71***	1.00		−	
NN NN PRP NN O	18	1	0.0002	0	0	5	0.0010	-	-	0.22	2.51	11.71***	2.46			+
NA NN	19	0	0	25	0.0023	9	0.0018	-	0.80	-	-	0.34	11.46***	−		
<P> PRN VBN ADV	20	12	0.0027	73	0.0067	38	0.0077	0.41	1.16	0.35	9.98**	0.53	11.34***	−		
VBN <P> CONJ	21	3	0.0007	1	0.0001	7	0.0014	7.50	15.57	0.48	3.69	11.11***	1.21			+

Table 6.12: Reduced syntactic tag n-gram analysis, Psychoticism.

Note. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

only the first tag for the word was retained, and the rest of the semantic tag information disregarded. To gather more general information about the semantic tagging of a text, further versions of the semantically tagged subcorpora were built by further reducing the tags so that they consisted of the letter code ('Reduced' version) and the highest-order number, and letter code alone ('Most Reduced' version).

Each of these different versions of the semantically tagged subcorpora were then subject to the stratified corpus comparison analysis, to give frequency and relative frequency information for High, Low and Mid groups, with additional relative-frequency ratio and log-likelihood information for High-Low, High-Mid, and Low-Mid comparisons.

6.3.2 Unigram Results

The results showing the unigram features are displayed in the following tables (reaching a critical value of 6.63 or greater). The analysis using the full USAS semantic tags for Extraversion, Neuroticism, and Psychoticism is presented in tables 6.13, 6.14, and 6.15, the analysis using the reduced tags is shown in tables 6.16, 6.17, and 6.18, and finally, the analysis using the most reduced semantic tags is shown in tables 6.19, 6.20, and 6.21. Critical values of 15.13 and 10.83 are marked on these tables with a rule where appropriate.

The unigram results for the full semantic tags (Tables 6.13, 6.14, and 6.15) show that the only distinguishing reference for Extraversion is to Cigarettes and drugs (F3), which is overused by Introverts; Less significant use of semantic references shown by Extraverts are the overuse of general living creatures (L2mfn), liking (E2), linear order (N4), and an underuse of degree approximators (A13.4), and new clauses (<NC>); Introverts show an overuse of arts and crafts (C1), entertainment generally (K1), and an underuse of references to comparing different (A6.1), and personal names (Z1mf).

For Neuroticism, we find that significant patterns are an overuse of references to medicines and medical treatment (B3), and of unmatched words (Z99) by the Low Neurotics; for High Neurotics there is an overuse of references to negative affiliation and group processes (S5−); both High and Low Neurotics underuse happy (E4.1) references. Less significantly, we find that High Neurotics overuse references to intimate or

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
F3	1	0	0	0	0	8	0.0030	-	-	-	-	26.76****	20.83****			+
C1	2	2	0.0003	1	0.0001	6	0.0022	3.23	25.96	0.12	0.99	14.75***	7.90**			+
A13.4	3	4	0.0006	33	0.0029	10	0.0037	0.20	1.31	0.15	14.13***	0.54	11.83***	−		
A6.1−	4	32	0.0045	45	0.0039	1	0.0004	1.15	0.10	11.96	0.36	12.42***	13.96***			−
L2mfn	5	0	0	7	0.0006	5	0.0019	-	3.09	-	-	3.34	13.02***	−		
K1	6	37	0.0052	48	0.0041	27	0.0101	1.25	2.43	0.51	1.00	12.28***	6.64**			+
Z1mf	7	30	0.0042	54	0.0047	2	0.0007	0.90	0.16	5.60	0.22	11.89***	9.29**			−
E2	8	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
<NC>	9	291	0.0406	595	0.0514	119	0.0445	0.79	0.87	0.91	11.05***	2.14	0.68		+	
N4	10	85	0.0119	82	0.0071	24	0.0090	1.68	1.27	1.32	11.04***	0.99	1.54	+		
N6+	11	23	0.0032	46	0.0040	23	0.0086	0.81	2.16	0.37	0.71	8.24**	10.72**			+
T1.1.1	12	35	0.0049	104	0.0090	19	0.0071	0.54	0.79	0.69	10.61**	0.94	1.65		+	
T1.3−	13	5	0.0007	0	0	0	0	-	-	-	9.62**	-	3.17	+		
K2	14	15	0.0021	13	0.0011	0	0	1.87	-	-	2.70	-	9.52**			−
Z4	15	69	0.0096	168	0.0145	30	0.0112	0.66	0.77	0.86	8.63**	1.80	0.47		+	
Y1	16	7	0.0010	1	0.0001	0	0	11.32	-	-	8.40**	-	4.44*	+		
X2.2+	17	8	0.0011	36	0.0031	6	0.0022	0.36	0.72	0.50	8.34**	0.59	1.58		+	
N3.2−	18	1	0.0001	9	0.0008	5	0.0019	0.18	2.40	0.07	4.09*	2.22	8.25**	−		
S2.2m	18	13	0.0018	16	0.0014	0	0	1.31	-	-	0.53	-	8.25**			−
X2.4	19	4	0.0006	13	0.0011	8	0.0030	0.50	2.66	0.19	1.66	4.26*	8.09**			+
A9+	20	136	0.0190	158	0.0136	49	0.0183	1.39	1.34	1.04	7.88**	3.07	0.05	+		
T1.2	21	27	0.0038	19	0.0016	5	0.0019	2.30	1.14	2.02	7.87**	0.07	2.42	+		
Z6	22	69	0.0096	165	0.0143	43	0.0161	0.68	1.13	0.60	7.83**	0.48	6.60*	−		
S2mf	23	0	0	5	0.0004	3	0.0011	-	2.60	-	-	1.53	7.81**	−		
T3−−−	23	0	0	10	0.0009	3	0.0011	-	1.30	-	-	0.15	7.81**	−		
S4	24	12	0.0017	5	0.0004	0	0	3.88	-	-	7.30**	-	7.62**	+		
B5	24	12	0.0017	11	0.0009	0	0	1.76	-	-	1.84	-	7.62**			−
Y2	24	11	0.0015	4	0.0003	2	0.0007	4.45	2.16	2.06	7.62**	0.72	1.03	+		
A2.1+	25	7	0.0010	32	0.0028	10	0.0037	0.35	1.35	0.26	7.58**	0.66	7.45**	−		
X2.2−	26	8	0.0011	2	0.0002	0	0	6.47	-	-	7.31**	-	5.08*	+		
T3−	27	22	0.0031	24	0.0021	14	0.0052	1.48	2.52	0.59	1.76	6.80**	2.31			+
P1	28	27	0.0038	54	0.0047	4	0.0015	0.81	0.32	2.52	0.83	6.73**	3.72		+	

Table 6.13: Semantic (full) tag unigram analysis, Extraversion.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
B3	1	2	0.0005	2	0.0002	18	0.0025	2.84	14.58	0.19	1.04	23.59****	7.31**			+
S5–	2	12	0.0029	2	0.0002	1	0.0001	17.02	0.81	21.01	21.99****	0.03	18.14****	+		
Z99	3	44	0.0108	92	0.0079	104	0.0146	1.36	1.83	0.74	2.66	17.82****	2.89			+
E4.1+	4	8	0.0020	57	0.0049	10	0.0014	0.40	0.28	1.40	7.45**	17.60****	0.50		+	
S3.2	5	17	0.0042	22	0.0019	6	0.0008	2.19	0.44	4.96	5.58*	3.62	13.42***	+		
Z2	6	12	0.0029	88	0.0076	48	0.0067	0.39	0.88	0.44	12.04***	0.48	7.61**	–		
Y1	7	6	0.0015	1	0.0001	9	0.0013	17.02	14.58	1.17	11.00***	11.80***	0.09		–	
A13.7	8	2	0.0005	4	0.0003	14	0.0020	1.42	5.67	0.25	0.16	11.74***	4.64*			+
E5–	9	7	0.0017	3	0.0003	12	0.0017	6.62	6.48	1.02	8.42**	10.99***	0.00		–	
I2.2	10	3	0.0007	21	0.0018	2	0.0003	0.41	0.15	2.63	2.67	10.45**	1.15		+	
E6–	11	11	0.0027	8	0.0007	3	0.0004	3.90	0.61	6.42	8.55**	0.58	10.42**	+		
X8+	12	12	0.0029	8	0.0007	8	0.0011	4.25	1.62	2.63	10.18**	0.92	4.59*	+		
A13.1	13	5	0.0012	5	0.0004	0	0	2.84	-	-	2.60	-	10.12**			–
M5	13	5	0.0012	7	0.0006	0	0	2.03	-	-	1.37	-	10.12**			–
A10+	14	0	0	10	0.0009	11	0.0015	-	1.78	-	-	1.74	9.94**	–		
T1.3	15	71	0.0174	299	0.0258	156	0.0218	0.67	0.85	0.80	9.70**	2.95	2.59		+	
A2.1+	16	2	0.0005	32	0.0028	18	0.0025	0.18	0.91	0.19	9.50**	0.10	7.31**	–		
Q4.3	17	14	0.0034	19	0.0016	6	0.0008	2.09	0.51	4.08	4.13*	2.27	9.32**	+		
S9	18	0	0	31	0.0027	10	0.0014	-	0.52	-	-	3.52	9.04**	–		
X2.6+	18	0	0	12	0.0010	10	0.0014	-	1.35	-	-	0.48	9.04**	–		
X4.2	19	4	0.0010	1	0.0001	7	0.0010	11.34	11.34	1.00	6.36*	8.42**	0.00		–	
O4.3	20	6	0.0015	2	0.0002	4	0.0006	8.51	3.24	2.63	8.34**	1.99	2.30	+		
S6+	21	59	0.0144	108	0.0093	61	0.0085	1.55	0.91	1.69	6.96**	0.31	8.20**	+		
Z8mf	22	241	0.0590	609	0.0526	332	0.0464	1.12	0.88	1.27	2.26	3.36	7.90**	+		
O1.1	23	0	0	1	0.0001	6	0.0008	-	9.72	-	-	6.78**	5.42*			+
S2.2m	24	4	0.0010	16	0.0014	2	0.0003	0.71	0.20	3.50	0.40	6.68**	2.26		+	

Table 6.14: Semantic (full) tag unigram analysis, Neuroticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
B3	1	1	0.0002	2	0.0002	12	0.0023	1.25	13.44	0.09	0.03	18.20****	9.78**			+
T2–	2	37	0.0080	33	0.0029	15	0.0029	2.81	1.02	2.76	18.16****	0.00	12.23***	+		
I3.1	3	30	0.0065	57	0.0049	54	0.0104	1.32	2.12	0.62	1.45	15.24****	4.53*			+
T1.1.2	4	69	0.0149	117	0.0101	38	0.0073	1.48	0.73	2.03	6.35*	3.06	12.91***	+		
Z3c	5	16	0.0035	10	0.0009	5	0.0010	4.01	1.12	3.58	12.20***	0.04	7.36**	+		
T1.3+	6	0	0	35	0.0030	9	0.0017	-	0.58	-	-	2.41	11.50***	–		
S3.2	6	0	0	22	0.0019	9	0.0017	-	0.92	-	-	0.05	11.50***	–		
Z1f	7	6	0.0013	51	0.0044	12	0.0023	0.29	0.53	0.56	10.95***	4.51*	1.42		+	
B5	8	0	0	11	0.0009	8	0.0015	-	1.63	-	-	1.06	10.22**	–		
P1	9	11	0.0024	54	0.0047	33	0.0064	0.51	1.37	0.37	4.77*	1.96	9.19**	–		
X2.2–	10	3	0.0006	2	0.0002	7	0.0014	3.76	7.84	0.48	2.14	8.40**	1.23			+
[	11	23	0.0050	25	0.0022	13	0.0025	2.30	1.16	1.98	8.03**	0.20	4.05*	+		
G3	12	5	0.0011	1	0.0001	1	0.0002	12.52	2.24	5.59	7.81**	0.32	3.38	+		
Y1	12	5	0.0011	1	0.0001	2	0.0004	12.52	4.48	2.80	7.81**	1.62	1.69	+		
T1.2	13	19	0.0041	19	0.0016	8	0.0015	2.50	0.94	2.66	7.74**	0.02	5.93*	+		
A1.9	14	0	0	3	0.0003	6	0.0012	-	4.48	-	-	4.86*	7.67**			+
]	15	22	0.0048	24	0.0021	12	0.0023	2.30	1.12	2.05	7.62**	0.10	4.21*	+		
K2	16	8	0.0017	13	0.0011	1	0.0002	1.54	0.17	8.95	0.89	4.74*	7.01**			–

Table 6.15: Semantic (full) tag unigram analysis, Psychoticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

sexual relationships (S3.2), and refer less to geographical names (Z2); Low Neurotics overuse degree minimisers (A13.7); both groups overuse science and technology terms (Y1), and negative fear/bravery/shock words (E5−), relative to the Mid group.

For Psychoticism, we find that again most significantly an overuse of references to medicines and medical treatment (B3) is an important feature, along with references to general work and employment (I3.1) is characteristic of Low Psychotics; High Psychotics overuse of time ending references (T2−). Less significant features for High Psychotics are an overuse of present and simultaneous general time references (T1.1.2), along with proper names (Z3c), and an underuse of longer time periods (T1.3+), and intimate or sexual relationship references (S3.2); Both High and Low groups underused female proper names (Z1f), relative to the Mid.

When the unigram data is reduced to just the initial letter indicating overall category, and initial number indicating first subdivision of this (Tables 6.16, 6.17, and 6.18), we find that this affects the dimensions differently, in the case of Extraversion, this increases the number of features achieving the conservative 15.13 critical level, but reduces the number of features demonstrating significance below this level. For Psychoticism, this reduces dramatically the number of features in both cases.

For Extraversion, cigarettes and drugs (F3) is still significantly overused by the Introverts and this is joined by general references to living creatures (L2), however other significant features are the High Extravert overuse of references to kin (S4), and the Mid group underuse of physical attribute (O4) references. Less significant is the Low Extravert overuse of arts and crafts (C1) and entertainment (K1) categories and the Mid group overuse of newclauses (<NC>) and underuse of linear order references (N4).

For Neuroticism, references to medicines and medical treatments (B3) is again most significant (overused by Low Neurotics), with High Neurotics overusing references to aircraft and flying (M5), with both groups overusing references which were unmatched (Z99) in comparison with the Mid group. Less significant is the High Neurotic overuse of references to relationships (S3), and underuse of geographical names (Z2); The Mid group overused happy or sad references (E4), those relating to business (I2), and underused references to fear, bravery or shock (E5), and references to science

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
F3	1	0	0	0	0	8	0.0030	-	-	-	-	26.76****	20.83****			+
O4	2	73	0.0102	55	0.0047	22	0.0082	2.15	1.73	1.24	18.50****	4.34*	0.81		–	
L2	3	3	0.0004	11	0.0009	11	0.0041	0.44	4.33	0.10	1.82	10.88***	16.00****			+
S4	4	43	0.0060	27	0.0023	5	0.0019	2.57	0.80	3.21	15.38****	0.22	8.24**	+		
C1	5	2	0.0003	1	0.0001	6	0.0022	3.23	25.96	0.12	0.99	14.75***	7.90**			+
K1	6	37	0.0052	48	0.0041	27	0.0101	1.25	2.43	0.51	1.00	12.28***	6.64**			+
<NC>	7	291	0.0406	595	0.0514	119	0.0445	0.79	0.87	0.91	11.05***	2.14	0.68		+	
N4	8	85	0.0119	82	0.0071	24	0.0090	1.68	1.27	1.32	11.04***	0.99	1.54	+		
K2	9	17	0.0024	15	0.0013	0	0	1.83	-	-	2.91	-	10.79**			–
F2	10	11	0.0015	48	0.0041	8	0.0030	0.37	0.72	0.51	10.63**	0.79	1.95		+	
Z1	11	90	0.0126	128	0.0111	16	0.0060	1.14	0.54	2.10	0.86	6.29*	8.84**			–
Z4	12	69	0.0096	168	0.0145	30	0.0112	0.66	0.77	0.86	8.63**	1.80	0.47		+	
Y1	13	7	0.0010	1	0.0001	0	0	11.32	-	-	8.40**	-	4.44*	+		
P1	14	31	0.0043	59	0.0051	4	0.0015	0.85	0.29	2.90	0.55	8.12**	5.22*			–
Z6	15	69	0.0096	165	0.0143	43	0.0161	0.68	1.13	0.60	7.83**	0.48	6.60*	–		
K5	16	11	0.0015	30	0.0026	13	0.0049	0.59	1.88	0.32	2.37	3.26	7.73**			+
B5	17	12	0.0017	11	0.0009	0	0	1.76	-	-	1.84	-	7.62**			–
Y2	17	11	0.0015	4	0.0003	2	0.0007	4.45	2.16	2.06	7.62**	0.72	1.03	+		
N6	18	40	0.0056	61	0.0053	27	0.0101	1.06	1.92	0.55	0.08	7.19**	5.36*			+

Table 6.16: Semantic (reduced) tag unigram analysis, Extraversion.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
B3	1	2	0.0005	2	0.0002	18	0.0025	2.84	14.58	0.19	1.04	23.59****	7.31**			+
M5	2	9	0.0022	10	0.0009	0	0	2.55	-	-	3.95*	-	18.21****	+		
Z99	3	44	0.0108	92	0.0079	104	0.0146	1.36	1.83	0.74	2.66	17.82****	2.89			+
S3	4	41	0.0100	78	0.0067	30	0.0042	1.49	0.62	2.39	4.09*	5.17*	13.37***	+		
E5	5	7	0.0017	3	0.0003	13	0.0018	6.62	7.02	0.94	8.42**	12.48***	0.02		−	
E4	6	10	0.0024	74	0.0064	20	0.0028	0.38	0.44	0.88	10.27**	12.37***	0.12		+	
I2	7	4	0.0010	27	0.0023	3	0.0004	0.42	0.18	2.33	3.22	12.24***	1.25		+	
Z2	8	12	0.0029	88	0.0076	48	0.0067	0.39	0.88	0.44	12.04***	0.48	7.61**	−		
Y1	9	6	0.0015	1	0.0001	9	0.0013	17.02	14.58	1.17	11.00***	11.80***	0.09		−	
E6	10	11	0.0027	11	0.0009	3	0.0004	2.84	0.44	6.42	5.72*	1.81	10.42**	+		
X8	11	12	0.0029	8	0.0007	8	0.0011	4.25	1.62	2.63	10.18**	0.92	4.59*	+		
A10	12	0	0	23	0.0020	11	0.0015	-	0.77	-	-	0.50	9.94**	−		
S9	13	0	0	31	0.0027	10	0.0014	-	0.52	-	-	3.52	9.04**	−		
A2	14	7	0.0017	52	0.0045	36	0.0050	0.38	1.12	0.34	7.26**	0.28	8.49**	−		
S6	15	60	0.0147	108	0.0093	62	0.0087	1.58	0.93	1.69	7.58**	0.21	8.36**	+		
N6	16	12	0.0029	61	0.0053	19	0.0027	0.56	0.50	1.11	3.87*	7.54**	0.07		+	
M4	17	1	0.0002	7	0.0006	14	0.0020	0.41	3.24	0.13	0.89	6.96**	7.33**			+
X4	18	6	0.0015	7	0.0006	14	0.0020	2.43	3.24	0.75	2.42	6.96**	0.36			+
Q4	19	16	0.0039	35	0.0030	10	0.0014	1.30	0.46	2.80	0.72	5.24*	6.77**			−

Table 6.17: Semantic (reduced) tag unigram analysis, Neuroticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

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
I3	1	30	0.0065	57	0.0049	58	0.0112	1.32	2.28	0.58	1.45	19.02****	6.21*			+
B3	2	1	0.0002	2	0.0002	12	0.0023	1.25	13.44	0.09	0.03	18.20****	9.78**			+
Z3	3	19	0.0041	13	0.0011	8	0.0015	3.66	1.38	2.66	13.16***	0.50	5.93*	+		
P1	4	13	0.0028	59	0.0051	38	0.0073	0.55	1.44	0.38	4.25*	3.01	10.16**	−		
X4	5	12	0.0026	7	0.0006	5	0.0010	4.29	1.60	2.68	9.79**	0.62	3.81	+		
[	6	23	0.0050	25	0.0022	13	0.0025	2.30	1.16	1.98	8.03**	0.20	4.05*	+		
G3	7	5	0.0011	1	0.0001	1	0.0002	12.52	2.24	5.59	7.81**	0.32	3.38	+		
Y1	7	5	0.0011	1	0.0001	2	0.0004	12.52	4.48	2.80	7.81**	1.62	1.69	+		
]	8	22	0.0048	24	0.0021	12	0.0023	2.30	1.12	2.05	7.62**	0.10	4.21*	+		
S2	9	24	0.0052	40	0.0035	10	0.0019	1.50	0.56	2.68	2.39	3.00	7.61**	+		

Table 6.18: Semantic (reduced) tag unigram analysis, Psychoticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05, **p < .01, ***p < .001, ****p < .0001, df = 1$.

and technology (Y1).

Psychoticism shows just Low Psychotics overuse of references to work and employment (I3), and medicines and medical treatment (B3); High Psychotics, solely overuse references to other proper names (Z3).

The most reduced unigram analysis just reduces the tags to the initial letter indicating broad semantic grouping (Tables 6.19, 6.20, and 6.21). This has little effect upon Extraversion, with a slight reduction of features showing greater levels of significance, whereas for the Neuroticism results this is more serious, as the features do not reach even the lower critical value of 10.83; for Psychoticism, only a few features reach this critical value.

Thee Mid Extraversion group show underuse of references to substances, materials, objects and equipment (O) (also slightly overused by High Extraverts), with High Extraverts overusing science and technology references (Y). Features reaching a lower critical value are the overuse of arts and crafts (C), and food and farming (F) references and the underuse of movement, location travel and transport (M) references by the Low Extraverts; the Mid group also overused newclauses (<NC>). Additionally features with a critical value of greater than 6.63, are the Mid group's overuse of time references (T), and the Low Extravert overuse of references to life and living things (L), and entertainment, sports and games (K), and the underuse of education references (P).

Although the Neurotics do not show highly significant patterns, the most significant features (with a critical value greater than 6.63) are the High Neurotic overuse of social action, states and processes (S) references, and the Low Neurotics overuse of references to substances, materials, objects and equipment; the Mid group underuse science and technology (Y), and money and commerce in industry (I) references, and underuse references to the body and the individual (B).

For Psychoticism, we find that greatest significance is found for the High Psychotics' overuse of science and technology (Y) references, and the Low Psychotic overuse of references to the body and the individual (B) (critical value $\geq$10.83; findings reaching the critical value of 6.63 are the High Psychotic underuse of education references (P), and their use of parentheses ([]).

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
O	1	93	0.0130	74	0.0064	25	0.0093	2.03	1.46	1.39	20.84****	2.52	2.27	+		
Y	2	18	0.0025	5	0.0004	2	0.0007	5.82	1.73	3.36	15.36****	0.39	3.63	+		
C	3	2	0.0003	1	0.0001	6	0.0022	3.23	25.96	0.12	0.99	14.75***	7.90**			+
F	4	39	0.0054	105	0.0091	36	0.0135	0.60	1.48	0.40	7.93**	3.90*	14.65***	−		
M	5	219	0.0306	424	0.0366	63	0.0235	0.84	0.64	1.30	4.78*	11.92***	3.50		+	
<NC>	6	291	0.0406	595	0.0514	119	0.0445	0.79	0.87	0.91	11.05***	2.14	0.68		+	
T	7	511	0.0713	976	0.0843	196	0.0732	0.85	0.87	0.97	9.47**	3.32	0.10		+	
L	8	11	0.0015	23	0.0020	14	0.0052	0.77	2.63	0.29	0.51	7.32**	9.14**			+
P	9	31	0.0043	59	0.0051	4	0.0015	0.85	0.29	2.90	0.55	8.12**	5.22*			−
K	10	74	0.0103	103	0.0089	41	0.0153	1.16	1.72	0.67	0.96	7.96**	3.92*			+

Table 6.19: Semantic (most reduced) tag unigram analysis, Extraversion.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
Y	1	9	0.0022	5	0.0004	12	0.0017	5.10	3.89	1.31	8.97**	7.33**	0.38		−	
S	2	175	0.0429	406	0.0351	227	0.0318	1.22	0.91	1.35	4.80*	1.44	8.74**	+		
A	3	537	0.1315	1539	0.1329	1066	0.1491	0.99	1.12	0.88	0.04	8.29**	5.72*			+
I	4	31	0.0076	129	0.0111	51	0.0071	0.68	0.64	1.06	3.96*	7.69**	0.07		+	
B	5	32	0.0078	76	0.0066	73	0.0102	1.19	1.56	0.77	0.69	7.20**	1.61			+

Table 6.20: Semantic (most reduced) tag unigram analysis, Neuroticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
Y	1	12	0.0026	5	0.0004	3	0.0006	6.01	1.34	4.47	12.86***	0.16	6.84**	+		
B	2	20	0.0043	76	0.0066	54	0.0104	0.66	1.59	0.41	2.97	6.58*	12.65***			+
P	3	13	0.0028	59	0.0051	38	0.0073	0.55	1.44	0.38	4.25*	3.01	10.16**	−		
[	4	23	0.0050	25	0.0022	13	0.0025	2.30	1.16	1.98	8.03**	0.20	4.05*	+		
]	5	22	0.0048	24	0.0021	12	0.0023	2.30	1.12	2.05	7.62**	0.10	4.21*	+		

Table 6.21: Semantic (most reduced) tag unigram analysis, Psychoticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

6.3.3 Combined N-gram Results

The results showing the combined 1–5 n-gram analysis are shown in the following tables. This analysis uses the full USAS semantic tags, and the findings for Extraversion, Neuroticism, and Psychoticism are presented in tables 6.22, 6.23, and 6.24. Due to the greater levels of significance for these features, here we only display features with critical values greater than 15.13.

Here we include the n-gram analysis of the semantic tags for completeness, since semantics in its conveyance of meaning, is not generally concerned with the structure of a text, in the way that the ordering of syntactic information is dependent on order in English (however, this is not the case in all languages, e.g. Latin). However, here we look to the semantic patterning of the words in our corpus for evidence of further language variation across the different personality groups.

To interpret the contextual information features showing critical values greater than 15.13, we start by looking for the co-occurrence of semantic features with larger linguistic markers such as the newclause (<NC>) boundary marker, and punctuation (<P>). We then chart patterns of co-occurrence across semantic tags within the clause level.

For Extraversion, we find that Low Extraverts underuse grammatical words (Z5), such as prepositions, adverbs or conjunctions following punctuation (<P>) or a newclause marker (<NC>); High Extraverts show an underuse of geographical names (Z2), and an overuse of terms depicting greater quantities (N4++) before punctuation (<P>); both the High and Low groups showed an overuse of terms relating to food (F1) before punctuation (<P>).

Turning to pronoun references for Extraversion, and we find that whilst Low Extraverts underuse pronouns (Z8) preceding general grammatical words (Z5) ([*Z8 Z5*]), High Extraverts underuse them prior to terms of positive modality (A7+) ([*Z8 A7+*]); in the case of personal pronouns (Z8mfn), we find that Low Extraverts overuse these in combination with grammatical words (Z5) ([*Z5 Z8mfn*], [*Z5 Z8mfn Z5*]), and also with references to entertainment (K1) ([*Z8mf K1*]), but show underuse in relation to general future time references ([*Z8mf T1.1.3*]); High Extraverts overuse personal pronouns when preceding evaluative terms of positive authenticity (A4.5+) ([*Z8mf A4.5+*]).

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
Z5 A1.1.1	1	42	0.0059	0	0	0	0	-	-	-	80.80****	-	26.66****	+		
<P> Z5	2	116	0.0162	163	0.0141	0	0	1.15	-	-	1.32	-	73.64****			−
A9+ Z5	3	71	0.0099	56	0.0048	0	0	2.05	-	-	16.24****	-	45.07****	+		
Z8 Z5	4	62	0.0087	81	0.0070	0	0	1.24	-	-	1.58	-	39.36****			−
I3.1 Z5	5	17	0.0024	0	0	0	0	-	-	-	32.71****	-	10.79**	+		
<NC> Z5	6	48	0.0067	76	0.0066	0	0	1.02	-	-	0.01	-	30.47****			−
N1 Z5	7	0	0	0	0	9	0.0034	-	-	-	-	30.11****	23.44****			+
Z5 Z8mfn	7	0	0	0	0	9	0.0034	-	-	-	-	30.11****	23.44****			+
Z5 T3−	8	0	0	0	0	8	0.0030	-	-	-	-	26.77****	20.83****			+
F3	9	0	0	0	0	8	0.0030	-	-	-	-	26.76****	20.83****			+
Q1.2 Z5	10	13	0.0018	0	0	0	0	-	-	-	25.01****	-	8.25**	+		
Z5 A1.1.1 Z5	10	13	0.0018	0	0	0	0	-	-	-	25.01****	-	8.25**	+		
Z5 A9−	11	0	0	0	0	7	0.0026	-	-	-	-	23.42****	18.23****			+
Z8mf K1	11	0	0	0	0	7	0.0026	-	-	-	-	23.42****	18.23****			+
T1.1.3 A9+	12	12	0.0017	0	0	0	0	-	-	-	23.09****	-	7.62**	+		
Z2 <P>	13	0	0	19	0.0016	8	0.0030	-	1.82	-	-	1.85	20.83****	−		
F1 <P>	14	8	0.0011	0	0	6	0.0022	-	-	0.50	15.39****	20.08****	1.58		−	
<NC> N4	15	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.35*	+		
T1 Z5	15	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.35*	+		
Z1m Z5	15	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.35*	+		
Z5 Z1m	15	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.35*	+		
Z5 A3+	16	30	0.0042	53	0.0046	0	0	0.92	-	-	0.15	-	19.04****			−
Z8mf T1.1.3	17	29	0.0040	37	0.0032	0	0	1.27	-	-	0.90	-	18.41****			−
A9+ A9+	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
N4 Z5	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
N5++ <P>	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
X8+ Z5	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
Z5 T2−	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
Z8mf A5.4+	18	9	0.0013	0	0	0	0	-	-	-	17.32****	-	5.71*	+		
A3+ S6+	19	0	0	0	0	5	0.0019	-	-	-	-	16.73****	13.02***			+
N4 T1	19	0	0	0	0	5	0.0019	-	-	-	-	16.73****	13.02***			+
Z5 Z8mfn Z5	19	0	0	0	0	5	0.0019	-	-	-	-	16.73****	13.02***			+
Z6 A9+	19	0	0	0	0	5	0.0019	-	-	-	-	16.73****	13.02***			+
M6 Z5	20	26	0.0036	37	0.0032	0	0	1.14	-	-	0.25	-	16.50****			−
Z5 X3.4	21	25	0.0035	29	0.0025	0	0	1.39	-	-	1.46	-	15.87****			−
Z5 N1 Z5	22	0	0	6	0.0005	6	0.0022	-	4.33	-	-	5.94*	15.63****			+
Z8 A7+	22	0	0	19	0.0016	6	0.0022	-	1.37	-	-	0.42	15.63****	−		
Z5 F2	22	0	0	25	0.0022	6	0.0022	-	1.04	-	-	0.01	15.63****	−		
X5.2+ Z5	23	8	0.0011	0	0	0	0	-	-	-	15.39****	-	5.08*	+		
N4 N4	24	24	0.0034	10	0.0009	0	0	3.88	-	-	14.61***	-	15.24****	+		
Z1f Z5	24	24	0.0034	23	0.0020	0	0	1.69	-	-	3.19	-	15.24****			−

Table 6.22: Semantic tag n-gram analysis, Extraversion.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
Z5 Z8mf	1	71	0.0174	206	0.0178	0	0	0.98	-	-	0.03	-	143.69****			−
Z8mf Z5	2	0	0	193	0.0167	111	0.0155	-	0.93	-	-	0.35	100.30****	−		
Z5 Z8	2	0	0	180	0.0155	111	0.0155	-	1.00	-	-	0.00	100.30****	−		
Z5 A1.1.1	3	0	0	0	0	45	0.0063	-	-	-	-	86.69****	40.66****			+
T1.3 Z5	4	26	0.0064	0	0	0	0	-	-	-	69.92****	-	52.62****	+		
A1.1.1 <P>	5	15	0.0037	0	0	23	0.0032	-	-	1.14	40.34****	44.31****	0.16		−	
Z8 Z5	6	0	0	81	0.0070	47	0.0066	-	0.94	-	-	0.11	42.47****	−		
<P> <P> <P>	7	22	0.0054	4	0.0003	2	0.0003	15.60	0.81	19.26	39.26****	0.06	32.57****	+		
Z5 Z99	8	0	0	39	0.0034	43	0.0060	-	1.79	-	-	6.85**	38.86****			+
<P> <P> <P> <P>	9	17	0.0042	1	0.0001	1	0.0001	48.23	1.62	29.77	38.60****	0.12	27.59****	+		
<P> <P> <P> <P> <P>	10	14	0.0034	0	0	0	0	-	-	-	37.66****	-	28.34****	+		
A9+ Z5	11	0	0	56	0.0048	40	0.0056	-	1.16	-	-	0.49	36.14****	−		
Z5 A9+	11	0	0	58	0.0050	40	0.0056	-	1.12	-	-	0.29	36.14****	−		
Z5 N1	12	0	0	0	0	17	0.0024	-	-	-	-	32.75****	15.36****			+
Z99 Z5	13	0	0	36	0.0031	34	0.0048	-	1.53	-	-	3.13	30.72****	−		
<NC> A5.1+	14	11	0.0027	0	0	0	0	-	-	-	29.58****	-	22.26****	+		
T1.1.1 <P>	15	0	0	0	0	15	0.0021	-	-	-	-	28.90****	13.55***			+
Z8mf M1	16	0	0	0	0	14	0.0020	-	-	-	-	26.97****	12.65***			+
A9+ Z8	17	10	0.0024	0	0	0	0	-	-	-	26.89****	-	20.24****	+		
T2− Z5	17	10	0.0024	0	0	0	0	-	-	-	26.89****	-	20.24****	+		
Z5 Q4.3	18	13	0.0032	12	0.0010	0	0	3.07	-	-	7.59**	-	26.31****	+		
Z5 K1	18	13	0.0032	19	0.0016	0	0	1.94	-	-	3.20	-	26.31****			−
Z5 A1.1.1 Z5	19	0	0	0	0	13	0.0018	-	-	-	-	25.04****	11.75***			+
Z8 A1.1.1	20	9	0.0022	0	0	0	0	-	-	-	24.20****	-	18.21****	+		
X8+ Z5	20	9	0.0022	0	0	8	0.0011	-	-	1.97	24.20****	15.41****	1.94		−	
B3	21	2	0.0005	2	0.0002	18	0.0025	2.84	14.58	0.19	1.04	23.59****	7.31**			+
S5−	22	12	0.0029	2	0.0002	1	0.0001	17.02	0.81	21.01	21.99****	0.03	18.14****	+		
A1.1.1 Z8	23	8	0.0020	0	0	0	0	-	-	-	21.51****	-	16.19****	+		
Z5 N5	23	8	0.0020	0	0	0	0	-	-	-	21.51****	-	16.19****	+		
Z5 Z8mfn	23	8	0.0020	0	0	0	0	-	-	-	21.51****	-	16.19****	+		
K1 <P>	24	0	0	0	0	11	0.0015	-	-	-	-	21.19****	9.94**			+
Z5 F1	25	0	0	30	0.0026	22	0.0031	-	1.19	-	-	0.37	19.88****	−		
N5+ Z5	25	0	0	34	0.0029	22	0.0031	-	1.05	-	-	0.03	19.88****	−		
A9+ A9+	26	7	0.0017	0	0	8	0.0011	-	-	1.53	18.82****	15.41****	0.67		−	
Z6 A9+	26	7	0.0017	0	0	5	0.0007	-	-	2.45	18.82****	9.63**	2.38		−	
Z99	27	44	0.0108	92	0.0079	104	0.0146	1.36	1.83	0.74	2.66	17.82****	2.89			+
E4.1+	28	8	0.0020	57	0.0049	10	0.0014	0.40	0.28	1.40	7.45**	17.60****	0.50		+	
F1 <P>	29	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.13**			+
Z5 M2	29	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.13**			+
T1.1.3 A3+	30	0	0	35	0.0030	19	0.0027	-	0.88	-	-	0.21	17.17****	−		
A5.1+ <P>	30	0	0	29	0.0025	19	0.0027	-	1.06	-	-	0.04	17.17****	−		
<P> <NC> A5.1+	31	11	0.0027	3	0.0003	6	0.0008	10.40	3.24	3.21	16.85****	2.99	5.61*	+		
T1 <P>	32	6	0.0015	0	0	0	0	-	-	-	16.13****	-	12.14***	+		
T1.1.3 S6+	32	6	0.0015	0	0	0	0	-	-	-	16.13****	-	12.14***	+		
A9+ N5+	32	6	0.0015	0	0	5	0.0007	-	-	2.10	16.13****	9.63**	1.50		−	
M1 M6	33	4	0.0010	0	0	8	0.0011	-	-	0.88	10.76**	15.41****	0.05		−	
I3.1 <P>	33	0	0	0	0	8	0.0011	-	-	-	-	15.41****	7.23**			+

Table 6.23: Semantic tag n-gram analysis, Neuroticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

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
Z8mf Z5	1	0	0	193	0.0167	103	0.0199	-	1.20	-	-	2.10	131.59****	–		
Z5 Z8mf	2	68	0.0147	206	0.0178	0	0	0.83	-	-	1.91	-	102.09****			–
Z8 Z8mf	3	20	0.0043	0	0	0	0	-	-	-	50.17****	-	30.03****	+		
I3.1 Z5	4	0	0	0	0	21	0.0041	-	-	-	-	49.37****	26.83****			+
Z8mf M1	5	0	0	0	0	18	0.0035	-	-	-	-	42.32****	23.00****			+
Z5 Q2.1	6	10	0.0022	0	0	16	0.0031	-	-	0.70	25.08****	37.62****	0.81		–	
Z5 N1	6	0	0	0	0	16	0.0031	-	-	-	-	37.62****	20.44****			+
Z8mf A3+	7	22	0.0048	35	0.0030	0	0	1.57	-	-	2.67	-	33.03****			–
<NC> T1.3	8	13	0.0028	0	0	0	0	-	-	-	32.61****	-	19.52****	+		
X7+ Z5	9	0	0	34	0.0029	24	0.0046	-	1.58	-	-	2.85	30.66****	–		
I3.1 <P>	10	0	0	0	0	13	0.0025	-	-	-	-	30.56****	16.61****			+
A9+ Z5	11	20	0.0043	56	0.0048	0	0	0.89	-	-	0.19	-	30.03****			–
Z5 Z5 Z5	12	12	0.0026	11	0.0009	29	0.0056	2.73	5.91	0.46	5.65*	29.25****	5.49*			+
Z8m Z5	13	11	0.0024	0	0	0	0	-	-	-	27.59****	-	16.51****	+		
A9+ Z8	14	0	0	0	0	10	0.0019	-	-	-	-	23.51****	12.78***			+
Z5 A2.1+	14	0	0	0	0	10	0.0019	-	-	-	-	23.51****	12.78***			+
M1 M6	15	9	0.0019	0	0	5	0.0010	-	-	2.01	22.58****	11.76***	1.65		–	
<NC> T1.3 <P>	15	9	0.0019	0	0	2	0.0004	-	-	5.03	22.58****	4.70*	5.64*	+		
<P> <NC> T1.3 <P>	15	9	0.0019	0	0	2	0.0004	-	-	5.03	22.58****	4.70*	5.64*	+		
Z5 Z8mf Z5	16	4	0.0009	63	0.0054	23	0.0044	0.16	0.82	0.19	22.05****	0.70	12.74***	–		
X8+ Z5	17	0	0	0	0	9	0.0017	-	-	-	-	21.16****	11.50***			+
Z5 Z5 Z5 Z5	17	1	0.0002	0	0	9	0.0017	-	-	0.12	2.51	21.16****	6.50*			+
Z99 <P>	18	14	0.0030	22	0.0019	0	0	1.59	-	-	1.78	-	21.02****			–
F2 <P>	19	8	0.0017	0	0	0	0	-	-	-	20.07****	-	12.01***	+		
H4 Z5	19	8	0.0017	0	0	0	0	-	-	-	20.07****	-	12.01***	+		
K1 <P>	20	0	0	0	0	8	0.0015	-	-	-	-	18.81****	10.22**			+
Z5 Z8f	20	0	0	0	0	8	0.0015	-	-	-	-	18.81****	10.22**			+
Z8 P1	20	0	0	0	0	8	0.0015	-	-	-	-	18.81****	10.22**			+
Z8 <P>	20	0	0	0	0	8	0.0015	-	-	-	-	18.81****	10.22**			+
B3	21	1	0.0002	2	0.0002	12	0.0023	1.25	13.44	0.09	0.03	18.20****	9.78**			+
T2–	22	37	0.0080	33	0.0029	15	0.0029	2.81	1.02	2.76	18.16****	0.00	12.23***	+		
S3.1 Z5	23	12	0.0026	21	0.0018	0	0	1.43	-	-	0.95	-	18.02****			–
Z1m <P>	24	7	0.0015	0	0	0	0	-	-	-	17.56****	-	10.51**	+		
K1 Z5	25	0	0	20	0.0017	13	0.0025	-	1.46	-	-	1.08	16.61****	–		
A7+ Z6	26	11	0.0024	22	0.0019	0	0	1.25	-	-	0.36	-	16.51****			–
<P> B1	27	0	0	0	0	7	0.0014	-	-	-	-	16.46****	8.94**			+
Z5 X8+	27	0	0	0	0	7	0.0014	-	-	-	-	16.46****	8.94**			+
Z5 X3.4	28	0	0	29	0.0025	12	0.0023	-	0.93	-	-	0.05	15.33****	–		
I3.1	29	30	0.0065	57	0.0049	54	0.0104	1.32	2.12	0.62	1.45	15.24****	4.53*			+

Table 6.24: Semantic tag n-gram analysis, Psychoticism.

Note. See Tables B.2– B.4 for a description of the tags. $*p < .05$, $**p < .01$, $***p < .001$, $****p < .0001$, $df = 1$.

Here we also note the behaviour of getting and giving possession terms (A9), since there is significant difference in behaviour between the High and Low Extravert groups: the Low Extraverts show the only use of negation (N6) in relation to allocating or acquiring possession, (A9+) overusing [*N6 A9+*], but otherwise overuse the semantic tag indicating relinquishing or receiving possession (A9−) along with a grammatic term (Z5) ([*Z5 A9−*]); High Extraverts on the other hand only overuse the form relating to allocating or acquiring possession and general future time references ([*A9+ Z5*], [*T1.1.3 A9+*]), and also show duplication as in ([*A9+ A9+*]).

Finally we observe the behaviour of time references across the groups: High Extraverts overuse terms relating to lack of age (T3−) with grammatical words ([*N5 T3−*]), and terms of linear order (N4) with general time references (T1), and underuse personal pronouns (Z8mf) with general future time references (T1.1.3) ([*Z8mf T1.1.3*]); High Extraverts overuse general future time references (T1.1.3) in combination with terms indicating the allocation or acquiring of possession (A9+) ([*T1.1.3 A9+*]).

For Neuroticism we find that in addition to their multiple use of punctuation features, we find that High Neurotics overuse positive evaluative terms at the start of a clause ([*<NC> A5.1+*], [*<P> <NC> A5.1+*]), but underuse this feature before punctuation ([*A5.1+ <P>*]), and overuse general time references (T1) before punctuation ([*T1 <P>*]); A neat contrast in time references is found for Low Neurotics who overuse references to general past time (T1.1.1) ([*T1.1.3 <P>*]). They additionally overuse references to entertainment (K1), and general work and employment (I3.1) preceeding punctuation ([*K1 <P>*], [*I3.1 <P>*]); Both High and Low groups overuse general actions relating to making (A1.1.1) prior to punctuation ([*A1.1.1 <P>*]).

In terms of pronoun use (Z8), High Neurotics show underuse in combination with grammatical features (Z5) ([*Z5 Z8*], [*Z8 Z5*]), but this alternated with an overuse of general making actions (A1.1.1) ([*Z8 A1.1.1*], [*A1.1.1 Z8*]). They also overuse pronouns in relation to allocating or acquiring (A9+) ([*A9+ Z8*]). For personal pronouns, we find alternation in relation to use with grammatical words: the Low group underuse personal pronouns following the grammatical words ([*Z5 Z8mf*]), whereas the High group underuse the reversed form for male or female pronouns ([*Z8mf Z5*]), and overuse with the neuter pronoun ([*Z5 Z8mfn*]); Low Neurotics also overuse terms of

coming or going movement (M1), in relation to pronouns ([*Z8mf M1*]).

In addition to the different time references—Low Neurotics preferring references to past rather than general time—prior to punctuation noted above, High Neurotics show overuse of time period references followed by grammatical words ([*T1.3 Z5*]), and also overuse general future references when followed by terms indicating high levels of obligation or necessity (S6+) ([*T1.1.3 S6+*]), and an underuse when followed by expressions of being or existing ([*T1.1.3 A3+*]); Finally, High Neurotics also overused time references relating to ending or ceasing (T2−), in combination with grammatical words ([*T2− Z5*]).

The pattern for allocating or acquiring references (A9+) appears to not be clearly contrastive for Neuroticism: whilst High Neurotics overuse this in conjunction with pronouns ([*A9+ Z8*]), they underuse this term with grammatical words ([*A9+ Z5*], [*Z5 A9+*]); we also find that both High and Low groups overuse this term relative to the Mid group, notably when duplicated ([*A9+ A9+*]), when negated ([*Z6 A9+*]) and when used in conjunction with references to large quantities ([*A9+ N+*]).

For Psychoticism, we find that the High group overuses time references relating to age or maturity at the start of a clause, which is usually followed by punctuation ([*<NC> T1.3*], [*<NC> T1.3 <P>*], [*<P> <NC> T1.3 <P>*]). Looking at punctuation we find that for High Psychotics precede this with an overuse of terms relating to drinks (F2), and male names (Z1m) ([*F2 <P>*], [*Z1m <P>*]), whereas Low Psychotics underuse unclassified words (Z99), and overuse entertainment references (K1) and pronouns (Z8) when preceding punctuation ([*Z99 <P>*], [*K1 <P>*], [*Z8 <P>*]), and overuse anatomy or physiology references (B1) when following it (([*B1 <P>*]).

In addition to the Low Psychotic overuse of pronouns preceding punctuation, we find that use of personal pronouns distinguish the groups further: In relation to grammatical words (Z5), the Low group underuse grammatical words followed by pronoun ([*Z5 Z8mf*]), whereas the High group underuse the reverse of this pattern most significantly ([*Z8mf Z5*]), but also underuse ([*Z5 Z8mf Z5*]); in the case of gender differentiated pronouns, High Psychotics overuse male references ([*Z8m Z5*]), and Low Psychotics overuse female references ([*Z5 Z8f*]), which neatly map onto the underuses of the respective groups noted above in relation to general personal pronoun use; Ad-

ditionally Low Psychotics underuse personal pronouns when followed by references to being or existing ([*Z8mf A3+*]), and show overuse when followed by coming or going movement, or education references ([*Z8mf M1*], [*Z8mf P1*]).

With reference to the other features we find that in addition to High Psychotics showing an overuse of starting their clauses with references to time periods, they also overuse the unigram reference to time ending or ceasing (T2−). Furthermore, references to allocating or acquiring (A9+) are shown by the Low group to be underused when followed by a grammatical word ([*A9+ Z5*]), and overused by this group when followed by a pronoun ([*A9+ Z8*]). Negation is overused by the Low Psychotic group when following positive modal references ([*A7+ Z6*]).

6.4 Discussion

6.4.1 Summary of analysis techniques

The analyses which we have explored in this chapter have built upon the corpus stratification and the simple annotation techniques proposed in the previous chapter. Here we have adopted higher-level syntactic and semantic annotation methods which have allowed us to address several of the issues raised in the previous chapter, namely the relatively modest size of our stratified subcorpora groups and the possibility of topic specificity. By abstracting away from individual word forms or word stems (lemmas) we have been able to examine the relative differences in usage of broad linguistic word categories and concepts. Additionally, in our analysis we have applied simple scripting tools to enable the further reclassification of syntactic and semantic tags to allow analysis at still more general conceptual levels.

Furthermore by applying n-gram analyses to our data—prior to the comparison of our corpus groups—we have been able to show the way in which significantly collocating grammatical constructions are characteristic of our personality groups. This is a significant advancement upon simply analysing the proportions of grammatical category use which has previously been applied to the study of personality language. Furthermore we have applied these n-gram comparison techniques to the semantic concept analysis. The semantic analysis showed significant differences in semantic

patterning. However, since this is a novel technique we suggest that it requires further investigation. In summary, however, we view semantic tagging analysis as a significant advancement upon content analysis based approaches.

6.4.2 Summary of findings and hypotheses

For our results overall, we found that the unigram part-of-speech analysis identified relatively few features which showed a moderate significance, but that the n-gram analysis revealed a greater number of significant patterns. In each case analyses using the general grammatical categories showed lower levels of significance. The semantic tag analysis identified significant features characteristic of the personality groups, for both the n-gram analysis, and to a lesser extent, the unigram analysis. We now examine these results with regard to our hypotheses (summarised in Table 6.25).

For Extraversion, although our unigram analysis was not highly significant, it showed a greater use of past tense and past participle verbs; we also demonstrated differences in the patterning of adverbs for High Extraverts, and nouns and prepositions for Low Extraverts (this partly confirms our Grammatical hypothesis). From the semantic analysis, we find that High Extraverts refer more to friends (partially our Theory hypothesis); we also note that Low Extraverts make more references to drugs, and arts, crafts and entertainment.

Although we did not have any specific grammatical hypotheses for Neuroticism and Psychoticism, we found the following: High Neurotics are characterised by multiple punctuation patterns and an avoidance of proper noun references, Low Neurotics are differentiated by adverb and verb patterns; Psychoticism, we noted, was less distinguishable by broad grammatical category patterning, however the high and low groups do show different collocations of verbs, adverbs and nouns.

Additionally for the semantic analysis, we found: High Neurotics make more references to negative affiliation towards groups and intimate relationships, Low Neurotics refer more to medicine; High Psychotics make more references to time and proper names, and refer less to intimate relationships (partially confirming our Theory hypothesis); Low Psychotics refer more to medicine and employment. Although these findings do not necessarily address our hypotheses, they provide additional insight

into the characteristic language behaviour of the personality groups.

The personality projection features covered here, along with those from the previous chapters, are presented and summarised together in the conclusion. Please see Section 8.2.2, and Table 8.1 for this comprehensive presentation of projection—and also perception—findings.

6.5 Conclusion

In this chapter we have built upon the previous stratified corpus analysis which had used words or minor annotations, and applied these techniques to higher level linguistic analysis of the personality corpus using syntactic and semantic information. In addition to the unigram anlysis which examined the relative usage of different syntactic or semantic categories in the corpus, we also applied n-gram analysis to our data. Although this is an established technique for the analysis of syntactic information, and has previously been applied to corpus comparison studies, we extended this analysis to the semantic data. In addition to the semantic preferences revealed by the unigram analysis, we were also able to demonstrate instances where contextual information can provide further insight for semantic investigation.

We have therefore demonstrated that personality is projected through language. In the next chapter we turn to address the second main hypothesis of the thesis, namely whether personality can be perceived through language.

Extraversion											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	$\oplus^{n,z}$	$\odot^{p,o}_{n,z}$	o	$\oplus^{z}$							
Real.	o	o	o								
Fluency	o	o	o								
Gramm	$\oplus^{n}$	$\bullet^{p,o}$	$\oplus^{p,o}$	o	$\odot^{p,o}_{q}$	$\oplus^{n}$	$\bullet^{p,o}$	o			+Coord. conj.^{p}; [Adverb, Noun and Conjunction patterns]^{q,n}
Conv.	o	o	o	o	o	o	$\oplus^{z}$	o			
Lexis	$\oplus^{z}$	o	$\oplus^{z}$	$\oplus^{z}$	o	o	o	o	o	o	[+Linear order; +Music; +Education; +Science; +Clothes; +Information technology; −Cigarettes and drugs; −Living creatures; −Arts and crafts; −Entertainment; −Sports; −Frequency]^{z}
Percept.	o	o									

Neuroticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	$\oplus^{z}$	$\ominus^{z}$	$\odot^{p,o}_{,q,n}$	$\odot^{p,q}_{n}$	$\odot^{o,n}$						[−Coord. conj.; +Ellipsis]^{p}; [Adverb, Noun and Verb patterns]^{q,n}
Conv.	o	$\oplus^{z}$									
Lexis	$\odot^{o,n}$	$\oplus^{z}$	$\oplus^{z}$	o	o						[+Aircraft; +Relationships; +Trying; +Obligation; +Media; −Medical; −Places; −Revealing; −Religion; −Water activity;−Mental concerns]^{z}
Percept.	o	o									

Psychoticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	$\odot^{o,n}_{z}$	$\odot^{o,n}_{z}$	o	o	o	o	$\odot^{p,o}_{n}$				[+Ellipsis; −Foreign word; +Interjection; +Parentheses]^{p}; [Adverb, Noun and Verb patterns]^{q,n}
Conv.	o										
Lexis	$\odot^{o,n}$	o	$\bullet^{p,o}$	o	o	o	o	o	o	o	[+Mental concerns; +Warfare; +Science; −Work; −Medical; −Education]^{z}
Percept.	o	o									

Table 6.25: Review of hypotheses

Note. o indicates an hypothesis; $\oplus$ confirmation of hypothesis; $\ominus$ inverse of hypothesis; $\odot$ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{p} PENN tag unigram analysis, ^{o} reduced tagset unigram analysis, ^{q} PENN tag n-gram analysis, ^{n} n-gram analysis (reduced tags),^{z} semantic analysis (reduced). Please refer to Section 2.6 for a full description of the hypotheses.

Chapter 7

Rating E-mail Personality

In previous chapters we have outlined the collection of the corpus which forms the focus of investigation for this thesis, and also different linguistic analyses which have highlighted language features characteristic of personality. Indeed, some of these analyses have proved more sensitive to the identification of the personality language of some traits than others.

In this chapter we turn to the role of human perception in the identification of personality: whereas previously we have focussed on identifying specific features which are characteristic of personality, here we investigate the ability of human raters to identify personality on the basis of e-mail texts. Furthermore, we also investigate how subjective perceptions of the judges relate to the personality of the author.

Therefore, to test the second hypothesis of the thesis, we investigate the perception of personality through e-mail text.[1]

[1] Some of the work reported in this chapter is published as Gill and Oberlander (2003b).

7.1 Introduction

This thesis has two main hypotheses: The first relates to whether personality is projected through language, and this has been addressed in the previous chapters; The second hypothesis seeks to investigate whether personality can be perceived through language. We now address this second hypothesis, using the short e-mail texts written by authors of identified personality collected during the first stage of experimentation (Chapter 3).

Here we assess the ability of judges to rate and subjectively evaluate the personalities of the authors of these texts using a variety of measures. In this chapter we firstly describe our experimental methodology which is then followed by our results. We conclude with a discussion of these rating and evaluation results and relate them to our hypotheses.

7.2 Method

7.2.1 The Judges

The 30 judges were undergraduate or postgraduate students, or recent graduates currently living in Edinburgh (15 males, 15 females; mean age = 21.6 years, SD = 1.24). All were experienced e-mail users (rating themselves between 7 and 10 on a scale of 1-10, with 10 being 'a great deal'; mean = 9.23, SD = 0.77), and all were naive raters of personality (18 had no experience of personality, although 9 had 'some' experience (having read books on psychology) and 3 had studied psychology or personality psychology as part of their degree). No one had previously taken part in any personality rating experiments. Completion of EPQ-R (short form; Psychoticism Mean score: 3.17, SD 2.4; Normative score: M = 3.08, F = 2.35, Extraversion Mean score: 7.30, SD 2.6; Normative score: M = 6.36, F = 7.60, Neuroticism Mean score: 5.30, SD 3.1; Normative score: M = 4.95, F - 5.90, and Lie Scale Mean score: 3.27, SD 2.0; Normative score: M = 3.86, F = 2.71). We also collected data using the NEO-PI (short form), but we do not discuss this here.

7.2.2 Materials

The rating booklet sections were similarly structured for each personality trait: First a description of the personality trait was given, and then on each subsequent page after an introduction to the task, there was a target text followed by several questions relating to the judge's perception of the text's author.

7.2.2.1 The target texts

The target texts were all taken from the data collected previously. Six texts were chosen to represent a range of scores for each of the three-factor personality dimensions of Psychoticism, Extraversion and Neuroticism, on the basis of information for the group of 105 texts as a whole. Two texts were chosen whose authors scored greater than $+1$ standard deviation from the mean for that personality dimension, and two were chosen which were greater than -1 standard deviation (in each case these texts scored less than 1 SD either side of the mean on the other personality dimensions). Two further texts were selected which were within < 1 SD, but $> .5$ SD of the mean (one each above and below the mean - in these cases, the texts were within 1 SD (.5 SD where possible) of the mean on the other personality dimensions).

In this experiment texts detailing 'past' activities were selected for the rating exercise as these were generally longer than those outlining future plans (Mean length of texts in words: P=258.67, E=261.33, N=261.00; Further information about these texts and the participants who produced them can be found in Tables A.1 and A.2). These selected texts were presented in random order of personality score for each dimension at a time.

7.2.2.2 The questionnaire

The rating questionnaire was divided into three sections, each relating to a different personality trait (Psychoticism, Extraversion, or Neuroticism[2]) with the order in which these sections were presented determined by a Latin square technique to avoid an or-

[2]Note, however, that the terms Tough-mindedness and Emotionality were used instead of Psychoticism and Neuroticism; see also below and Section 2.1.2 for further details.

dering effect. These booklets were given an identification code which was used when referring to judges in order to maintain their anonymity.

The rating questionnaire booklet was prefixed by an explanatory page informing judges of the format of the experiment, and emphasising our interest in how they 'think the author comes across', the need for them to answer 'honestly and accurately' and 'not to spend too long thinking about each question' and to instead concentrate on giving their 'initial response'. For each personality dimension a description based upon those of Eysenck and Eysenck (1975) was included (as found in Section 2.1.1). These descriptions received minor re-wording to enhance intelligibility, minimise issues of social desirability, and to make them more understandable to a wider audience (as recommended by Eysenck and Eysenck, 1975, p. 12). Although it is more usual to rate personality using a standard set of questions (cf. Ten-Item Personality Inventory; Gosling et al., 2003), Sneed et al. (1998) have found that 'most laypersons can easily grasp the nature of the factors and their behavioural manifestations and can spontaneously recognise their grouping when presented with clear exemplars' (p. 115).

Judges were at first asked to rate the personality of the author for the trait which has been described at the beginning of the section, using the following question 'How [Tough-Minded/ Extravert/ Emotionally-Stable] is the the author of the e-mail', with the extremes of the scale labelled 'Not at All' and 'Very [Tough-Minded/ Extravert/ Emotionally-Stable]'. The judges were then asked 'How easy was it to come to this conclusion?' (about the e-mail author's personality) rated on a scale of 1–10 labelled 'Very Difficult' and 'Very Easy' respectively, and then to assess 'What aspects of the e-mail were most informative in reaching this conclusion about the e-mail author's personality: Topic (what they chose to write about); Vocabulary (words used); Style (how sentences were put together and followed on from each other)'. Each of these questions were rated on a scale of 1–10, identified as 'Very Difficult' and 'Very Easy' respectively. These ratings were then followed by two blank lines for which the judge was invited to 'Please explain/give examples'.

Two further questions were asked, firstly 'Please supply 5 words which you feel best describe this e-mail and/or its author' which was followed by 5 blank lines (which we do not report here), and lastly 'How similar would you say is this personality of

this e-mail's author to yours?' (again using a scale of 1–10, 'Very Different' – 'Very Similar').

7.2.3 Procedure

All 30 judges worked through the rating booklet at their own speed, and although there was no official time limit, they were encouraged to work 'quickly and efficiently' so that the participant didn't spend too much time thinking about their responses and also so that they remained well motivated. In all cases several judges participated in the experiment at the same time over-seen by the experimenter. However, they were informed that exam-type conditions should be maintained, and that responses to the questionnaire should not be discussed with each other during the experiment.

Equal numbers of participants were randomly assigned to each questionnaire. These questions are detailed above. After completing the rating booklet, there followed a debriefing section which asked judges to confirm that they were native English speakers, detail their experience of personality psychology, and to rate their 'previous experience using e-mail' (1–10; Very little – A great deal). EPQ-R and NEO-PI (both short form versions) personality questionnaires were administered to the judges, and upon submission of all these materials they received £10 for participating in the experiment.

7.3 Results

7.3.1 Consistency and Agreement of Judges' Ratings

All 6 authors for each of the three personality traits were scored on a scale of 1-10 by each judge. Concordance between the judges was measured using Kendall's W, and in all cases the Kendall coefficient reached a level of statistical significance, indicating relative agreement among judges concerning the trait score of each text. The value of these coefficients were: Psychoticism 0.2870 $[W(5) = 43.0453, p < 0.0001]$; Extraversion 0.4710 $[W(5) = 70.6445, p < 0.0001]$; Neuroticism 0.2664 $[W(5) = 38.9048, p < 0.0001]$.

In addition to using Kendall's W coefficient of concordance which describes judge

Judge	Psychoticism		Extraversion		Neuroticism		Mean r_s
1	0.396	(2)	0.199	(1)	−0.007	(0)	0.196
2	0.227	(0)	0.407	(0)	0.448	(1)	0.361
3	0.176	(0)	0.497	(2)	0.351	(1)	0.341
4	0.489	(0)	0.367	(0)	0.230	(0)	0.362
5	−0.142	(0)	0.014	(0)	0.466	(0)	0.113
6	0.482	(2)	0.594	(5)	0.253	(1)	0.443
7	0.378	(1)	0.682	(6)	0.341	(1)	0.467
8	0.362	(0)	0.155	(1)	0.090	(0)	0.202
9	0.413	(2)	0.533	(3)	0.246	(1)	0.397
10	0.309	(0)	0.537	(3)	0.442	(1)	0.429
11	0.367	(0)	0.666	(4)	0.220	(0)	0.418
12	0.333	(1)	0.422	(0)	0.300	(1)	0.352
13	0.092	(0)	0.429	(0)	0.490	(2)	0.337
14	0.493	(0)	0.178	(0)	0.540	(0)	0.404
15	0.510	(2)	0.400	(0)	0.237	(1)	0.382
16	0.463	(2)	0.314	(0)	0.285	(1)	0.354
17	0.380	(0)	0.501	(2)	0.383	(1)	0.421
18	0.327	(1)	0.520	(2)	0.299	(1)	0.382
19	0.100	(0)	0.569	(1)	−0.086	(0)	0.194
20	0.379	(2)	0.652	(6)	0.531	(1)	0.521
21	0.369	(1)	0.562	(2)	0.267	(0)	0.399
22	0.218	(1)	0.581	(6)	0.459	(0)	0.419
23	0.298	(1)	0.320	(0)	0.436	(1)	0.351
24	0.176	(0)	0.682	(7)	0.417	(1)	0.425
25	0.288	(0)	0.626	(7)	0.352	(1)	0.422
26	0.471	(3)	0.666	(6)	0.175	(1)	0.437
27	0.340	(1)	0.642	(3)	−0.112	(0)	0.290
28	0.403	(1)	0.541	(2)	0.449	(0)	0.464
29	0.429	(0)	0.602	(2)	0.349	(0)	0.460
30	0.472	(3)	0.613	(5)	0.374	(2)	0.486
Mean r_s	0.333		0.482		0.308		0.374

Note. Agreement is described by the mean correlation of each judge with other judges for each scale. The number of statistically significant positive correlations (at the $p < .05$ level) is shown in brackets, maximum 29 per cell.

Table 7.1: Inter-Judge Agreement correlations for raters

consistency overall, it is also possible to examine how the each judge agrees with each of the other judges in the experiment (cf. Morris et al., 2002). Correlations were performed for each judge with each of the other judges, with the mean overall correlation reported for each judge (counts of correlations achieving significance are also noted for each cell out of a maximum of 29). Although the personality questionnaire results can usually be regarded as interval data (Kline, 1983), the ordinal nature of the rating scale responses meant that Spearman rank correlations were used throughout the following analyses, since this is more appropriate for such data (Butler, 1985).

The final row of Table 7.1 gives the average rank correlations for each trait across all judges. Extraversion is shown to have the greatest inter-judge agreement, and therefore in terms of inter-judge agreement appears to be the easiest trait to rate (mean $r_s = 0.482$). This is followed by Psychoticism (mean $r_s = 0.333$), and finally Neuroticism (mean $r_s = 0.308$) which both show lower levels of agreement and therefore suggests that they are harder to rate. The greater agreement shown between judges for ratings of Extraversion is also reflected in the total number of significant correlations found for the trait (76), which is much greater than that found for either Psychoticism (26) or Neuroticism (20).

Since we calculate Spearman rank correlations, here we have reported the means of these correlations (Morris et al., 2002), rather than use Fischer's r to z conversion (e.g., Funder and Colvin, 1988; Funder et al., 1995; Spain et al., 2003). Therefore in order in order to establish the significance of agreement between judges, intraclass correlations were calculated across the thirty judges for their ratings of P, E, and N targets, since this statistic is regarded as the equivalent of performing correlations between all possible pairs of raters (McCrae and Costa, 1987). Similarly to the findings reported in Table 7.1, Extraversion showed the highest agreement with an intraclass correlation of 0.4025, and although Neuroticism and Psychoticism both showed relatively low agreement, this was actually slightly lower for Psychoticism (0.2055) than for Neuroticism (0.2476; all significant at $p < 0.0001$).

7.3.2 Are All Judges Equally Good?

The level of agreement between judges across all three personality traits is also shown in Table 7.1. From this it can be seen that the best judges, *in terms of agreeing most with the others* were judges 20, 30, 7, 28, and 29, and the worst judges were 5, 19, and 1. The mean level of agreement across P, E, and N dimensions was .374.

Turning to each trait individually, for Psychoticism judges 15, 4, 6, 30, and 16 showed the most agreement, whilst judges 5, 13, 19, 24, and 3 showed relatively little agreement. For Extraversion, judges 7, 24, 11, 26, and 20 demonstrated greatest agreement, whereas for judges 5, 14, and 1 the levels achieved were much lower. For Neuroticism it can be seen that judges 14, 20, 13, and 5 all show the most agreement, whereas judges 27, 19, and 1 actually show disagreement with other judges.

The level of agreement between target and judge ratings can also indicate how accurate judges are, and information about this can be found in Table 7.2. Here it can be seen that the best judges when defined as agreeing most with targets across all personality dimensions are judges 21, 17, 6, 11, 18, and 28 and the worst judges are 8 and 12 who both correlated negatively, and judges 13 and 5.

For each individual trait, starting with Psychoticism, judges 28, 14, 21, and 17 all agreed highly with the targets, whereas judges 5, 3, and 22 showed a negative correlation with the target self reports of personality. The trait of Extraversion elicited even higher levels of target-judge agreement for judges 20, 22, 25, and 26, with only judges 1 and 8 showing a negative correlation. However, for Neuroticism lower levels of agreement were found for judges 18 and 21, with many judges showing a negative correlation (16 in total), with some of the greatest disagreement found for judges 12 and 13. Additionally we analysed inter-judge and judge-target agreement by the personality traits of the judges (EPQ-R and NEO-PI-R), but this appeared to demonstrate little effect on levels of agreement.

7.3.3 Are All Targets Equally Good?

If one text on a particular personality trait was much more difficult to rate than any of the others, we would expect judges to show a much greater variability in their rat-

ings for it. Levene's test for homogeneity (or equality) of variance was used to investigate whether there was significant variance in ratings for texts belonging to each trait. Although significant differences were not found for Extraversion or Neuroticism, they were found for Psychoticism [$F(5,174) = 2.8682, p < .05$]. In this case, the texts which showed the greatest variance were P6 (M=4.4, SD=2.3; mid-high-P), P5 (M=4.4, SD=2.0; high-P), and P3 (M=5.2, SD=2.0; high-P), and therefore appear to be the most difficult to rate. The texts showing least variance were P4 (M=2.7, SD=1.4; mid-low-P), P1 (M=2.8, SD=1.6; low-P), and P2 (M=3.5, SD=1.9; low-P). This demonstrates that the High Psychotic texts showed greater variation in ratings, and may indicate that they were harder to rate, therefore resulting in the lower intra-class correlation results for ratings of Psychoticism.

7.3.4 Target-Judge Correlation

To gain an overall sense of how the individual judges had performed, mean correlations of judge-target agreement were calculated. For each of the judges, each of their six ratings of the texts for P, E, and N were correlated with the original personality scores of the authors, and their mean performance for rating P, E, and N also noted (Table 7.2). Looking at the correlations of the individual judges for each dimension, we can see that the largest number of significant correlations (out of a possible 30) were found for Extraversion (5), followed by Psychoticism (2), with none of the correlations between judges and targets reaching significance for ratings of Neuroticism.

To ensure increased agreement and accuracy of target-judge correlation, the aggregate measure of personality ratings across multiple raters was then calculated, since McCrae and Costa (1987) suggest that this takes into account how the target is seen by the judgement group as a whole. Therefore Spearman correlations were performed taking the mean of the judges ratings for each text, along with the original personality scores of the targets. Correlation of the target's raw EPQ-R with the mean of the judges ratings (1–10), gave the following correlations (Spearman, pairwise, two-tailed, 6 cases): Extraversion $r_s = .8857$; Psychoticism $r_s = .7537$; Neuroticism $r_s = -.3769$; of these, only ratings of Extraversion showed significant target-judge agreement ($p < 0.05$).

Judge	Psychoticism	Extraversion	Neuroticism	Mean r_s
1	0.729	−0.114	−0.186	0.143
2	0.200	0.714	−0.614	0.100
3	−0.200	0.700	0.386	0.295
4	0.571	0.314	−0.257	0.209
5	−0.229	0.329	0.100	0.067
6	0.771	0.829	0.157	0.586
7	0.386	0.886	0.300	0.524
8	0.071	−0.143	−0.329	−0.134
9	0.586	0.714	0.214	0.505
10	0.000	0.814	−0.243	0.190
11	0.500	0.800	0.429	0.576
12	0.114	0.286	−0.557	−0.052
13	0.171	0.329	−0.486	0.005
14	0.929*	0.329	0.343	0.534
15	0.686	0.629	−0.229	0.362
16	0.543	0.457	0.157	0.386
17	0.829	0.757	0.300	0.629
18	0.357	0.814	0.500	0.557
19	0.214	0.700	0.443	0.452
20	0.629	0.986*	−0.157	0.486
21	0.886	0.757	0.529	0.724
22	−0.057	0.929*	0.157	0.343
23	0.500	0.457	−0.243	0.238
24	0.429	0.971*	−0.300	0.367
25	0.700	0.929*	−0.100	0.510
26	0.600	0.929*	−0.443	0.362
27	0.500	0.814	−0.186	0.376
28	0.943*	0.671	0.057	0.557
29	0.571	0.714	−0.071	0.405
30	0.629	0.771	−0.357	0.348
Aggregate r_s	0.754	0.886*	−0.377	

Note. Significance denoted by * is at the $p < .05$ level.

Table 7.2: Target-Judge agreement correlations

7.3.5 Judge Perception of Target Rating

7.3.5.1 Perceived similarity of target-judge

In order to investigate how judges perceived the target author personalities relative to their own, analysis of the similarity ratings of texts was performed.

These analyses were carried out with the six target texts for each personality dimension grouped into three categories of high, mid, and low. A within subjects analysis of variance (ANOVA) revealed effects of text personality type on ratings of similarity for Psychoticism [$F(2,58) = 7.999, p < .001, MSE = 1.6126$], and also this time for Extraversion [$F(2,58) = 4.052, p < .05, MSE = 1.6238$], but not Neuroticism texts. Tukey HSD tests revealed that significant differences in similarity ratings were found between LowP ($M = 5.6$) and HighP ($M = 4.3$), and also HighP ($M = 4.3$) and MidP ($M = 5.1$) Psychoticism texts and between the HighE ($M = 5.3$) and MidE ($M = 4.3$) Extraversion texts (all significant at $p < .05$).

When texts were further categorised into either high or low on their particular personality dimension, ANOVA showed effects of personality on ratings of similarity only for Psychotic texts [$F(1,29) = 12.090, p < .001, MSE = 1.1612$] (LowP $M = 5.6$, HighP $M = 4.3$).

These analyses have so far not taken into account the effects of judge personality on the ratings of similarity, but have grouped the judges as a whole. Therefore, judges were categorised as either 'high' or 'low' on the personality dimension in question using a mean split, and author personality of the target texts was categorised into the 'high', 'mid', or 'low' groups since this reduced the data yet retained broad information. A two factor mixed-design ANOVA revealed for Psychoticism main effects of judge personality type [$F(1,28) = 6.555, p < .05, MSE = 3.0586$] and as would be expected personality of text author [$F(2,56) = 8.063, p < .001, MSE = 1.5999$], however no interaction effect was found between judge personality and text author personality in the ratings of similarity. For Extraversion, as expected, a main effect was found for text personality type on similarity rating [$F(2,56) = 4.390, p < .05, MSE = 1.4982$], and also an interaction effect for rater and text personality upon similarity ratings [$F(2,56) = 3.430, p < .05, MSE = 1.4982$]. No effects were found for Neuroticism.

In order to investigate possible interaction effects further, we examine the simple

main effects of text author personality for the high and low personality groups of judges individually. The within subjects ANOVA shows—as expected from the significant interaction—effects of text type on the ratings of similarity for High Extravert judges [$F(2,26) = 5.082, p < .05, MSE = 1.8988$]. Tukey tests reveal significant effects ($p < .05$): The High Extravert judges rated the HighE texts as most similar to themselves ($M = 6.1$) and the MidE texts as least similar ($M = 4.5$).

However, findings for Psychoticism also show an effect of text type on similarity rating for the Low Psychotic judges [$F(2,32) = 5.753, p < .01, MSE = 1.8848$]. Tukey tests revealed significant results ($p < .05$), with Low Psychotic judges rating themselves as most similar to the LowP texts ($M = 6.2$), and most dissimilar to the HighP texts ($M = 4.6$). For High Psychotic judges, MidP texts were regarded as most similar ($M = 4.8$), and HighP texts most dissimilar ($M = 3.8$), but this effect of text type was found to be border line significant at $p < 0.1$ [$F(2,24) = 3.299, p < 0.1, MSE = 1.2201$]. No significant effects were found for judges grouped by Neuroticism.

If the actual personality scores of the texts being rated for similarity are disregarded, and the personality scores of the raters are considered (again divided at the mean as either high or low), then between subjects ANOVA shows that only rater Psychoticism has an influence on ratings of Psychotic texts [$F(1,28) = 6.556, p < .05, MSE = 1.0195$]. This means that LowP judges rated the texts (*all texts*, high *and* low P) as more similar ($M = 5.4$) than HighP judges ($M = 4.4$) ($p < .05$).

7.3.5.2 Perceived ease of rating personality

Indications of how judges perceived ease of rating personality of texts were gained from the subjective scores. Within subjects ANOVAs were performed for ratings of ease compared with the personality of the text author categorised into 'high', 'mid' and 'low'. ANOVAs show that significant effects of the personality of the text upon rating difficulty for Extraversion [$F(2,58) = 13.155, p < .001, MSE = 3.1689$] and Psychoticism [$F(2,58) = 10.368, p < .001, MSE = 1.8522$]. Tukey tests show that significant differences for Extraversion exist between LowE ($M = 5.7$) and HighE ($M = 7.8$), and between HighE ($M = 7.8$) and MidE ($M = 5.8$) texts, and for Psychoticism between LowP ($M = 7.4$) and HighP ($M = 5.9$), and between HighP ($M = 5.9$) and MidP

($M = 7.1$) texts (all $p < .05$).

When the personality categorisation of the rated texts is further reduced to 'high' and 'low' ANOVAs again reveal significant effects for personality of text, and perceived ease of rating for Extraversion [$F(1,29) = 17.540, p < .001, MSE = 2.1893$] and Psychoticism [$F(1,29) = 21.856, p < .001, MSE = 1.1011$]. The means for these significantly different groups are LowE ($M = 5.6$) and HighE ($M = 7.2$), and LowP ($M = 7.4$) and HighP ($M = 6.1$).

These analyses so far have grouped the judges as a whole and not taken into account the effects of judge personality. Although effects of judge personality are not expected to have as great an effect on judgements of difficulty, as they do on similarity—as this is built into the measure—for completeness this analysis was carried out. As before, judges were categorised as either 'high' or 'low' on the personality dimension in question using a mean split, and author personality of the target texts was categorised into the 'high', 'mid', or 'low'. A two factor mixed-design ANOVA revealed main effects of personality of text author for Psychoticism [$F(2,56) = 10.592, p < .001, MSE = 1.8130$], and Extraversion [$F(2,56) = 12.820, p < .001, MSE = 3.2516$] but not for Neuroticism, as would be expected from the previous analyses. However, no main effects of judge personality or interaction effects of judge personality and text author personality were found for any of the traits.

7.4 Discussion

7.4.1 Ratings of Inter-Judge and Target-Judge Agreement

These results demonstrate that judges reliably agree with each other when rating a text for a specific personality trait. However the level of agreement is greatest for Extraversion, followed to a lesser extent by Psychoticism and then Neuroticism.

That Extraversion shows the greatest inter-judge agreement is compatible with previous literature, and suggests that this may be due to its more observable and less evaluative properties (see Section 2.3 for a discussion). However, in the case of John and Robbins's analysis, Neuroticism (termed Emotional Stability in their model; along with Intellect, or Openness to Experience) shows quite good agreement, with this re-

duced for Conscientiousness and lower still for Agreeableness (John and Robbins, 1993). In the present study, because the three factor (EPQ-R) personality model was used, Psychoticism has replaced the Intellect, Conscientiousness and Agreeable traits, which has left Neuroticism as the trait showing least inter-judge reliability. Because we are trying to compare two different models of personality, it is difficult to assess whether in the current study Neuroticism has been shown to demonstrate less agreement in judges than in previous studies, or whether in fact Psychoticism is more observable and less evaluative than the individual traits of Conscientiousness and Agreeableness.

However, since the actual ratings in the current study are using a different novel source of information as the target (a short sample of e-mail text rather than having met the person in real life or through observation; cf. Markey and Wells, 2002, who used an interactive CMC chatroom environment), this difference in rating agreement, for both Neuroticism (Emotional Stability) and Psychoticism (Intellect/ Conscienciousness/ Agreeableness) may be due to the properties of e-mail text as not being 'good information' for personality judgement of Neuroticism (Funder, 1995, see Section 2.3 for further discussion).

Turning to the agreement between the judges' and targets' rating of personality, and a similar pattern emerges to that of inter-judge agreement, with ratings for both Extraversion and Psychoticism showing a relatively stronger positive correlation, but Neuroticism bearing a non-significant negative relationship. This again points to Extraversion being an observable, but relatively unevaluative trait, its evaluative neutrality emphasised by self-peer agreement. The weaker target-judge relationship for Psychoticism would suggest that it is both less observable and more evaluative. However the lack of strong target-judge relationship for Neuroticism relative to Psychoticism (given their similar inter-judge agreement) would suggest its much greater evaluativeness results in a distortion of self-perception, or alternatively that e-mail does not provide good information for its accurate judgement.

An alternative explanation may be that the judges are attending to the wrong information. In a study which looked at personality perception through speech, Scherer (1972) found that despite the high rate of inter-rater reliability for the trait of Extraver-

sion, there was little target-judge agreement for this trait. He concluded from this that judges were instead attending to stereotyped cue information for socially desirable traits projected by the targets. Therefore in our case of Neuroticism, judges similarly may be attending to misinformed stereotyped cues. However, given that Neuroticism is generally regarded as more evaluative and less desirable, it may be that they attend to less desirable stereotyped features.

When the performance of individual judges is examined, it can be seen that inter-judge agreement can be differentiated across the traits: on some this can be quite high, and on others—especially Neuroticism—this can be quite low. In the case of target-judge agreement the pattern is more consistent, with judges generally showing either generally higher or lower levels of agreement across all traits. This greater consistence of agreement is to be expected due to the judge's ratings only being correlated with those of the target rather than all of the other judges. As expected from the mean ratings of judges overall, most judges show a noticeably poorer performance for Neuroticism.

7.4.2 Judge Perception Rating Measures

Additionally, we also collected novel subjective ratings of similarity between rater and target, and perceived ease of rating the text for personality. This data is informative because it allows us investigate how perceptions of the rating exercise and of own and other personality compare to objective measures.

For the similarity ratings there was a general pattern of the judges distancing themselves from the undesirable high end of the trait. Even when judge personality was taken into consideration, the judges were still seen to identify with low Psychoticism, meaning that, whilst the Low Psychotic judges (accurately) rated the low Psychotic texts as most similar, the High Psychotic judges also (incorrectly) rated the low Psychotic texts as most similar.

Although it may be the case that highly Psychotic judges are for some reason less able to accurately judge author Psychoticism, it would appear to be more likely that they were influenced by the evaluativity of this trait. Indeed, it may be that as a result of higher levels of judge Psychoticism, such judges are more likely to consciously or unconsciously provide inaccurate information about themselves. Level of judge

Psychoticism also had an overall effect on the similarity scores, with Low Psychotic judges regarding themselves as more similar in general to the authors of the texts. Given that lower Psychoticism scorers are more likely to be interpersonally oriented it should not be too surprising that they more readily identify with the authors of the texts, regardless of how similar their personality scores actually were.

For the judges of Extraversion overall, a relationship was only shown between the texts when grouped into three categories, with the high Extravert text regarded as more similar than the mid text. When personality information is added to this analysis, an interaction effect emerges between the personality of the judge and the author of the text. Separate analysis of the high and low Extravert similarity ratings shows that the High Extraverts view the high Extravert texts as most similar by quite some way (followed, surprisingly, by the Introvert texts). On the other hand, this interaction is mirrored by Low Extravert judges (not significantly) rating the Introvert texts as most similar, followed shortly after by the high Extravert texts. Since both groups accurately rate the texts which are most similar to themselves, this contributes to the interaction effect. However because Low Extravert judges rate the high Extravert texts as still relatively similar, this contributes to the overall effect for High Extravert texts being rated as similar for the group as a whole.

Since an interaction of judge and author personality occurs, this suggests that effects of trait desirability, or undesirability, are less important for ratings of Extraversion, and this is confirmed by it being regarded as a less evaluative trait. Furthermore, the fact that high Extraverts more readily identified their similarity more accurately may be a result of their greater interpersonal ability associated with higher Extraversion. However, the fact that Low Extraverts are less likely to distinguish themselves as low Extraverts as opposed to high Extraverts may be an effect of a lower interpersonal awareness or a remnant of weak desirability effects of higher levels of Extraversion.

So far we have discussed the accuracy and relative desirability effects present in the similarity ratings for Psychoticism and Extraversion, without reference to Neuroticism. Whilst Psychoticism and Extraversion have shown several broad patterns relating to similarity ratings, perception of similarity to Neuroticism show few patterns and again demonstrate a mixed up picture.

Extraversion											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	○	○	○	○							
Real.	○	○	○								
Fluency	○	○	○								
Gramm.	○	○	○	○	○	○	○	○			
Conv.	○	○	○	○	○	○	○	○			
Lexis	○	○	○	○	○	○	○	○	○	○	
Percept.	$\oplus^c$	$\oplus^c$									
Neuroticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	○	○	○	○	○						
Conv.	○	○									
Lexis	○	○	○	○	○						
Percept.	$\oplus^c$	$\oplus^c$									
Psychoticism											
Hypotheses	1	2	3	4	5	6	7	8	9	10	Other Findings
Theory	○	○	○	○	○	○	○				
Conv.	○										
Lexis	○	○	○	○	○	○	○	○	○	○	
Percept.	$\oplus^c$	$\oplus^c$									

Table 7.3: Review of hypotheses

Note. ○ indicates an hypothesis; ⊕ confirmation of hypothesis; ⊖ inverse of hypothesis; ⊙ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{c} perception study.

Turning to the perceived ease of rating texts, we find that these findings are consistent across both Extraversion and Psychoticism dimensions: high Extravert texts and low Psychotic texts are regarded as the easiest to rate, regardless of judge personality. These findings are consistent in that they show that texts belonging to the more desirable end of both scales, are seen as easier to rate, and therefore are not an artifact of the rating scale description (in which case the higher ends of the scales would have been regarded as easier to rate). We therefore suggest that it appears to be the case that individuals have a better concept of behaviour which is desirable, rather than undesirable.

7.4.3 Summary of Findings and Evaluation of Hypotheses

Given our results presented above, we now relate them to our hypotheses: Firstly—for the second hypothesis of the thesis, *that personality is perceived through language*—

we have shown that this is indeed the case. However, we note that the ability to perceive personality is not consistent for all traits, here we found that Extraversion, and to a lesser extent, Psychoticism were accurately perceived (i.e., there was relatively high target-judge, and also inter-judge agreement). Although we found reasonable inter-judge agreement for Neuroticism, there was target-judge *dis*-agreement. Secondly, we can relate these findings to our Perception hypotheses presented in our review of literature. These were as follows (reproduced from Section 2.3.5):

Extraversion This trait will be the most easily perceived due to its high visibility and low evaluativeness. We therefore expect it to show the highest levels of inter-judge (1) and target-judge agreement (2), even in CMC at zero-acquaintance.

Neuroticism We expect that agreement (within judges [1], and target-judge [2]) will be lowest for Neuroticism, due to its high evaluativeness and low visibility, which we predict will be most affected by the lack of information available in the CMC and zero-acquaintance conditions.

Psychoticism Since we propose that Psychoticism is visible, but evaluative, we expect agreement to be higher than for Neuroticism, but lower than for Extraversion (inter-judge [1], and target-judge [2]). We also expect that the conditions will only have moderate lowering effect upon agreement.

We now address these hypotheses with reference to our findings (summarised in Table 7.3). For individual trait beviour, our hypotheses were confirmed since Extraversion displayed the highest levels of inter-judge and target-judge agreement in the CMC and zero-acquaintance environment; Psychoticism showed the next highest levels, presumably due to its relatively high observability combined with high evaluativeness; For Neuroticism, both measures of agreement were the lowest, which was expected due to its high evaluativeness and low observability.

Additionally using our subjective judge perceptions, we found that for the traits which appeared perceptible, the more socially desirable ends of the scale (i.e., high Extraversion and low Psychoticism) were regarded as easier to rate. Furthermore, the greater evaluativeness of Psychoticism was confirmed by judges rating themselves as

more similar to the desirable low end of the scale; since Extraversion is less evaluative, this resulted in judges accurately rating themselves as being more similar to targets who in fact were similar in levels of Extraversion.

7.5 Conclusion

From a text of around 300 words, 30 judges were able to consistently agree (both with each other and with the target individual's self-rating), on the personality of the text's author when rating them for Extraversion and also to a slightly lesser extent, for Psychoticism. In both cases, judges used a general subjective rating of personality rather than an itemised personality questionnaire. Additionally, judges rated ease of assigning personality and also perceived target similarity, which confirmed the judge's ability to perceive personality consciously and subconsciously, and also the relative evaluativeness and desirability of these traits.

Although judges generally agreed with each other regarding ratings of Neuroticism, little consistency was found with the author's own personality assessment, or with ease of rating or similarity. We propose that this is partly due to characteristics of the trait itself, and also the quantity and quality of information which the e-mails in this experiment made available to the judges.

Chapter 8

Conclusion

In the conclusion we draw together the different work which has been presented in this thesis. We briefly address the main hypotheses of the thesis, before summarising it chapter-by-chapter. Then, we draw particular attention to the contributions that this thesis has made: After presenting specific findings for personality and language, we show how these can be used to inform the current state-of-the-art for both natural language processing and models of language production.

Finally, we note the boundaries of this thesis, before examining how these restrictions—and other questions raised as a result of this study—can be addressed in future work. We close with some final words.

8.1 Summary of the thesis

The primary hypotheses that we address in this thesis are captured in its title: *Personality and Language: The projection and perception of personality in computer-mediated communication*. The first hypothesis is that personality is projected in some way by the language of the author; the second hypothesis is that the personality of an author of a piece of writing is perceptible to the reader.

In both cases we support the hypotheses. We trace out the journey through the thesis which led us to these conclusions as follows:

In the first chapter, the introduction, we describe why personality is important and its implications for behaviour. In doing so, we demonstrated the relevance of personality to cognitive science, social psychology, and computational linguistics. Furthermore we described how personality and language—particularly in computer-mediated communication—can inform person perception and natural language generation. The objectives, the boundaries, and the structure of the thesis were then presented.

The second chapter formed the review of literature, which we divided into two sections by relevance to, firstly, topic or, secondly, methodology. We introduced the topic section by establishing the concept of personality traits and theory which form the basis of this thesis. This was followed by a discussion of previous studies which have examined the way personality is projected through language. Finally, we discussed studies which have looked at the perception of personality, which we framed in terms of Funder's 'Realistic Accuracy Model'.

In our review of methodological literature, we first examined issues surrounding our chosen medium of investigation, the computer-mediated environment: here we reviewed the implications for running psychological experiments, and for personality, language, and communication. Finally we discussed linguistic analysis methods relevant to corpus studies. We introduced the basic approach, annotation, and relevant analysis methods, before describing the main theory-driven and data-driven techniques, with reference to the literature.

The third chapter described the building of the e-mail personality corpus which forms the basis of the experimentation conducted in this thesis. Here we discussed the methodology which was adopted for data collection, and we then used factor analysis

to enable the comparison of our data to a previous study. This demonstrated a general replication of the earlier work, despite differences in the experimental task, communication medium, variety of English used, and size of data-set. When we correlated the resulting factors and the original variables from the content analysis with our measurement of personality, we generally replicated their results, but with lower levels of significance. We explained this in terms of the smaller size of our corpus, with other differences due to variation between the corpora. Additionally, we observed that multi-dimensional methods may not be most appropriate for such analyses, especially given the modest size of our corpus. Alternative approaches are investigated in the next three chapters.

The fourth chapter examined alternative approaches to the multi-dimensional technique used in chapter three. Specifically we adopt techniques which will retain the maximum linguistic information from the content analysis methods of the previous chapter, relative to author personality. Here we use multiple regression analysis to determine the combination of content-analysis features which show the greatest relationship to each of the personality dimensions. The generalisability of different analyses are discussed.

The content analysis used so far is based on human-rated psychological categories. In the second half of this chapter we investigate more empirical approaches. Firstly, we note that the lexical density measured as part of the content analysis does not account for text length. We therefore explore several lexical density measures which are independent of text length. Secondly, we use a dictionary resource of experimentally derived psycholinguistic properties, to analyse our texts. Again, multiple-regression analysis is used to identify the combination of features which show the greatest relationship to each of the personality dimensions.

The results from these analyses identify rich psychological and psycholinguistic properties for Psychoticism, and to a lesser extent for Neuroticism. However, these techniques are less successful for Extraversion. In summary, we discuss potential explanations for the difference in findings based on personality theory and the top-down and content-based analysis methods used. In the following chapter we explore data-driven approaches.

The fifth chapter introduces our data-driven approach. This involves dividing our corpus according to the personality of a text's author, and enables the use of corpus-comparison techniques on these personality subcorpora. In this chapter we evaluate the different ways in which the authors can be grouped: initially, we isolate the extreme personality scorers on each dimension and compare these high and low groups. However, this does not allow us to view linguistic behaviour along the middle section of a dimension, and we also note that the behaviour of the extreme groups is not independent of other dimensions. We therefore propose the redefinition of the extreme groups so that they consist of authors who show independence across the personality dimensions, that is, they are only extreme scorers on one dimension. Furthermore, to trace behaviour across the middle section of a trait, we also employ a subcorpus of authors who are not extreme scorers for any of the personality dimensions. Although this more restricted classification of authors gives smaller subcorpora, we regard these analyses as more informative.

A further issue for the data-driven approach involves exploring linguistic analysis methods suitable for comparing the subcorpus groups. Firstly we adapted an approach used to identify terminology specific to textual genres, and applied this to the comparison of the groups at the extremes of the dimensions. From this analysis we were able to identify major areas of language use which systematically varied across the personality dimensions. We refined this analysis further, to give us a greater selection of features in the analysis of our more restricted subcorpus groups. Additionally, we investigated more sophisticated methods of corpus annotation which enable greater abstraction away from content-specific features.

In summary, this chapter describes data-driven analyses which show greater sensitivity, and allow more sophisticated linguistic analysis which takes into account contextual information. For each of the personality dimensions, we identify a number of characteristic features or behaviours. Annotation methods are adopted which reduce content-dependence. Here we also refine methodological techniques to allow further examination of linguistic behaviour along each of the personality dimensions, with these behaviours independent of other dimensions. However, this analysis method leads to smaller subcorpus sizes. Therefore, in the next chapter we investigate further

annotation techniques.

The sixth chapter expands further upon the data-driven analysis methods developed in the previous chapter. In order to overcome the potential limitations of this technique for our data—namely data sparsity and content-specificity—we explore more sophisticated, higher-level annotation methods. Firstly, we analyse a version of our corpus which has been tagged for part of speech (grammatical) information. A basic analysis which solely examined the relative usage of a reduced set of grammatical categories showed few features which distinguished the groups. These results therefore failed to replicate some of the previous findings for personality and non-native speakers. However, when we used more specific syntactic categories and our more sophisticated analysis technique which considers contextual information, we found grammatical sequences which characterised the different personality groups.

The second part of the chapter uses semantic analysis which categorises and tags words according to their meaning, which to some extent relates to content analysis. Here we again examined coarser-grained tag classifications and contextual information, and again these proved to be less successful in characterising the differences in linguistic behaviour between personality groups. Again, contextual information and fine-grained semantic and grammatical classifications proved to be valuable in distinguishing author personality.

In this chapter we perform the most advanced syntactic and semantic analyses of our personality corpus. These are the least prone to data-sparsity, and given their content-independence, show the greatest generalisability. For all personality dimensions we found characteristic linguistic behaviour, however, for Psychoticism this appears to be mainly encoded in fine-grained grammatical and semantic choices.

The seventh chapter turns our focus of attention from the projection of personality, to its perception through language. Here we describe an experiment which evaluates how good people are at judging personality from a short written text. In the first half of the chapter we look at inter-judge and target-judge agreement. The results show that, as expected, more visible and less evaluative traits show greater agreement. Computer-mediated communication at zero-acquaintance leads to lower levels of inter-judge agreement. However for target-judge agreement, the impoverished cues

additionally tend to exaggerate the effects of visibility and evaluativeness.

This chapter additionally explores other information gathered from the judges relating to their subjective perceptions. In summary, we show here that personality can be accurately perceived from short written texts, although levels of accuracy vary according to the dimension being rated. These results mirror previous findings for personality rating in other contexts.

8.2 Significant findings of the thesis

The previous discussion gives a broad overview of the thesis and notes broad findings and methodological concerns. We now turn to discuss the specific findings of the thesis and relate these back to the hypotheses which we presented at the beginning of the thesis. We start by presenting again the hypotheses of the thesis; then follows a summary of the findings and how these relate to the hypotheses; we then highlight the most significant and interesting results.

8.2.1 Re-presentation of hypotheses

8.2.1.1 Extraversion

Theory We expect the language of high Extraverts to reflect their sociability by referring to other people (1), and to express their activity by using words associated with actions (2; cf. Grammatical Hypotheses, increased use of verbs), and by saying more (3). We also expect them to use language which suggests positive affect (4).

Realisation Extravert 'loudness' will be realised in the increased use of capital letters (1) and exclamation marks (2); Worse pronunciation will result in worse spelling and more typographical errors (3).

Fluency Higher speech rate of Extraverts will be realised in longer sentences (1); shorter pauses and less hesitation will result in more ellipses (...)(2) and hyphens (-)(3) being used to separate clauses rather than the full stop.

Grammatical Extravert language will contain more adverbs (1), pronouns (2), and verbs (3)(i.e., more 'implicit'), and have a lower lexical density (TTR) (4); it will contain fewer nouns (5), modifiers (6) and prepositions (7)(less 'explicit'), and be less formal (8).

Conversational Extraverts will write more (1) (cf. Theory hypotheses); initiate more laughter, perhaps indicating this by explicit references ('ha') or by exclamation (2); they will refer to themselves more (3); they will use more terms indicating pleasure and agreement (4), pay more compliments (5); they will use fewer hedges (6) and references to problems (7), and will show less anxiety during the communication (8).

Lexis Extraverts will show a greater use of social (1), inclusive (2), and positive emotion words (3), and use fewer negations (4), tentativity (5), exclusive (6), causation (7), negative emotion words (8), and articles (9); In terms of factors, Extraversion will have a negative relationship with Making Distinctions (10).

Perception Extraversion will be the most easily perceived due to its high visibility and low evaluativeness, we therefore expect it to show the highest levels of inter-judge (1) and target-judge agreement (2), even in CMC at zero-acquaintance.

8.2.1.2 Neuroticism

Theory The language of high Neurotics, we expect to be highly emotional—particularly expressing negative affect (1), but also positive affect (2; cf. LIWC which predicts fewer positive emotion words)—and this is also revealed through intensified language (e.g., adjectives [3] and adverbs [4]). Since the individual tends to focus more on themselves, we also expect this self-preoccupation to be expressed through an increased reference to self (5).

Conversational High Neurotics will use a lower lexical density (1) (TTR), and show greater anxiety during communication, realised explicitly through references to 'worry' or 'stress' (2).

Lexis High Neurotics will use more first person singular (1) and negative emotion words (2), and fewer positive emotion words (3), and articles (4); Neuroticism will also correlate positively with the Immediacy factor (5).

Perception We expect that agreement (within judges [1], and target-judge [2]) will be lowest for Neuroticism, due to its high evaluativeness and low visibility, which we predict will be most affected by the lack of information available in the CMC and zero-acquaintance conditions.

8.2.1.3 Psychoticism

Theory We expect highly Psychotic individuals to reflect their lack of sociability and detachedness by making fewer references to themselves (1) or to others (2), and to demonstrate their harshness and toughness by avoiding emotional words (positive [3] and negative [4]). Since they are creative and enjoy unusual things, we predict that they will use more unusual language, realised both in words and constructions (i.e., lexically and syntactically). This we predict would result in for example, the use of less frequent words (5), a higher type-token ratio (6), and use of passive constructions (7).

Conversational We expect high Psychotics to show less anxiety during communication, for example, through fewer explicit references to 'stress' or 'worry' (1).

Lexis Here we predict Psychoticism will show an inverse relationship to Agreeableness and Conscientiousness: based on Agreeableness, we expect fewer first person singular (1) and positive emotion words (2), and more articles (3) and negative emotion words (4), and also a negative correlation with the factor Immediacy (5); on the basis of the findings for Conscientiousness, we expect fewer positive emotion words (=2), and more negations (6), negative emotion (=4), causation (7), exclusive words (8), and discrepancies (9), and a positive relationship with the factor Making Distinctions (10).

Perception Since we propose that Psychoticism is visible, but evaluative, we expect agreement to be higher than for Neuroticism, but lower than for Extraversion

(inter-judge [1], and target-judge [2]). We also expect that the conditions will only have moderate lowering effect upon agreement.

8.2.2 Evidence for the hypotheses

Chapter by chapter evidence from the thesis which informs the hypotheses is presented together in Table 8.1. Overall information relating to whether a hypothesis is proved or disproved is indicated with an appropriate key, as is the experimental source of this information. In the case where evidence from different sources is brought to bear on a hypothesis, the large key indicates the overall outcome, with super- or subscript notation describing the evidence in more detail (in the case of contradictory evidence, positive [$\oplus$] and negative [$\ominus$] evidence results in partial evidence [$\odot$], however $\oplus$ or $\ominus$ take precedence over $\odot$). Findings derived from this thesis in addition to those directly addressing the hypotheses, are listed for Chapters 3–6 in subsequent columns.

Overall there is confirmation for the majority of the hypotheses, although it is notable, that this is a result of evidence accumulated on the whole throughout the thesis, rather than from one particular set of analyses (the exception, of course, being the perception hypotheses). Overall the most successfully addressed hypotheses were those concerned with conversational features or theoretical predictions (or in the case of Extraversion, concerning grammar). Perhaps the least successfully addressed predictions were those derived from previous LIWC findings and were largely concerned with lexis or content. This is not to say that content or lexical information is not a useful indicator of personality—indeed our semantic analysis found many additional features for all personality types—however it appears to underline the topic-dependence of content analysis. As we noted previously, syntactic analysis abstracts away from issues of content, and this is confirmed by the relatively more successful addressing of the grammatical hypotheses. We now note the key findings for each trait in turn; more extensive information can be found by consulting Table 8.1, or the relevant chapter, directly.

Predictions that Extraverts use language suggesting positive affect, make fewer references to problems, and say more were confirmed by our analysis. However we found mixed evidence for the sociable Extravert referring more to other people, and unexpectedly found that they don't tend to use language associated with actions or hy-

phens to punctuate their writing. Overall, for Extraversion the most successful analysis appeared to be the tokenised and lemmatised corpus comparison.

For Neuroticism, we note the strong preference for self-reference, and we reject the theory hypothesis that they use more positive affect words in favour of that based on the LIWC results which predicted the converse. We note also the success of the semantic tag analysis in addressing these hypotheses, although observe that our analysis of their use of intensifying language appeared inconclusive.

Our hypotheses for Psychoticism which were based on theory appeared to be confirmed by the results, namely that they are more detached and refer less to themselves, and that they use language more unusually, in terms of vocabulary, and also grammatical constructions. Addressing hypotheses derived from previous LIWC results revealed that they do use more words associated with negative emotions, however, they do not appear to use more negations. Here we find that for Psychoticism both content-based and corpus comparison techniques are informative.

Perhaps the most intriguing results presented here are those derived using the semantic tagging analysis. Although this is able to identify meaning associated with language—like content analysis—it performs such analysis from a computational corpus linguistics perspective, thus enabling techniques such as corpus comparison to be used. This approach was only briefly mentioned in this thesis, however the additional findings derived using this method are illustrated in the final column of Table 8.1, and here we can see how they compare to that of the MRC and LIWC analysis, and also how they relate to what may be expected for the personality groups. In future work it would be interesting to examine the possibilities of semantic and corpus comparison analysis across larger and more varied data sets.

In terms of the perception of personality through e-mail communication, we note that, as predicted, Extraversion was the most easily detectable. This is consistent with findings from other forms of communication. However, we also note that Neuroticism was the least easily perceived via e-mail, however, the effects of the e-mail medium resulted in even lower levels of agreement that were expected on the basis of previous studies.

Extraversion													
Hypotheses	1	2	3	4	5	6	7	8	9	10	Ch3	Ch4	Ch5	Ch6
Theory	$\odot^{\oplus k,l}_{\oplus n,z}$	$\ominus^{\ominus k,l}_{\odot p,o,n,z}$	$\oplus^{\oplus k,l}_{\odot t,g}$	$\oplus^{\oplus k,l,z}$									−Quantification	
Real.	∘	$\oplus^{\oplus k}$	∘											
Fluency	∘	$\oplus^{\oplus k,l}$	$\ominus^{\ominus k}$											
Gramm.	$\oplus^{\odot k,l}_{\oplus n}$	$\odot^{\odot k,l}_{\bullet p,o}$	$\oplus^{\oplus p,o}$	$\bullet^{\bullet r}$	$\odot^{\odot p,o,q}$	$\oplus^{\oplus n}$	$\bullet^{\bullet p,o}$	$\oplus^{\oplus k}$						+Coord. conj.^{p}; [Adverb, Noun and Conjunction patterns]^{q,n}
Conv.	$\oplus^{\oplus k,l}_{\odot t,g}$	$\oplus^{\oplus k}$	$\oplus^{\oplus k}$	$\oplus^{\oplus k,l}$	∘	$\oplus^{\oplus k,l}$	$\oplus^{\odot k,l}_{\oplus z}$	$\odot^{\odot k,l}$						
LIWC	$\oplus^{\oplus k,z}$	∘	$\oplus^{\oplus k,z}$	$\oplus^{\oplus z}$	$\oplus^{\oplus k}$	∘	∘	∘	∘	∘		(−Numbers)^{t,g}; (−Sports; +Affect. proc.)^{o}; −Mean conc.^{m}		[+Linear order; +Music; +Education; +Science; +Clothes; +Information technology; −Cigarettes and drugs; −Living creatures; −Arts and crafts; −Entertainment; −Sports; −Frequency]^{z}
Percept.	$\oplus^{\oplus c}$	$\oplus^{\oplus c}$												

Neuroticism														
Hypotheses	1	2	3	4	5	6	7	8	9	10	Ch3	Ch4	Ch5	Ch6
Theory	$\oplus^{\oplus z}$	$\ominus^{\ominus z}$	$\odot^{\odot p,o,q,n}$	$\odot^{\odot p,q,n}$	$\oplus^{\oplus k,l,s}_{\odot o,n}$								−Temporal; −Quantification; −Modals	[−Coord. conj.; +Ellipsis]^{p}; [Adverb, Noun and Verb patterns]^{q,n}
Conv.	$\oplus^{\oplus r}$	$\oplus^{\oplus z}$											+Surface features	
LIWC	$\oplus^{\oplus k,l,s}_{\odot o,n}$	$\oplus^{\oplus z}$	$\oplus^{\oplus z}$	∘	∘						+Inclusive words	+Inclusive words^{s}; +Mean Conc.^{m}; +Mean Brown verb. freq.^{m}		[+Aircraft; +Relationships; +Trying; +Obligation; +Media; −Medical; −Places; −Revealing; −Religion; −Water activity; −Mental concerns]^{z}
Percept	$\oplus^{\oplus c}$	$\oplus^{\oplus c}$									The Social Past			

Psychoticism														
Hypotheses	1	2	3	4	5	6	7	8	9	10	Ch3	Ch4	Ch5	Ch6
Theory	$\oplus^{\oplus k,l}_{\odot o,n,z}$	$\odot^{\odot o,n,z}$	∘	∘	$\oplus^{\oplus m}$	$\oplus^{\oplus r}$	$\odot^{\odot p,o,n}$						+Temporal	[+Ellipsis; −Foreign word; +Interjection; +Parentheses]^{p}; [Adverb, Noun and Verb patterns]^{q,n}
Conv.	$\odot^{\odot l}$													
LIWC	$\oplus^{\oplus k,l,w}_{\odot o,n}$	∘	$\bullet^{\bullet p,o}$	$\oplus^{\oplus w}$	∘	$\ominus^{\ominus k}$	∘	∘	∘	∘		+Cog. mech.^{s}; (+Anger)^{g}; (−Motion; +Certainty; −Feeling; +Sad)^{g,l}; (+Death)^{g}; Std Dev of Kucera & Francis writ. Freq.^{m}	+Modality	[+Mental concerns; +Warfare; +Science; −Work; −Medical; −Education]^{z}
Percept.	$\oplus^{\oplus c}$	$\oplus^{\oplus c}$												

Table 8.1: Table displaying evidence for the Hypotheses

Note. ∘ indicates an hypothesis; ⊕ confirmation of hypothesis; ⊖ inverse of hypothesis; ⊙ partial evidence (direction unclear); • hypothesis tested but no evidence found. ^{w} LIWC factor analysis; ^{a} 'all' LIWC multiple regression; ^{t} topic controlled LIWC regression; ^{g} topic and genre controlled LIWC regression; ^{s} topic, genre and sparsity controlled LIWC regression; ^{m} MRC database analysis; ^{r} Type-token ratio analysis. ^{k} tokenised corpus comparison; ^{l} lemmatised corpus comparison. ^{p} PENN tag unigram analysis, ^{o} reduced tagset unigram analysis, ^{q} PENN tag n-gram analysis, ^{n} n-gram analysis (reduced tags), ^{z} semantic analysis (reduced); ^{c} perception study. Please refer to Section 8.2 for a full description of the hypotheses.

8.3 Contributions of the thesis

The main contributions of this thesis are as follows. In order of importance, we have:

1. Demonstrated that personality is projected and perceived through language in a CMC environment.
2. Explored the specific linguistic features associated with different personality dimensions.
3. Extended perception studies to asynchronous computer-mediated communication at zero-acquaintance.
4. Examined the relative contributions of different levels of linguistic analysis, using both theory- and data-driven techniques.
5. Implemented innovative methods of corpus comparison.
6. Investigated similarity ratings of perception.
7. Built and annotated a personality language corpus of e-mail communication.

We now discuss how these findings can inform models of human and computational language production.

8.4 Language processing and personality

In this thesis we have explored the relationship between language behaviour and personality. Here we discuss how this relates to a cognitive model of language production. Adapting Marr's (1982) framework for vision which consists of computational, algorithmic and implementational process levels, we note the following:

At the computational level, this thesis shows that personality systematically mediates language behaviour. Therefore, using the data derived from our analyses as the basis, it would be possible to implement a surface-level personality language generation system. The rationale for this would be that the personality parameters of the system are pre-defined—along for example, scales of Extraversion, Neuroticism and

Psychoticism—and the generation system then favours some of the (over-generated) linguistic candidates over others, on the basis of how strongly they are associated with the pre-defined personality parameters. This would inform the kind of stochastic text generation proposed by Oberlander and Brew (2000), which may be implemented through a probabilistic NLG system (Langkilde and Knight, 1998a,b), an instance-based approach (Varges, 2002), or using a hybrid symbolic-statistical system based on a CCG realiser (White and Baldridge, 2003).

These generation techniques, which we have just discussed, only take into account the surface realisation of the data. However, closer investigation of the data can be used to inform the algorithmic level of a language production model (cf. Levelt, 1989). Each of the personality dimensions appear to influence different levels of language behaviour, as revealed by our various approaches to analysis. For example, the content-based analyses appeared more informative for behaviour related to Neuroticism than Extraversion, but the reverse is true for the analyses which took into account contextual information. Psychoticism demonstrates characteristic behaviour in both content and contextual analyses. Therefore, these results suggest that Extraversion plays a larger role in the surface realisation of language, rather than particular concerns about topic; especially with regard to quickly generating constructions, linguistic behaviour which in conversational terms, aids the Extravert's bid to hold (or gain) the floor (Edelsky, 1981), and leads to their perceived greater conversational ability (Matthews and Deary, 1998). Neuroticism is more closely related to the content, or topic, of the language, and in particular concerns or issues relevant to the individual. On the basis of our results, we observe that Psychoticism influences both levels of language behaviour. Specifically, since this trait is related to characteristics such as aggression, which in turn has implications for socialisation, we propose that it will be particularly relevant to interpersonal aspects of language behaviour, such as convergence, priming or adherence to maxims of politeness or relevance (Giles, 1973; Brown and Levinson, 1987; Pickering and Branigan, 1998). Additionally Psychoticism is related to creativity, which we propose is reflected in attention to linguistic content and construction. Therefore as a result of these predispositions, the high Psychotic allocates more resources to the language production process, and additionally pays less attention to considering other

interlocutors (although we do not necessarily propose that this is a *direct* diversion of resources between processes).

The algorithmic information, which we have inferred from the results of this thesis, will therefore allow the more specific manipulation of language production at its different levels, to take personality into account. However, this still leaves the implementational level of our cognitive model unconsidered. Here we briefly sketch out the possible biological architecture which integrates the personality and language algorithms discussed above.

Recent neurocognitive work on emotion and human language processing suggests biological bases which could provide alternatives to those discussed by Eysenck (1970). In particular we note that hemispheric asymmetry underlies emotional and affective responses, which associate the left hemisphere with approach and positive affect, and the right hemisphere with withdrawal and negative affect (Davidson, 2001). Similarly hemispheric asymmetry has been proposed to influence language processing behaviours, with the left hemisphere responsible for syntactic and surface ordering features, and the right hemisphere for semantic and particular lexical processing (Beeman and Chiarello, 1998; Embick et al., 2000; Indefrey et al., 2001). Therefore, we suggest that the left hemisphere is important in Extravert language behaviour, and we suggest that the right hemisphere is important for Neurotic language behaviour. Above we note that the influence of Psychoticism upon language production is more general, and propose that this relates to interpersonal responses, based upon the trait's association with aggression and socialisation, and a focus on the language itself based on creativity. In this case we acknowledge the possible role of trait levels of stress hormones which relates back to the earlier proposals of Eysenck (1970). We leave the exploration of these hypotheses as the subject of future work.

8.5 Applications of the thesis

In the previous section we explored the theoretical implications of this work in terms of developing a model of language production which could account for personality. Here we turn to the possible technological applications of this work. In the previous

discussion, using Marr's framework, we outlined at the computational level how it would be possible to generate text which projected personality, having already specified such parameters to the system. This thesis contributes towards such a system in the following ways: Firstly, it has confirmed the value of generating language which indicates personality, since this thesis has demonstrated that personality can be projected and perceived through language; Secondly, this thesis has described language behaviour which would inform the language behaviour of such a system, whether this is in terms of specific features such as personal reference or emotion words, or in terms of higher mechanisms such as at what level a particular personality trait would influence behaviour, whether this is interpersonal orientation or the syntactic constructions used.

Potential uses for such personality language generation may be to enable user interfaces to adapt in a more sophisticated manner to their user. For example, Nass et al. (1995) indicated that users preferred interfaces whose language which projected personality (albeit manually manipulated) matched their own, and Amichai-Hamburger (2002) proposes that the internet should adapt to the user's personality. In both these cases, dynamic personality language generation would make them a reality. However, in both of these potential applications there are a number obstacles. Firstly, given the results presented here it should be possible to generate personality language which belongs to the genre of e-mail, although quite how far these features will allow the linguistic projection of personality in other genres is debatable. Secondly, to enable an interface to dynamically generate personality language in response to a user, it would require information about the personality of the user. However, if a computer user was required to complete a 50-item personality questionnaire before using a machine, then this may well undermine any potential benefits gained through its generation of personality language to match that of the user.

Another potential application and solution to the problem of not knowing a computer user's personality may be the automatic detection of personality through text. This would involve selecting the most informative and characteristic features for each personality type which have been presented in this thesis, and classifying a target text on the basis of which (proportion of) features it contained. As may be expected, this

would probably require additional data to be collected to enable the classification of texts other than e-mail. However there are additional complications: Firstly, such classification may only be successful on larger texts in order for enough of the features used for classification to be detected; Secondly, some of these features may be characteristic of more than one personality type, for example, the increased use of self reference may be a feature of high Neurotics or low Psychotics. Furthermore, if such a classification system was to be used in a naturalistic setting, as in a computer interface (rather than as a scientific or psychometric tool), there is also the issue of how a suitable text for analysis could be derived. One possible solution is the analysis of outgoing e-mails, or of written work produced by the user, although this is by no means an ideal solution.

Finally, one further application which is being actively investigated by the author is that of the 'Personality Style Checking Tool'. This system has personality parameters which are defined by the user. The tool is then used to check that the style of the text matches that which would be used by the desired personality type. In cases where the style does not match, more appropriate alternatives are suggested, in much the same way that alternative or more appropriate spellings are suggested by a spelling checker. For more information, the interested reader is directed to Gill et al. (pat pend).

8.6 Boundaries of the thesis

In the introduction of this thesis, we outlined boundaries determined to restrict the scope of the thesis. Here we revisit those boundaries in light of the findings we have presented. These are as follows:

First, the thesis has restricted the focus of investigation to language in a computer-mediated environment—specifically e-mail. Therefore, this raises questions about the generalisability of these findings to other forms or varieties of language. We have already noted that e-mail bears similarities to both writing and speech, and in particular CMC is regarded as being similar in genre to public interviews and letters, although beyond this is a matter of further investigation.

Secondly, in our study of perception, judges rate personality on the basis of an informal e-mail. Given these conditions, it was possible to perceive the author's per-

sonality through their writing. However, we acknowledge that levels of agreement between judges may differ if they are asked to judge more formal or tightly constrained e-mails or texts.

Finally, in studying personality, we have limited this to a trait approach, and in particular Eysenck's three-factor model of personality. Although it may be complicated to integrate this model with alternative—non-trait—theories of personality, this measure is highly regarded and relates to the similarly well-known five-factor model in well documented and defined ways.

8.7 Future work

In this section we outline possible future work which would firstly allow us to address some of the boundaries of the thesis discussed above. Secondly, in this section, we detail some of the questions which have been raised by this thesis, and suggest possible ways of investigating them.

We address the restrictions outlined above in terms of further generalisation studies and experimental manipulations. To investigate the generalisability of the current findings, further data collection and systematic variation of the experimental parameters is required. In the first instance, we may wish to replicate the same experiment, but instead using the five-factor model of personality. This would allow us to test the way the two personality models relate to each other and to language, and would allow the direct comparison with other studies which have used the five-factor model. Additionally, it would also be desirable to replicate our previous data collection, but using a different population. For example, in the current study we used a specific population, namely current or recent university students, most of whom were British. However it would be useful to examine how well this data represents e-mail users of a different educational level who speak a different variety of English.

Furthermore, informal e-mail may give results specific to this genre of writing, therefore, we would be able to test this by collecting texts of different formality or genre—preferably by the same authors—to examine how stable linguistic personality characteristics are across different forms of language.

Additionally, using experimental manipulations, we propose to identify the factors which lead to different levels of agreement between judges rating the personality of a text's author. One method of doing this would be to take the texts of different genres or formality collected as part of our generalisation work, and use these as targets for personality judgement. This would enable us to identify whether salience of author personality varied across these different conditions. On the basis of these results, we may be able to identify specific linguistic features which varied across text type, and relate these to levels of agreement in the judges.

To enable the more precise identification of linguistic features critical to personality judgement, we propose a more sophisticated method, namely the artificial manipulation of the target texts. To examine the importance of specific linguistic features in personality judgement, we would be able to modify the features of a text in tightly controlled conditions using a specific editing algorithm. By so doing, we would be able to make further inferences about the way personality is perceived through language, by controlling other factors in the data, such as topic, or stylistic features.

In this thesis we have touched on a variety of areas, each of which raises interesting possibilities for future research. But from a cognitive science viewpoint, perhaps the most obvious next steps relate to the computational implementation of these findings: for example, above we discuss the ways in which a natural language generation system may incorporate personality in determining the form of its output; alternatively we may want to automatically classify texts on the basis of the personality of—or projected by—the author, in much the same way that research has investigated classification by authorship or gender.

Additionally, we speculated about the form in which a biological model of personality may be incorporated with language production. This is to be regarded very much as an invitation and stimulus for future work (e.g., Gill et al., to appear), rather than a fully-fledged theory.

8.8 Final Words

To reiterate the central findings of this thesis: personality is indeed projected and perceived through language in a computer-mediated environment. This is significant because in such an environment—unlike face-to-face situations—textual communication is the sole source of available cues. This finding therefore has important implications for knowing about how we are perceived and how we project ourselves when only minimal cues are available. This is indeed something to be borne in mind when next writing an e-mail to a new acquaintance, or by extension of these findings to other situations, when writing a letter or making a job application. Additionally this better knowledge of how personality is projected through language can be used to inform natural language generation systems and also for text classification.

To end, we return to Louis Milic (Milic, 1966, pp. 79–80). As we have noted, he took personality to be related to specific individuals, rather than types of individuals, but he summarises what we have demonstrated here for personality dimensions.

> The personality of a writer is an inferential structure built upon what we know or can guess about his subjects of interest, his reasoning, his feelings, his linguistic decisions, his attitudes. [...] The greater the writer, usually the more numerous and impressive these differences, and the stronger the sense of personality conveyed. Personality may thus be thought of as the reverse of humanity: it is the identity of a human unit as an individual, not his identification with the race in general. Personality, therefore, and one of its literary reflections, style, is the combination of drives to break away from the uniformity of the human mass and to establish, by expressing, one's particular indefinable uniqueness. Today, when all the forces of society, technology and industry combine to reduce human beings to equivalent easily-handled units, a strong interest has arisen in asserting claims of individual personality. It is perhaps ironic that machines which have a causative share in human depersonalizations should be called to assist in the rescue of the individual literary style.

Appendix A

Participant Information

Subject ID	Pers. score				Age	Gender	Nationality	Student	Text word count		Hi-Lo split group	Hi-Mid-Low group	Rating text
	P	E	N	L					Past	Next			
28Aug2001_09:14:41+0100	2	8	7	4	44	F	UK	PG	251	217		MidSD	
11Jan2001_16:04:13GMT	4	8	1	0	20	M	UK	UG	377	207	LowN	LowNsd	
27Jul2001_17:03:29+0100	2	12	10	5	31	F	UK	PG	383	307	HighE, HighN		
29Aug2001_11:33:29+0100	3	7	0	4	29	M	UK	PG	230	205	LowN	LowNsd	
24Aug2001_17:52:23+0100	3	6	2	3	30	M	UK	PG	263	217	LowN	LowNsd	
17May2001_16:14:29+0100	3	6	9	4	24	M	UK	PG	144	161	HighN	HighNsd	
28Aug2001_15:13:35+0100	3	2	8	2	24	M	UK	PG	468	358	LowE	LowEsd	
27Aug2001_15:20:08+0100	3	8	3	6	23	M	Aus	PG	201	140		MidSD	N5
24Aug2001_16:12:27+0100	3	12	1	3	23	M	Ire	PG	441	252	HighE, LowN		
20Aug2001_17:39:02+0100	2	7	2	5	24	F	UK	PG	455	331	LowN	LowNsd	
14Dec2000_14:37:24GMT	2	6	5	3	28	F	UK	PG	99	58		MidSD	E6
27Jul2001_02:25:28+0100	2	8	3	8	24	F	Can	UG	177	120		MidSD	
28Feb2001_13:28:56GMT	2	8	1	3	21	F	UK	UG	231	140	LowN	LowNsd	N1
15Jan2001_20:13:44GMT	2	8	1	2	23	F	Aus	PG	376	405	LowN	LowNsd	
20Jul2001_08:57:13+0100	2	9	5	6	22	F	UK	UG	353	374		MidSD	P4
15Dec2000_05:52:09GMT	2	9	6	0	22	F	UK	UG	335	287		MidSD	
25Jul2001_17:21:17+0100	2	10	5	2	23	F	UK	PG	206	273		MidSD	
11Jan2001_09:31:10GMT	2	10	4	5	21	F	UK	UG	194	125		MidSD	
10Mar2001_13:58:13GMT	2	10	7	2	21	F	UK	UG	320	159		MidSD	
15Jan2001_14:49:22GMT	2	10	11	2	22	F	UK	PG	343	369	HighN	HighNsd	
5Jan2001_13:00:37GMT	2	12	5	1	24	F	UK	PG	335	280	HighE	HighEsd	E2
3Jun2001_22:25:29+0100	2	12	4	4	24	F	UK	UG	209	119	HighE	HighEsd	
22Dec2000_15:40:12GMT	2	12	3	1	22	F	UK	UG	238	191	HighE	HighEsd	
18Jul2001_19:10:15+0100	2	12	0	1	27	F	NZ	PG	516	311	HighE, LowN		
9Jan2001_16:12:48GMT	2	12	10	4	24	F	UK	PG	403	320	HighE, HighN		
28Aug2001_11:06:26+0100	2	4	7	2	32	M	UK	PG	110	71	LowE	LowEsd	
21Jul2001_19:57:10+0100	2	4	10	4	37	M	UK	PG	403	306	HighN, LowE		
27Jul2001_15:28:49+0100	2	12	7	7	33	M	UK	PG	366	306	HighE	HighEsd	
11Jan2001_21:17:00GMT	3	12	8	2	20	M	UK	UG	186	135	HighE	HighEsd	
30Jul2001_22:22:52+0100	1	4	12	7	31	F	UK	PG	146	174	LowP, LowE, HighN		
9Jan2001_15:05:11GMT	1	12	9	1	38	F	UK	PG	285	284	LowP, HighE, HighN		
11Aug2001_12:05:29+0100	1	5	10	1	25	F	UK	PG	277	273	HighN, LowP		
16Jan2001_11:57:29GMT	1	3	11	2	21	F	UK	UG	179	118	LowP, LowE, HighN		
15Jan2001_12:05:41GMT	1	9	1	8	24	F	UK	PG	547	387	LowP, LowN		
5Jan2001_14:26:30GMT	1	10	3	2	25	F	UK	PG	491	419	LowP	LowPsd	
24Jan2001_15:24:33GMT	2	6	6	2	20	M	UK	UG	157	127		MidSD	
28Jul2001_20:49:57+0100	2	0	9	2	18	M	UK	UG	558	405	LowE, HighN		
30Jan2001_12:36:20GMT	1	8	8	4	20	F	UK	UG	165	113	LowP	LowPsd	P1
1Feb2001_16:22:26GMT	1	7	10	5	20	F	UK	UG	147	104	LowP, HighN		
13Dec2000_22:07:47GMT	1	6	2	7	20	F	UK	UG	229	184	LowP, LowN		
18Jan2001_11:11:44GMT	1	11	7	0	19	F	UK	UG	246	192	LowP	LowPsd	
28Aug2001_18:42:23+0100	1	6	5	3	25	M	UK	UG	242	285	LowP	LowPsd	
10Jan2001_15:44:04GMT	1	5	7	3	21	M	UK	UG	269	203	LowP	LowPsd	
14Dec2000_10:24:30GMT	1	3	7	8	24	M	UK	PG	247	180	LowP, LowE		
28Aug2001_13:04:55+0100	1	1	4	9	24	M	UK	PG	156	93	Low, PLowE		
1Mar2001_11:53:56GMT	0	6	11	0	24	F	UK	PG	344	363	LowP, HighN		
29Jan2001_15:39:13GMT	0	5	8	2	23	F	UK	UG	336	259	LowP	LowPsd	
16Jul2001_13:08:24+0100	0	3	6	6	26	F	UK	PG	235	165	LowP, LowE		
23Jan2001_13:11:53GMT	0	1	7	4	23	F	UK	UG	217	167	LowP, LowE		
26Jul2001_22:09:59+0100	0	8	6	6	25	F	Can	UG	343	201	LowP	LowPsd	P2
30Jan2001_18:14:20GMT	0	9	3	4	18	F	UK	UG	208	17	LowP	LowPsd	
11Jan2001_15:07:02GMT	0	9	4	4	21	M	UK	UG	239	71	LowP	LowPsd	
11Jan2001_17:16:07GMT	0	4	10	4	19.5	M	UK	UG	117	95	LowP, LowE, HighN		

Table A.1: Demographic information for authors of e-mail corpus

Subject ID	Pers. score				Age	Gender	Nationality	Student	Text word count		Hi-Lo split group	Hi-Mid-Low group	Rating text
	P	E	N	L					Past	Next			
30Jul2001_03:55:52+0100	3	7	10	1	31	M	Can	UG	370	194	HighN	HighNsd	N4
10Jan2001_10:14:27GMT	3	12	0	7	32	M	UK	PG	171	231	HighE, LowN		
15Jan2001_15:52:37GMT	3	4	12	3	20	F	UK	UG	432	324	LowE, HighN		
26Jul2001_18:31:54+0100	3	11	8	3	19	F	UK	UG	320	253			E3
23Jul2001_10:16:39+0100	3	12	4	1	19	F	UK	UG	151	138	HighE	HighEsd	
11Jan2001_10:38:51GMT	4	5	5	2	21	M	UK	UG	107	91		MidSD	
18Jan2001_01:07:23GMT	4	9	1	4	22	M	UK	UG	384	288	LowN	LowNsd	N2
18Jul2001_11:44:11+0100	4	9	9	1	23	M	UK	PG	281	194	HighN	HighNsd	
10Jan2001_11:33:48GMT	4	12	0	6	22	M	UK	UG	254	86	HighELowN		
1Aug2001_20:48:23+0100	3	7	5	8	23	F	UK	UG	185	126		MidSD	
15Feb2001_18:08:31GMT	3	7	7	2	21	F	UK	UG	288	244		MidSD	N6
26Aug2001_12:55:51+0100	3	5	5	1	29	F	UK	PG	73	59		MidSD	E1
29Jul2001_19:55:29+0100	3	4	0	6	24	F	Can	UG	222	174	LowE, LowN		
12Jan2001_10:41:26GMT	3	3	7	9	24	F	UK	PG	348	202	LowE	LowEsd	E4
30Aug2001_11:52:23+0100	3	8	6	4	24	F	UK	PG	373	234		MidSD	
9Jan2001_13:48:02GMT	3	9	5	3	27	F	UK	PG	167	144		MidSD	
14Dec2000_13:41:24GMT	3	10	1	1	24	F	UK	UG	430	323	LowN	LowNsd	
17Dec2000_19:10:40GMT	3	10	7	2	24	F	UK	PG	393	291		MidSD	E5
5Jan2001_17:24:11GMT	3	11	5	4	24	F	UK	PG	381	267		MidSD	
29Jan2001_18:28:45GMT	3	11	2	4	21	F	UK	UG	480	561	LowN	LowNsd	
5Jan2001_13:39:38GMT	3	12	4	3	22	F	UK	PG	267	214	HighE	HighEsd	
6Aug2001_16:00:21+0100	3	12	4	6	23	F	UK	PG	798	495	HighE	HighEsd	
27Aug2001_13:52:51+0100	3	12	3	7	30	F	UK	UG	285	288	HighE	HighEsd	
9Jan2001_20:03:52GMT	3	12	5	6	33	F	UK	PG	377	281	HighE	HighEsd	
30Jan2001_10:51:34GMT	4	6	4	5	19	F	UK	UG	153	128		MidSD	
14Dec2000_10:33:10GMT	5	2	8	2	24	M	UK	UG	135	94	HighP, LowE		
15Jan2001_18:31:17GMT	5	8	1	1	23	M	UK	UG	259	241	HighP, LowN		
18Jan2001_22:20:47GMT	5	10	6	7	24	M	UK	UG	241	211	HighP	HighPsd	
14Dec2000_11:29:43GMT	4	12	4	2	19	F	UK	UG	173	134	HighE	HighEsd	
29Aug2001_12:57:40+0100	4	2	12	5	36	M	UK	PG	346	232	LowE, HighN		
19Jan2001_16:46:02GMT	4	7	4	2	21	F	UK	UG	336	188		MidSD	
11Jan2001_16:23:58GMT	4	6	9	5	21	F	UK	UG	181	145	HighN	HighNsd	N3
9Jan2001_16:52:53GMT	4	0	7	4	26	F	UK	PG	361	309	LowE	LowEsd	
26Jul2001_22:44:05+0100	4	8	6	1	22	F	Can	UG	292	157		MidSD	P6
27Aug2001_18:39:05+0100	4	8	9	1	23	F	UK	PG	483	502		HighNsd	
10Jan2001_19:04:43GMT	4	9	4	5	22	F	UK	UG	154	94		MidSD	
15Jan2001_15:23:06GMT	4	11	5	3	30	F	UK	PG	235	191		MidSD	
27Aug2001_19:34:05+0100	6	7	6	2	27	M	UK	UG	438	394	HighP	HighPsd	
20Jul2001_16:37:47+0100	6	5	7	6	27	M	UK	UG	161	115	HighP	HighPsd	
16Jan2001_21:44:17GMT	6	9	0	1	25	M	UK	PG	185	154	HighP, LowN		
30Jan2001_20:49:18GMT	6	12	7	1	21	M	UK	PG	643	366	HighP, HighE		
11Jan2001_14:53:49GMT	5	10	3	3	31	M	UK	PG	127	108	HighP	HighPsd	P5
24May2001_15:33:45+0100	5	6	8	4	21	F	US	PG	201	145	HighP	HighPsd	
14Dec2000_15:57:53GMT	5	9	10	2	24	F	UK	PG	399	309	HighP, HighN		
19Jan2001_12:56:07GMT	5	10	10	3	22	F	UK	PG	202	365	HighP, HighN		
17Jan2001_14:21:36GMT	5	11	4	1	21	F	UK	UG	187	146	HighP	HighPsd	
14Jan2001_10:41:50GMT	5	12	2	0	23	F	UK	UG	191	121	HighP, HighE, LowN		
14Dec2000_10:51:07GMT	7	5	7	3	24	M	UK	PG	215	198	HighP	HighPsd	
10Feb2001_15:46:44GMT	6	7	5	5	32	M	Can	PG	272	192	HighP	HighPsd	P3
27Aug2001_15:03:21+0100	6	5	3	4	27	F	Ire	PG	259	181	HighP	HighPsd	
16Jan2001_12:52:34GMT	6	11	2	3	21	F	UK	UG	248	170	HighP, LowN		
16Jan2001_15:00:44GMT	9	12	0	1	21	M	UK	UG	307	292	HighP, HighE, LowN		

Table A.2: Demographic information for authors of e-mail corpus (cont.)

Appendix B

Key to Syntactic and Semantic Annotation

POS Tag	Description	Example
CC	coordinating conjunction	and
CD	cardinal number	1, third
DT	determiner	the
EX	existential there	*there* is
FW	foreign word	d'hoevre
IN	preposition/subordinating conjunction	in, of, like
JJ	adjective	green
JJR	adjective, comparative	greener
JJS	adjective, superlative	greenest
LS	list marker	1.
MD	modal	could, will
NN	noun, singular or mass	table
NNS	noun plural	tables
NNP	proper noun, singular	John
NNPS	proper noun, plural	Vikings
PDT	predeterminer	*both* the boys
POS	possessive ending	friend*'s*
PRP	personal pronoun	I, he, it
PRP$	possessive pronoun	my, his
RB	adverb	however, usually, good
RBR	adverb, comparative	better
RBS	adverb, superlative	best
RP	particle	give *up*
TO	to	*to* go, *to* him
UH	interjection	uhhuhhuhh
VB	verb, base form	take
VBD	verb, past tense	took
VBG	verb, gerund/present participle	taking
VBN	verb, past participle	taken
VBP	verb, sing. present, non-3d	take
VBZ	verb, 3rd person sing. present	takes
WDT	wh-determiner	which
WP	wh-pronoun	who, what
WP$	possessive wh-pronoun	whose
WRB	wh-abverb	where, when

Table B.1: Key to PENN Treebank POS tagset (Modified from Marcus et al., 1994)

Semantic Tag	Description
A1	GENERAL AND ABSTRACT TERMS
A1.1.1	General actions, making etc.
A1.1.2	Damaging and destroying
A1.2	Suitability
A1.3	Caution
A1.4	Chance, luck
A1.5	Use
A1.5.1	Using
A1.5.2	Usefulness
A1.6	Physical/mental
A1.7	Constraint
A1.8	Inclusion/Exclusion
A1.9	Avoiding
A2	Affect
A2.1	Affect:- Modify, change
A2.2	Affect:- Cause/Connected
A3	Being
A4	Classification
A4.1	Generally kinds, groups, examples
A4.2	Particular/general; detail
A5	Evaluation
A5.1	Evaluation:- Good/bad
A5.2	Evaluation:- True/false
A5.3	Evaluation:- Accuracy
A5.4	Evaluation:- Authenticity
A6	Comparing
A6.1	Comparing:- Similar/different
A6.2	Comparing:- Usual/unusual
A6.3	Comparing:- Variety
A7	Definite (+ modals)
A8	Seem
A9	Getting and giving; possession
A10	Open/closed; Hiding/Hidden; Finding; Showing
A11	Importance
A11.1	Importance: Important
A11.2	Importance: Noticeability
A12	Easy/difficult
A13	Degree
A13.1	Degree: Non-specific
A13.2	Degree: Maximizers
A13.3	Degree: Boosters
A13.4	Degree: Approximators
A13.5	Degree: Compromisers
A13.6	Degree: Diminishers
A13.7	Degree: Minimizers
A14	Exclusivizers/particularizers
A15	Safety/Danger
B1	Anatomy and physiology
B2	Health and disease
B3	Medicines and medical treatment
B4	Cleaning and personal care
B5	Clothes and personal belongings
C1	Arts and crafts
E1	EMOTIONAL ACTIONS, STATES AND PROCESSES General
E2	Liking
E3	Calm/Violent/Angry
E4	Happy/sad
E4.1	Happy/sad: Happy
E4.2	Happy/sad: Contentment
E5	Fear/bravery/shock
E6	Worry, concern, confident
F1	Food
F2	Drinks
F3	Cigarettes and drugs
F4	Farming & Horticulture
G1	Government, Politics and elections
G1.1	Government etc.
G1.2	Politics
G2	Crime, law and order
G2.1	Crime, law and order: Law and order
G2.2	General ethics
G3	Warfare, defence and the army; weapons

Table B.2: Key to USAS Semantic Tags (Modified from Archer et al., 2002)

Semantic Tag	Description
H1	Architecture and kinds of houses and buildings
H2	Parts of buildings
H3	Areas around or near houses
H4	Residence
H5	Furniture and household fittings
I1	Money generally
I1.1	Money: Affluence
I1.2	Money: Debts
I1.3	Money: Price
I2	Business
I2.1	Business: Generally
I2.2	Business: Selling
I3	Work and employment
I3.1	Work and employment: Generally
I3.2	Work and employment: Professionalism
I4	Industry
K1	Entertainment generally
K2	Music and related activities
K3	Recorded sound etc.
K4	Drama, the theatre and showbusiness
K5	Sports and games generally
K5.1	Sports
K5.2	Games
K6	Childrens games and toys
L1	Life and living things
L2	Living creatures generally
L3	Plants
M1	Moving, coming and going
M2	Putting, taking, pulling, pushing, transporting etc.
M3	Vehicles and transport on land
M4	Shipping, swimming etc.
M5	Aircraft and flying
M6	Location and direction
M7	Places
M8	Remaining/stationary
N1	Numbers
N2	Mathematics
N3	Measurement
N3.1	Measurement: General
N3.2	Measurement: Size
N3.3	Measurement: Distance
N3.4	Measurement: Volume
N3.5	Measurement: Weight
N3.6	Measurement: Area
N3.7	Measurement: Length & height
N3.8	Measurement: Speed
N4	Linear order
N5	Quantities
N5.1	Entirety; maximum
N5.2	Exceeding; waste
N6	Frequency etc.
O1	Substances and materials generally
O1.1	Substances and materials generally: Solid
O1.2	Substances and materials generally: Liquid
O1.3	Substances and materials generally: Gas
O2	Objects generally
O3	Electricity and electrical equipment
O4	Physical attributes
O4.1	General appearance and physical properties
O4.2	Judgement of appearance (pretty etc.)
O4.3	Colour and colour patterns
O4.4	Shape
O4.5	Texture
O4.6	Temperature
P1	Education in general
Q1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.1	LINGUISTIC ACTIONS, STATES AND PROCESSES; COMMUNICATION
Q1.2	Paper documents and writing
Q1.3	Telecommunications
Q2	Speech acts
Q2.1	Speech etc:- Communicative
Q2.2	Speech acts
Q3	Language, speech and grammar
Q4	The Media
Q4.1	The Media:- Books
Q4.2	The Media:- Newspapers etc.
Q4.3	The Media:- TV, Radio and Cinema

Table B.3: Key to USAS Semantic Tags (cont.)

Semantic Tag	Description
S1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.1	SOCIAL ACTIONS, STATES AND PROCESSES
S1.1.2	Reciprocity
S1.1.3	Participation
S1.1.4	Deserve etc.
S1.2	Personality traits
S1.2.1	Approachability and Friendliness
S1.2.2	Avarice
S1.2.3	Egoism
S1.2.4	Politeness
S1.2.5	Toughness; strong/weak
S1.2.6	Sensible
S2	People
S2.1	People:- Female
S2.2	People:- Male
S3	Relationship
S3.1	Relationship: General
S3.2	Relationship: Intimate/sexual
S4	Kin
S5	Groups and affiliation
S6	Obligation and necessity
S7	Power relationship
S7.1	Power, organizing
S7.2	Respect
S7.3	Competition
S7.4	Permission
S8	Helping/hindering
S9	Religion and the supernatural
T1	Time
T1.1	Time: General
T1.1.1	Time: General: Past
T1.1.2	Time: General: Present; simultaneous
T1.1.3	Time: General: Future
T1.2	Time: Momentary
T1.3	Time: Period
T2	Time: Beginning and ending
T3	Time: Old, new and young; age
T4	Time: Early/late
W1	The universe
W2	Light
W3	Geographical terms
W4	Weather
W5	Green issues
X1	PSYCHOLOGICAL ACTIONS, STATES AND PROCESSES
X2	Mental actions and processes
X2.1	Thought, belief
X2.2	Knowledge
X2.3	Learn
X2.4	Investigate, examine, test, search
X2.5	Understand
X2.6	Expect
X3	Sensory
X3.1	Sensory:- Taste
X3.2	Sensory:- Sound
X3.3	Sensory:- Touch
X3.4	Sensory:- Sight
X3.5	Sensory:- Smell
X4	Mental object
X4.1	Mental object:- Conceptual object
X4.2	Mental object:- Means, method
X5	Attention
X5.1	Attention
X5.2	Interest/boredom/excited/energetic
X6	Deciding
X7	Wanting; planning; choosing
X8	Trying
X9	Ability
X9.1	Ability:- Ability, intelligence
X9.2	Ability:- Success and failure
Y1	Science and technology in general
Y2	Information technology and computing
Z0	Unmatched proper noun
Z1	Personal names
Z2	Geographical names
Z3	Other proper names
Z4	Discourse Bin
Z5	Grammatical bin
Z6	Negative
Z7	If
Z8	Pronouns etc.
Z9	Trash can
Z99	Unmatched

Table B.4: Key to Semantic Tags (cont.)

Appendix C

Previous Results

Dictionary	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)
Present tense	.593		.596	
Words > 6 letters	−.683			
First-person Sing.	.823			
Insight				.627
Articles	−.765			
Exclusive		.674		
Negations		.579		
Tentative		.644		
Discrepancies	.485	.427		
Inclusive		−.463		
Past tense			.856	
Social			.425	
Positive emotion			−.469	
Negative emotion				−.443
Causation				.598

Note. Only loadings of .20 or above are shown. $N = 838$.

Table C.1: Rotated Factor Loadings for Exploratory Analysis of LIWC Dictionaries Ordered to match the current study (reproduced from Pennebaker and King, 1999, p. 1303).

	Five-Factor Dimension				
LIWC factor	Openn.	Agreeab.	Conscient.	Extrav.	Neurot.
Immediacy	**−.16****	**.07***	**−.02**	**.04**	**.10***
Present tense	−.15**	.04	.00	.01	−.06
Words > 6 letters	.16**	−.03	.06	−.04	−.03
First-person Sing.	−.13**	.07**	.01	.04	.13**
Articles	.13**	−.15**	−.04	−.09*	−.09*
Discrepancies	−.01	−.02	−.07*	−.03	.05
Making Dist.	**.06**	**−.05**	**−.13****	**−.14****	**.05**
Exclusive	.10*	−.06	−.08*	−.08*	.00
Negations	.00	−.04	−.15**	−.12**	.05
Tentative	.11**	−.02	−.06	−.14**	.06
Inclusive	.01	.03	.06	.07*	−.01
The Social Past	**.08***	**−.02**	**−.04**	**.00**	**.04**
Past tense	−.03	.06	−.06	.04	.03
Social	.02	.00	.02	.12**	−.01
Positive emotion	−.06	.07*	.07*	.15**	−.13**
Rationalisation	**−.03**	**.07**	**.04**	**.02**	**−.06**
Causation	−.08*	.00	−.07*	−.08*	.03
Negative emotion	.05	−.07*	−.15**	−.08*	.16**
Insight	.07*	.05	−.01	−.02	.03

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 constituent LIWC variables for the LIWC factors has been altered to match that of the present study.
$*p < .05$, $**p < .01$, two tailed.

Table C.2: LIWC Factors and Simple Correlations with Five-Factor Scores (reproduced from Pennebaker and King, 1999, p. 1307)

Appendix D

Published Papers

D.1 Gill and Oberlander (2002)

Taking Care of the Linguistic Features of Extraversion

Alastair J. Gill (agill@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We study how Extraversion or Introversion influences people's language production. A corpus of e-mail texts was gathered from individuals categorised via Eysenck's EPQ-R personality test. One experiment analysed the corpus using existing content analysis tools, and found relatively weak effects of Extraversion. A second experiment used more sensitive bigram-based techniques from statistical natural language processing to replicate earlier findings, and uncover novel patterns of behaviour.

Introduction

Casual acquaintance with Extraverts[1] and Introverts suggests that the former talk a lot more than the latter. But apart from this intuitive difference, how does this personality dimension influence an individual's language production? Before addressing this question, we need to clarify what we mean by Extraversion, and its relevance to cognitive science.

A typical Extravert tends to be sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert is quiet, retiring, reserved, plans ahead, and dislikes excitement (Eysenck and Eysenck, 1991).

The personality trait of Extraversion—and the complementary Introversion—is one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993). Beyond it, there is greater controversy.

For instance, Eysenck's EPQ-R personality test reflects a personality model which incorporates just two further dimensions: Neuroticism, which is mainly characterised by susceptibility to anxiety; and Psychoticism, which is more complicated, but generally related to aggression and individuality. By contrast, the NEO-PI-R model incorporates five factors (Costa and McCrae, 1992). As well as Extraversion and Neuroticism, there are Conscientiousness, Agreeableness and Openness. It is generally agreed that these relate to Psychoticism, but exactly how is still the subject of debate (cf. Matthews and Deary, 1998).

Extraversion, and its linguistic consequences—if there are any—is relevant to cognitive research for at least two reasons. First, there is considerable evidence that this personality dimension is related to preferred learning styles and educational achievement, via speed of exam completion, memory retrieval and recall tasks, creativity, mathematical ability, self monitoring and communication ability (Jonassen and Grabowski, 1993). Secondly, there is evidence that computer users attribute personality to interfaces, and respond to it in robust ways (eg. Nass, Moon, Fogg, and Reeves, 1995; Isbister and Nass, 2000). Even in a text-only environment, Extraverts apparently prefer interfaces presenting information using language associated with Extravert traits; Introverts prefer Introverted interfaces. An interface with a matching personality is judged more positively, and rated as more attractive, credible and informative (Nass *et al.*, 1995).

So the personality dimension has some validity, and appears relevant to the diagnosis and projection of personality in human-computer communication, and in computer-based learning. But how does Extraversion influence an individual's language production? In addressing this question, we first outline some hypotheses from the literature, before describing our collection of a controlled corpus of language, and our analysis of it. We then report the results—some unsurprising, others unexpected—and discuss some of their implications.

Previous hypotheses

Work on textual personality within the "Computers Are Social Actors" paradigm has taken the expressive hallmarks of Extraversion or *dominance* (one facet of the dimension) to be confidence, as shown by an avoidance of hedge-expressions such as *perhaps* and *maybe* (Nass *et al.*, 1995), and is related to the empirical work of Bradac and Mulac (1984) on perceptions of powerful and powerless speech.

From an intuitive perspective, Extraverts are described as individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts mo-

[1]The spelling of *Extravert* follows Eysenck, because this paper employs his EPQ-R as the measure of personality, but this does not represent a commitment to a specifically Eysenckian theory of personality.

nopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (cf. Carment, Miles, and Cervin, 1965). With reference primarily to speech, Furnham (1990) has proposed that Extravert language is less formal, has a more restricted code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely (see also Dewaele and Furnham, 1999, for a review of speech and writing studies).

Text analysis approaches have found that transcribed texts rated as belonging to the *warm* facet of Extraversion used fewer negative emotion words and unique words, and more present tense verbs, with *dominant* texts using fewer unique words, positive emotion words and self referents (Berry, Pennebaker, Mueller, and Hiller, 1997). Finally, study of the texts *written* by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclusive words, while using more social and positive emotion words (Pennebaker and King, 1999).

Data Collection

The approach to data collection follows Pennebaker and King (1999). Written texts were collected from 105 University students or recent graduates (37 males, 68 females; mean age = 24.3 years; SD = 4.6; all native English speakers). An introductory e-mail explained the experiment, and pointed subjects to the relevant web-page. After the completion of an online demographic questionnaire and a version of the Eysenck Personality Questionnaire (Revised short form; Eysenck, Eysenck, and Barrett, 1985) (mean score for E = 7.91, SD = 3.25; normative score = 7.42 (male), 7.60 (female)), subjects were asked to compose two e-mails *to a good friend whom they hadn't seen for quite some time*, the style of which is considered to be close to oral communication (Bälter, 1998). One message concerned their activities in the past week; the other discussed their plans for the next week. Subjects were advised to spend around ten minutes per message, composed online and submitted using an HTML form. It was highlighted that responses would be treated in confidence and that subjects could remain anonymous. No payment was made for participation, and integrity of responses was monitored by reading through the transcripts. One additional submission was rejected as not being serious; the resulting corpus contained 210 texts and 65,000 words.

Experiment 1: Dictionary techniques

LIWC and MRC Methods

LIWC Each respondent's texts were individually processed using the LIWC text analysis program (Pennebaker and Francis, 1999). Items were selected for principle components analysis using the same criteria as Pennebaker and King (1999), namely reliability, topic independence, independence from other variables, and a mean minimum usage of 1%. The validity of the current data was shown using varimax rotation to derive four factors which essentially replicate their prior findings. There was minor variation in some factor loadings, which we attribute to differences in the writing tasks. See Gill and Oberlander (prep) for a fuller discussion.

Table 1: Summary of E Score and LIWC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Numbers	−.21		
	Word Count	.20	**.08**	**.0144**

By correlating their resultant LIWC factors with personality dimensions, Pennebaker and King's results suggest broad style preferences for Extraverts. But this does not identify the relative importance of their categories for identifying text as Extravert.

Thus, to identify which LIWC variables best help identify an author's personality, stepwise linear multiple regression was performed. The variables entered were those which showed at least a small correlation with the personality type—with a significance of $p < .1$—and which satisfied the criteria for inclusion in the previous principle components analysis. However, since requiring variables to have a mean usage of 1% per essay for inclusion in the analysis did not leave any LIWC variables in the regression equation for Extraversion, this criterion was ignored for the results presented below. (Interestingly, by contrast, even with the application of this criterion, Psychoticism and Neuroticism both had several strongly significant LIWC predictor variables.)

MRC In addition to the LIWC-based tests, multiple regression analysis was also performed on psycholinguistic properties of the texts, derived from the MRC Psycholinguistic Database (Coltheart, 1981). Texts were first tagged for Parts of Speech,[2] and each word-POS pair was then looked up in the database. If the word and POS tag matched a pair in the database, psycholinguistic data was returned for that word. When all the words in the text had been processed, mean scores were calculated for categories such as verbal frequency, written frequency, concreteness, age of acquisition, along with additional global information, such as the percentage of a text's words which were captured by the database. As with the LIWC regression, variables showing a correlation with the personality type with a significance of $p < .1$ were entered in to the equation.

[2]Using the MXPOST tagger (Ratnaparkhi, 1996).

Table 2: Summary of E Score and MRC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Mean Concreteness	$-.21$	**.05**	**.0278**

Results

The multiple regression analysis of the LIWC variables (Table 1) shows that a greater overall word count for a text ($\beta = .20$), and the occurrence of fewer references to numbers within that text ($\beta = -.21$), indicate Extraversion ($p < .05$). So, Extraverts *do* appear to type more than Introverts, mirroring earlier results on speech (Carment *et al.*, 1965), with the avoidance of numbers embodying a 'looser', less precise use of language (Furnham, 1990). However, the variance accounted for by these variables is relatively low at 8%. In comparable analyses, both Psychoticism and Neuroticism regression equations explain variance greater than 10%.

Similarly, with the MRC Psycholinguistic analysis (Table 2), only the novel finding of a general lowering of a text's concreteness of vocabulary ($\beta = -.21$, $p < .05$) was seen to explain 5% of variance in Extraversion. Again, equations for Psychoticism and Neuroticism explained more than 10% of variance.

Discussion

In both of the dictionary-based analyses of the texts, rather few features appeared to distinguish Extravert/Introvert texts, especially when compared to the numerous LIWC and MRC features which associated with Psychoticism and Neuroticism traits.

How could this be? At least two explanations are possible. First, the LIWC dictionary is a subjectively constructed analysis tool. It is based on judgements by health psychologists of texts written by distressed individuals for therapeutic purposes (Pennebaker and Francis, 1999). For its original purposes, this is a strength; but it also imposes a *top down* limitation on LIWC's functioning. Given this therapeutic origin, it is tempting to suggest that the linguistic features associated with the personality traits of Psychoticism and Neuroticism were more important or relevant to the distressed individuals producing the texts—and that is why these features are better represented in LIWC's dictionary.

The MRC database is also fitted to its specific purposes—for example, matching psycholinguistic stimuli—but this again imposes constraints which might prove artificial when it is applied to a different area of investigation.

Secondly, both dictionaries necessarily operate using strings corresponding to individual words, subsequently classifying them in a predefined way. Neither takes into account the context of a word. Thus it may be that for Psychoticism and Neuroticism the choice of word, or some property of the word is informative—but for Extraverts, it may be that word order or collocations are more relevant.

Experiment 2: NLP techniques

Therefore, we recruit *bottom up* statistical text analysis techniques from corpus linguistics. Specifically, bigram analysis calculates the probability of pairs of adjacent terms, or bigrams, occurring together in that order in a given text. To determine the significance of a bigram's occurrence, a statistic—log likelihood—is calculated, taking into account all the other instances of each element in the bigram pair, and the other words with which they appear.

Since bigrams can be used to calculate the probabilistic space in which language occurs, they have been put to a variety of uses (Collins, 1996; Pedersen, 2001). However, this study uses them simply as an advancement on the classified unigram (that is, single-word) analysis in Experiment 1. Because bigrams contain information about the interconnection and dependencies of words, this second analysis retains some of the contextual information of language use. Equally importantly, since bigrams are not classified subjectively, they provide a form of analysis that is bottom-up, rather than top-down.

Method

The original corpus of texts was divided by degree of Extraversion by selecting respondents whose E score was greater or less than 1 s.d. of the mean (cf. Dewaele and Pavlenko, 2002), with the 21 High Extravert authors scoring more than 11, and the 17 Low Extravert authors scoring less than 5.

Bigrams were calculated for the resulting Extravert and Introvert subcorpora; the former contained over 12,000 words; the latter around 8,000. Bigram profiles were generated for each corpus and their co-occurence significance in the current texts ranked by log-likelihood statistic ($-2 \log \lambda$),[3] since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993). Rankings for each group are based on the top 50 bigrams with frequency of $N \geq 2$, and a significance of $p < .001$. Relative frequency ratios (Damerau, 1993) were then calculated for bigrams that were common to both the subcorpora, and a Spearman Rank correlation was also performed on these bigrams.

Results

Spearman Rank Correlation

The correlation coefficient score of .53 indicates that Extravert and Introvert use of the shared bigrams is significantly correlated at the $p < .005$ (one-tailed, N=28) level, and they are therefore not distinct.

[3]Ted Pederson's bigram software is available from: http://www.d.umn.edu/~tpederse/code.html.

Table 3: Shared Extravert and Introvert bigrams.

Bigram	Extr Cnt	Intr Cnt	Extr Ratio	Intr Ratio	Rel.F Ratio
looking forward	15	4	0.0011	0.0005	2.49
it was	46	22	0.0034	0.0025	1.39
next week	24	12	0.0018	0.0013	1.33
a bit	29	15	0.0022	0.0017	1.28
up with	19	10	0.0014	0.0011	1.26
! !	45	24	0.0033	0.0027	1.24
will be	24	13	0.0018	0.0015	1.22
i was	33	18	0.0025	0.0020	1.22
at the	27	16	0.0020	0.0018	1.12
to see	32	19	0.0024	0.0021	1.12
which is	15	9	0.0011	0.0010	1.11
for a	34	21	0.0025	0.0024	1.07
i have	44	29	0.0033	0.0032	1.01
to get	34	23	0.0025	0.0026	0.98
. i	99	69	0.0074	0.0077	0.95
on friday	11	8	0.0008	0.0009	0.91
, and	48	36	0.0036	0.0040	0.88
and then	23	19	0.0017	0.0021	0.80
in the	41	34	0.0031	0.0038	0.80
apart from	6	5	0.0005	0.0006	0.80
i am	33	28	0.0025	0.0031	0.78
i think	16	14	0.0012	0.0016	0.76
, but	35	31	0.0026	0.0035	0.75
a lot	10	9	0.0007	0.0010	0.74
going to	36	33	0.0027	0.0037	0.72
a few	12	11	0.0009	0.0012	0.72
to do	23	23	0.0017	0.0026	0.66
i've been	9	12	0.0007	0.0013	0.50

However, further analysis showed Extraverts to be more distinguishable from Ambiverts or Introverts.[4]

Extraverts versus Introverts

The results of the bigram analysis include: bigrams which occurred in both the Extravert and Introvert corpora (Table 3); bigrams which were found uniquely in the Extravert corpus (Table 4); and those found only in the Introvert corpus (Table 5). The shared bigrams are ordered by their relative frequency, with the highest ratios above 1.0 showing the strongest association with Extravert authors, and the smallest ratios less than 1.0 indicating a preference on the part of more Introverted authors (the breakpoint has been indicated by a separating rule). Features which are unique to each subcorpus group can be considered the most distinctive of authorial personality. For current purposes, we divide the features into eight groupings.

Surface Realisation Features These gross features are perhaps the most intuitive in their representation of the Extraverts or Introverts. For example, [<END> *hi*], the <END> (end–of–file marker) followed by *hi*, was unique to Extravert texts; and since the <END> marker separates concatenated files in the corpus, here we have a tendency towards message-initial *hi*. By contrast the more formal [<END> *hello*] was found solely in Introvert texts. Use of punctuation also differs between the two groups, with Extraverts preferring multiple exclamation marks [*! !*], and solely using multiple full stops [. .] as in the elliptical (. . .), again a feature of informal style, and 'looser' use of language.

[4]When comparing the groups High E ($\geq$1s.d.), Mid E ($< \pm$1s.d.) and Low E ($\leq -$1s.d.) (all P and N $< \pm$1s.d.) it was found that Low E and Mid E correlate very significantly ($p < .005$; $\rho = .67; N = 19$), whilst High E and Mid E do not significantly correlate at the $p < .05$ level ($\rho = .32; N = 24$).

Table 4: Bigrams unique to Extravert corpus.

Bigram	Rank	$-2 \log \lambda$	Count	Ratio
. .	8	183.48	152	0.0113
of the	33	79.47	40	0.0030
, which	20	100.89	25	0.0019
had a	16	115.60	22	0.0016
which was	24	95.69	19	0.0014
new year	7	192.22	18	0.0013
got a	45	66.65	17	0.0013
a good	46	64.45	16	0.0012
forward to	26	94.76	15	0.0011
need to	28	89.99	15	0.0011
i'll be	22	98.70	14	0.0010
on saturday	27	90.94	13	0.0010
we went	42	67.54	11	0.0008
as well	43	67.18	11	0.0008
couple of	30	84.18	10	0.0007
want to	41	68.01	10	0.0007
the moment	44	67.09	10	0.0007
<END> hi	21	99.44	9	0.0007
able to	50	61.19	9	0.0007
take care	23	96.00	8	0.0006
catch up	39	70.50	7	0.0005
other than	49	62.84	6	0.0005

Quantification In terms of quantification, Introverts generally tend to show a preference for a greater use of quantifiers, such as [*a lot*], [*a few*] and uniquely [*all the*], [*one of*], [*lots of*] and [*loads of*], whereas Extraverts show a preference for [*a bit*] and uniquely use [*couple of*]. Not only does this demonstrate an Extravert tendency to be looser and less specific, it also apparently reveals exaggeration on the part of the Introvert.

Social Devices The Extravert use of stylistic expressions such as [*catch up*] and [*take care*] indicate a relaxed and informal style; their omission points to a more socially restrained Introvert. A surprisingly neat equivalence in expression can be found between the Extravert use of [*other than*] rather than [*apart from*], although it is not immediately clear what might give rise to this.

Self/Other Reference References to self in the texts demonstrate differences between Extraverts and Introverts: Introverts make extensive use of the first person singular pronoun ([*i don't*], [*i went*], [*i'm going*], [*i can*], [*i've got*] are all unique to the Introvert text), and also show preference for the following shared bigrams: [*i've done*], [*i think*], [*i am*], [. *i*].

Table 5: Bigrams unique to Introvert corpus.

Bigram	Rank	$-2 \log \lambda$	Count	Ratio
. <END>	17	80.13	20	0.0022
i don't	18	78.77	18	0.0020
went to	25	63.53	15	0.0017
to go	34	56.65	14	0.0016
all the	47	43.06	12	0.0013
i went	50	42.70	12	0.0013
one of	32	57.45	11	0.0012
trying to	29	60.75	10	0.0011
i'm going	36	52.84	10	0.0011
i can	46	43.90	10	0.0011
on thursday	20	72.22	9	0.0010
don't know	21	69.76	9	0.0010
i've got	35	55.19	9	0.0010
lots of	26	62.29	8	0.0009
this week	39	48.51	8	0.0009
anyway ,	45	44.79	8	0.0009
should be	40	48.10	7	0.0008
on monday	41	47.91	6	0.0007
two weeks	31	58.65	5	0.0006
loads of	49	42.72	5	0.0006
<END> hello	44	45.05	4	0.0005
exam results	42	47.26	3	0.0003

For Extraverts, the only unique first person bigram is [*i'll be*], and they also show greater use of [*i was*] and [*i will*], although relatively less preferred than Introvert forms. This underscores the increased Introvert tendency to focus on self, whereas the only bigram containing a first person plural is unique to Extraverts ([*we went*]). The Extravert preference for the bigram [*up with*] typically indicates a shared experience (prompting the question *with whom?*) and greater sociability. These results apparently contradict Furnham (1990) on pronouns, but given that the vast majority of pronouns here are first-person singular, thus focusing on self, this is unsurprising.

Valence Bigrams containing negations were used significantly only by Introverts, as in [*i don't*] and [*don't know*] (indeed [*i don't*] is the bigram with most frequent use of *i*) , whilst Extraverts used the bigram [*a good*] which is suggestive of positive affect.[5] Similarly, the Extravert preference for [*looking forward*] and [*forward to*] (presumably as in *looking forward to*) also suggests a more positive disposition.

Ability Personal views on capability are suggested by the different collocations with infinitival *to*.[6] For Extraverts, their ability to do something should they choose is confidently and assertively relayed using *want–*, *need–*, and *able– (to)*; which they use uniquely. Introverts more timidly and tentatively state that they are [*trying to*] or possibly—and at some point in the future—they are [*going to*].

Modality Similarly, collocations with the verb *be* show a distinction in use of modal auxiliaries which has an effect on the projection of certainty. For example, Introverts are unique in their use of the weaker and more tentative *should be*, whereas Extraverts show a greater use of the stronger predictive [*will be*], and are unique in their use of the contracted form [*i'll be*] (*i will be*) (Coates, 1983).

Message Planning/Expression Looking towards surrogates of grammatical construction, Extraverts and Introverts differ in their use of connectives: Introverts show preference for the coordinating conjunctions [, *and*] and [, *but*], whilst Extraverts uniquely show use of the subordinating [, *which*], usually deployed in an evaluative sense.

Discussion

In summary, our results support earlier findings, and suggest some new conclusions.

We found that Extraverts produce texts with more words, which supports the previous findings for speech (Carment *et al.*, 1965), whilst the reduced concreteness of Extravert language is a novel finding. It may be a direct consequence of talking or writing more, if the pressure to produce words at a high rate (in order to hold the floor, for instance) diverts resources away from more detailed lexical planning. Introverts' greater preference for numbers and quantification fits with this, and is compatible with findings concerning the use of articles (Pennebaker and King, 1999), and suggestions of a more imprecise and 'looser' Extravert style (Furnham, 1990).

Extraverts' use of other or social referents, and Introverts' preference for self referents confirms Berry *et al.* (1997)'s previous findings for Extraversion and its *dominant/submissive* facets. Another possible manifestation of the increased Extravert social ability and ease in interaction is expressed by their use of surface features and social devices. We also note in passing the tendency of Extraverts to refer to days of the weekend, where Introverts refer to weekdays.

Our results on valence are consistent with previous findings on Introverts' preference for negations and negative emotion words, and the Extravert tendency for positive affect words is consistent with results for *warmth*. However, they do suggest that care should be taken over the relation between Extraversion and *dominant* facet features (cf. Isbister and Nass, 2000).

Expressions of definite modality and ability appear to be associated with Extraversion, although they may not be the same forms as those discussed in the context of powerful/less speech. Adoption of definite modalities can also be related to avoidance of tentativity (Pennebaker and King, 1999).

Turning to connectives, we note that our Introvert

[5] Further investigation shows that *good* is not directly negated (as in [*not good*]). Compare the Introvert [*i can*], which was generally followed by *not*. Although the effect of negation was not viewed as important by Pennebaker in the functioning of LIWC, it certainly has implications for models of language generation.

[6] This confirms the appropriacy of retaining functors usually filtered out by a stop list (cf. Damerau, 1993).

preference for [, *and*] and [, *but*] is consistent with studies using LIWC which found that the dictionary categories of Inclusion and Exclusion were both inversely correlate with Extraversion. However, [*other than*] and [*apart from*] would both fall into the same LIWC category, yet appear to distinguish opposite ends of the personality dimension.

Conclusion

By combining techniques from psycholinguistics and statistical natural language processing, we have been able to replicate previous findings on the expression of Extraversion through language, and uncover some new linguistic behaviours. Where existing content analysis tools could not detect reliable differences, more sensitive linguistic tools proved their worth.

Further, more technically sophisticated analyses can be carried out on this data, and we envisage the use of machine learning techniques to identify distinctive features from the texts, along with bigram analysis exploiting Parts of Speech tags. Additionally, the role of gender could be investigated.

Our findings could be exploited within the field of automatic language generation. As they stand, stochastic techniques would be needed; however, a cognitively-based personality model would allow a deeper approach, and that is our eventual goal.

Acknowledgements

Thanks to Elizabeth Austin, James Curran and our anonymous reviewers for advice and comments. This work was supported by the Economic and Social Research Council (Award R00429934162).

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph.D. thesis, Royal Institute of Technology, Stockholm.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Bradac, J. and Mulac, A. (1984). A molecular view of powerful and powerless speech styles. *Communication Monographs*, **51**, 307–319.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. Croom Helm, London.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proc of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**, 497–505.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, **29**, 433–448.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**(3), 509–544.

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**(2), 265–324.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Eysenck, H. and Eysenck, S. (1991). *Eysenck Personality Questionnaire-Revised*. Hodder, London.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**(1), 21–29.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (in prep.). Dictionary approaches to personality language. *in prep.*

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. J Human-Computer Studies*, **53**, 251–267.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *Int J Human-Computer Studies*, **43**, 223–239.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.

Pennebaker, W. and Francis, M. (1999). *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**(6), 1296–1312.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

D.2 Gill and Oberlander (2003a)

Looking forward to more Extraversion with N-grams

Alastair J. Gill and Jon Oberlander

School of Informatics, University of Edinburgh

2 Buccleuch Place, Edinburgh, EH8 9LW UK

agill@cogsci.ed.ac.uk J.Oberlander@ed.ac.uk

Abstract. We study how Extraversion or Introversion influences people's language production. Extending recent work, we show how the use of larger-scale co-occurrences of words distinguishes these personality groups. Along with previous findings, our results suggest that Extraverts could be "lazy" and use collocations of words to economise on discourse planning. We compare these results with previous findings for personality language. Implications of using co-occurrence techniques are discussed.

1. Introduction

We study the impact of personality on textual communication, in particular through computer-mediated means. The trait Extraversion-Introversion is especially relevant since this describes sociability, which is important for communication, and is readily perceived, even in computer-mediated communication (Gill and Oberlander, 2003).

Recent work using the MRC psycholinguistic database has shown that Extraverts use words which are less *concrete*, and more abstract, like *thoughts, flavours, pains*, rather than referring to entities which can be sensed like *table, spoon, girl* (Gill and Oberlander, 2002). In addition, we went on to demonstrate that Extravert and Introvert authors are distinguished by a range of two-word collocations (bigrams). A summary of these features can be found in Figure 1.

Surface Realisation: Extraverts are more informal, use *hi*, and use looser punctuation (!! or ...); Introverts use *hello.*

Quantification: Introverts show greater use of quantifiers (for exaggeration?); Extraverts are looser and less specific.

Social Devices: Stylistic expressions such as *catch up* and *take care* indicate the Extravert's relaxed social style.

Self/Other: Reference Introverts use more first-person singular (*i*), whereas Extraverts are more likely to use plural *we*.

Valence: Introverts prominently use negations; Extraverts use words suggestive of positive affect.

Ability: Extraverts are more confident and assertive (eg., *want-*, *able-*, *need-*(*to*)); Introverts are more tentative and timid (*trying-*, *going-*(*to*)).

Modality: Extraverts are more strongly predictive than Introverts (eg., modal auxiliaries *will-* vs. *should-*(*be*)).

Message Planning/Expression: Introverts prefer co-ordinating conjunctions (*and*, *but*), whereas only Extraverts use the subordinative *which* (usually for evaluation?).

Figure 1: Extravert and Introvert Language

So far, these two separate findings have been viewed in isolation. However, in this paper we aim to draw them together in an explanation of Extravert discourse behaviour. We propose that Extraverts direct resources away from precise lexical planning, in an endeavour to construct utterances more quickly. Their drive to seize the conversational floor leads to a certain linguistic *laziness*. This is, however, not laziness in the sense of indolence. Rather, it is an *efficiency of action*, whereby new or precise linguistic decisions are avoided in favour of pre-existing, remembered choices. In particular, such speakers are more likely to rely upon stereotypical expressions and previously used or pre-planned chunks of language: The collocations found in the previous bigram analysis suggests that Extraverts use regularly co-occurring pairs of words more frequently than Introverts.

To test this theory, we build upon the bigram analysis, and extend it so as to consider larger collocations of words. The structure of the paper is as follows: First we will introduce in more detail the concept of Extraversion and why it is such an important personality trait. We then briefly describe some findings for Extravert language use. Next, we introduce the experimental method used for the original bigram analysis and detail the extensions used in the current analysis. Then follows the discussion and conclusion.

1.1 The importance of being Extravert

Intuitively, we get the impression that Extraverts tend to talk loudly and say more, whereas Introverts are more softly spoken and reserved. Are such hypotheses borne out by fact, and how else does this personality dimension influence language production? Before approaching this question, we define more precisely what is meant by Extraversion, and why this trait is important.

Extraversion is a trait which is strongly related to interpersonal interaction and sociability, and as a result there is a greater awareness of this trait and its manifestation in behaviour. A typical Extravert is described as someone who is sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert is quiet, retiring, reserved, plans ahead, and dislikes excitement (Eysenck and Eysenck, 1991).

The trait of Extraversion is central to the two major theories of personality psychology: Eysenck's three factor model; and the five factor model developed by Costa and McCrae and others (Matthews and Deary, 1998). Indeed, the personality trait of Extraversion is one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993).

Despite the general agreement for the inclusion of Extraversion in personality theory, beyond this there is greater debate. For example, Eysenck's model of personality incorporates just two further dimensions: Neuroticism, which is mainly characterised by susceptibility to anxiety; and Psychoticism, which is more complicated, but generally related to aggression and individuality. By contrast, the NEO-PI-R model incorporates five factors (Costa and McCrae, 1992). In addition to Extraversion and Neuroticism, they proposed three other traits: Conscientiousness, Agreeableness and Openness, which are generally regarded as relating to Psychoticism; but this is still a matter of some debate (cf. Matthews and Deary, 1998).

But how does Extraversion influence an individual's language production? In addressing this question, we first outline some hypotheses from the literature, before describing our collection of a controlled corpus of language, and our analysis of it.

1.2 Previous hypotheses

From an intuitive perspective, Extraverts are described as individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts monopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (cf. Carment, Miles, and Cervin, 1965). With reference primarily to speech, Furnham (1990) has proposed that Extravert language is less formal, has a more restricted code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely (see also Dewaele and Furnham, 1999, for a review of speech and writing studies).

Text analysis approaches have found that transcribed texts rated as belonging to the *warm* facet of Extraversion used fewer negative emotion words and unique words, and more present tense verbs, with *dominant* texts using fewer unique words, positive emotion words and self referents (Berry, Pennebaker, Mueller, and Hiller, 1997). Finally, study of the texts *written* by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclusive words, while using more social and positive emotion words (Pennebaker and King, 1999).

2. Method

Our extension of *n*-gram analysis uses the same data and methods as our previously reported bigram analysis (Gill and Oberlander, 2002), namely: 210 texts produced by 105 University students or recent graduates (37 males, 68 females; mean age = 24.3 years; SD = 4.6; all native English speakers) of known personality (EPQ Revised short form; Eysenck, Eysenck, and Barrett, 1985; mean score = 7.91, SD = 3.25; normative score = 7.42 (male), 7.60 (female)). Note that these personality scores depend on subjects' self-assessment: they do not depend on peer-judgement, and hence do not depend on external judgments concerning the subjects' verbal behaviours. Each participant composed two e-mails *to a good friend whom they hadn't seen for quite some time*, spending around 10 minutes on each message. The first e-mail concerned their activities in the past week, the second discussed their plans for the next week. The total corpus size is around 65,000 words.

The original corpus of texts was divided by degree of Extraversion by selecting respondents whose E score was greater or less than 1 s.d. of the mean (cf. Dewaele and Pavlenko, 2002), with the 21 High Extravert authors scoring more than 11, and the 17 Low Extravert authors scoring less than 5. The resulting Extravert and Introvert sub-corpora contain around 12,000 words and 8,000 respectively, which resulted from the average length of Extravert texts being longer than that of Introvert texts (around 570 words versus 470 words). These subcorpora were used for the subsequent calculation of *n*-grams. This was performed using word co-occurrence window lengths of 3 and 5 words.[1]

The trigram data for each corpus was then ranked by their co-occurrence significance using the log-likelihood statistic ($-2 \log\lambda$), since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993). Rankings for each group are based on the top 50 trigrams with frequency of $N >= 2$, and a significance of $p<.001$. Relative frequency ratios

[1] Ted Pedersen's n-gram software is available from: http://www.d.umn.edu/~tpederse/code.html

(Damerau, 1993)[2] were then calculated for trigrams that were common to both the sub-corpora, and a Spearman Rank correlation was then performed on this data. Note that here the *n*-gram analysis and relative frequency ratios are used for different purposes than those of, for example, Damerau (1993), who uses them to distinguish texts on the basis of key words. Due to a scarcity of data and statistical tools for 5-grams, frequency and relative frequency alone were calculated.

3. Results

3.1 Spearman Rank Correlation

Extravert and Introvert use of the shared trigrams is not significantly correlated r_s =.236 (N=13) at the $p < .05$ and therefore indicates that the two groups' usage of these is distinct.

3.2 N-grams

The results of the relative frequency ratio analysis of the trigrams, and those unique to Extravert and Introvert corpora can be found in Tables 1, 2 and 3. The 5-grams with a frequency of at least 3 occurrences, are shown for the Extravert group in Table 7. Introvert 5-grams failed to reach this frequency. For reference, the previous findings of the bigram analysis are also presented in Tables 4, 5 and 6. These represent the relative frequency ratio data and bigrams unique to Extraverts and Introverts respectively.

Trigram	Extr Freq	Intr Freq	Extr R. Freq	Intr R. Freq	Rel. F Ratio
a bit of	10	2	0.0007	0.0002	3.31
. i have	12	3	0.0009	0.0003	2.65
! ! !	17	6	0.0013	0.0007	1.88
. it was	19	7	0.0014	0.0008	1.80
. i think	8	4	0.0006	0.0004	1.33
. i am	13	7	0.0010	0.0008	1.23
. i was	9	6	0.0007	0.0007	0.99
for a bit	3	2	0.0002	0.0002	0.99
i am going	6	6	0.0004	0.0007	0.66
i have to	7	9	0.0005	0.0010	0.52
need to get	3	4	0.0002	0.0004	0.50
i'm going to	5	8	0.0004	0.0009	0.41
that i am	2	4	0.0001	0.0004	0.33

Table 1: Shared Extravert and Introvert trigrams

[2] Note here that functors and rarer collocations are retained.

Trigram	Rank	-2logλ	Freq	Rel Freq
. . .	2	478.73	71	0.0053
looking forward to	5	328.73	15	0.0011
it was a	22	248.92	10	0.0007
really looking forward	19	270.95	6	0.0004
want to get	44	225.53	6	0.0004
next week .	47	222.33	6	0.0004
i will be	48	222.21	6	0.0004
going to get	9	306.28	5	0.0004
! it was	36	230.07	5	0.0004
i have been	37	228.85	5	0.0004
next week ,	38	228.05	5	0.0004
. . i	3	375.01	4	0.0003
! ! so	10	298.21	4	0.0003
the next week	20	267.20	4	0.0003
i'm looking forward	21	253.48	4	0.0003
a bit worried	23	247.71	4	0.0003
was a bit	26	242.06	4	0.0003
for next week	39	227.78	4	0.0003
going to do	18	273.76	3	0.0002
am looking forward	24	247.01	3	0.0002
will be able	28	240.29	3	0.0002
it was cool	32	234.94	3	0.0002
it was nice	34	233.68	3	0.0002
< END > next week	43	226.54	3	0.0002
and i am	46	223.07	3	0.0002
. . < END >	49	221.95	3	0.0002
it was really	50	221.69	3	0.0002
! ! not	11	290.85	2	0.0001
! ! it	12	289.13	2	0.0001
! ! ,	13	288.58	2	0.0001
! ! and	14	287.44	2	0.0001
! ! on	15	287.06	2	0.0001
so i am	29	240.01	2	0.0001
, it was	31	238.97	2	0.0001
i am looking	40	226.87	2	0.0001
but it was	41	226.82	2	0.0001
quite a bit	42	226.79	2	0.0001

Table 2: Trigrams unique to Extravert corpus.

Trigram	Rank	-2logλ	Freq	Rel Freq
going to the	13	188.42	7	0.0008
. but i	22	166.15	6	0.0007
i don't know	37	149.07	6	0.0007
going to be	10	197.73	5	0.0006
am going to	14	185.55	5	0.0006
. i don't	8	202.76	4	0.0004
managed to get	42	142.86	4	0.0004
in the evening	50	135.90	4	0.0004
going to see	2	239.05	3	0.0003
going to go	6	208.67	3	0.0003
. i got	24	163.75	3	0.0003
, and then	25	163.24	3	0.0003
. i played	26	162.53	3	0.0003
. i will	32	155.63	3	0.0003
. i wasn't	36	150.00	3	0.0003
, but it	41	145.19	3	0.0003
is a bit	45	137.72	3	0.0003
tomorrow i am	11	192.45	2	0.0002
i am not	16	183.39	2	0.0002
i am in	17	177.91	2	0.0002
probably going to	19	169.68	2	0.0002
it's going to	20	168.39	2	0.0002
going to book	21	167.60	2	0.0002
were going to	21	167.60	2	0.0002
, but it's	23	164.30	2	0.0002
trying to get	27	161.84	2	0.0002
be going to	28	161.26	2	0.0002
just going to	30	157.96	2	0.0002
was going to	31	155.84	2	0.0002
. i had	33	152.71	2	0.0002
went to see	34	152.68	2	0.0002
going to a	35	152.27	2	0.0002
, but that's	38	148.39	2	0.0002
, but there's	39	146.48	2	0.0002
, but he	40	146.29	2	0.0002
. i should	47	136.98	2	0.0002
. i still	48	136.57	2	0.0002
again . i	49	136.06	2	0.0002

Table 3: Trigrams unique to Introvert corpus.

Bigram	Extr Freq	Intr Freq	Extr R. Freq	Intr R. Freq	Rel. F Ratio
looking forward	15	4	0.0011	0.0005	2.49
it was	46	22	0.0034	0.0025	1.39
next week	24	12	0.0018	0.0013	1.33
a bit	29	15	0.0022	0.0017	1.28
up with	19	10	0.0014	0.0011	1.26
! !	45	24	0.0033	0.0027	1.24
will be	24	13	0.0018	0.0015	1.22
i was	33	18	0.0025	0.0020	1.22
at the	27	16	0.0020	0.0018	1.12
to see	32	19	0.0024	0.0021	1.12
which is	15	9	0.0011	0.0010	1.11
for a	34	21	0.0025	0.0024	1.07
i have	44	29	0.0033	0.0032	1.01
to get	34	23	0.0025	0.0026	0.98
. i	99	69	0.0074	0.0077	0.95
on friday	11	8	0.0008	0.0009	0.91
, and	48	36	0.0036	0.0040	0.88
and then	23	19	0.0017	0.0021	0.80
in the	41	34	0.0031	0.0038	0.80
apart from	6	5	0.0005	0.0006	0.80
i am	33	28	0.0025	0.0031	0.78
i think	16	14	0.0012	0.0016	0.76
, but	35	31	0.0026	0.0035	0.75
a lot	10	9	0.0007	0.0010	0.74
going to	36	33	0.0027	0.0037	0.72
a few	12	11	0.0009	0.0012	0.72
to do	23	23	0.0017	0.0026	0.66
i've been	9	12	0.0007	0.0013	0.50

Table 4: Shared Extravert and Introvert bigrams.

Bigram	Rank	-2 logλ	Freq	Rel Freq
. .	8	183.48	152	0.0113
of the	33	79.47	40	0.0030
, which	20	100.89	25	0.0019
had a	16	115.60	22	0.0016
which was	24	95.69	19	0.0014
new year	7	192.22	18	0.0013
got a	45	66.65	17	0.0013
a good	46	64.45	16	0.0012
forward to	26	94.76	15	0.0011
need to	28	89.99	15	0.0011
i'll be	22	98.70	14	0.0010
on saturday	27	90.94	13	0.0010
we went	42	67.54	11	0.0008
as well	43	67.18	11	0.0008
couple of	30	84.18	10	0.0007
want to	41	68.01	10	0.0007
the moment	44	67.09	10	0.0007
< END > hi	21	99.44	9	0.0007
able to	50	61.19	9	0.0007
take care	23	96.00	8	0.0006
catch up	39	70.50	7	0.0005
other than	49	62.84	6	0.0005

Table 5: Bigrams unique to Extravert corpus.

Bigram	Rank	-2 logλ	Freq	Rel Freq
· < END >	17	80.13	20	0.0022
i don't	18	78.77	18	0.0020
went to	25	63.53	15	0.0017
to go	34	56.65	14	0.0016
all the	47	43.06	12	0.0013
i went	50	42.70	12	0.0013
one of	32	57.45	11	0.0012
trying to	29	60.75	10	0.0011
i'm going	36	52.84	10	0.0011
i can	46	43.90	10	0.0011
on thursday	20	72.22	9	0.0010
don't know	21	69.76	9	0.0010
i've got	35	55.19	9	0.0010
lots of	26	62.29	8	0.0009
this week	39	48.51	8	0.0009
anyway .	45	44.79	8	0.0009
should be	40	48.10	7	0.0008
on monday	41	47.91	6	0.0007
two weeks	31	58.65	5	0.0006
loads of	49	42.72	5	0.0006
< END > hello	44	45.05	4	0.0005
exam results	42	47.26	3	0.0003

Table 6: Bigrams unique to Introvert corpus.

5-gram	Freq	Rel Freq
· · · it was	4	0.0003
really looking forward to seeing	3	0.0002
my plans for next week	3	0.0002
i'm really looking forward to	3	0.0002
· i'm looking forward to	3	0.0002
what i've been up to	3	0.0002

Table 7: 5-grams unique to Extravert corpus.

4. Discussion

Our discussion of these results will take the following form: Firstly we discuss the evidence from the trigrams and 5-grams which suggests different collocation usage by the Extravert and Introvert groups, and in particular whether a distinct pattern is present for the Extraverts; Secondly, we will evaluate the usefulness of the Extravert/Introvert characteristics summarised in Figure 1 which were formulated on the basis of the bigram data, and discuss whether they are supported in the current findings; Finally we assess the role of word collocation in personality language.

4.1 Extravert-Introvert collocations

The trigram analyses reveal an even more distinctive pattern of Extravert and Introvert language use, than was found for the bigrams. This is demonstrated firstly by the greater number of unique occurrences found for both personality types than was the case in the bigram analysis, and secondly by the non-significant correlation in the ordering of occurrence of trigrams shared between the two personality groups.

Turning to the 5-gram data, it can be seen that when a frequency cut-off of 3 occurrences is used, co-occurrence data is only found for the Extravert group. Given the modest data set, it is not surprising that few repeated 5-grams are found; indeed it could be argued that the relative difference in size between the Extravert and Introvert sub-corpora is responsible for this finding, although this in itself highlights the longer length of text produced by Extraverts, which is around 20% longer. However, when referring to data for 5-grams occurring with a frequency of 2, there are still disproportionately more of them for the Extravert group (n=56) than for the Introvert group (n=18). This pattern is also found from analysis of the whole of the trigram data occurring with a frequency of at least 2 and significance of $p < .005$. In this case, for the Extraverts 608 of 729 are unique, and for the Introverts this is 288 of 409, with 121 trigrams shared by both personality groups.

In order to better utilise the information that can potentially be provided by larger window *n*-gram analysis, a larger corpus would be preferable, along with a higher frequency cut off (eg. 5) and possibly also a statistical test of co-occurrence, like log-likelihood.

Before examining the trigram results in more detail, it is important that we clarify co-occurrence further. In the current analysis we have included or rather *not excluded* by way of stop list functors, punctuation, or rarer words and collocations, since the purpose of *n*-gram analysis in the current study is to find characteristic language patterns more generally, rather than the identification of, for example, key words.

We therefore distinguish co-occurrence more generally, into collocation, and colligation. Collocation, as we define it here, is what is perhaps more generally understood by the term *co-occurrence*, that is, 'the patterns of combinations of words (for example, with other words) in a text' (Oakes, 1998). Examples of collocation would be words which may occur separately, but occur together in a significant and meaningful way, in the way that *corpus linguistics* and *word frequency* may feature in the genre of corpus linguistics.

Colligation, on the other hand, is information which again is derived from co-occurrence information, eg. *n*-grams, but could not be described as collocation in the traditional sense. It is usually seen as more grammatically-oriented, covering the syntactic preferences of a word. For us, examples of colligation would be the positioning of words in relation to punctuation or other boundary markers, indicating that a particular word or token occurs in a text or sentence initial or final position. In the genre of formal letter writing, an example of a colligational co-occurrence might be *"start of document"* followed by *Dear*. Although punctuation is generally used to signal a sentence or phrase boundary, and is thus useful in determining colligation, we further distinguish between punctuation when used for a purely syntactic purpose, and when it is used to encode additional meaning, as is often the case in e-mails for example: multiple full stops or exclamation marks.

Given our hypothesis that Extraverts are more likely to use and re-use chunks of language, we would expect that collocations will constitute a larger proportion of total co-occurrences (and colligations a smaller proportion) for Extraverts, compared with Introverts.

Therefore in examining the co-occurrence data, we turn first to the trigrams which are shared by both Extraverts and Introverts. Here we can see that almost half of the trigrams contain elements of punctuation. Five of these provide colligations concerning (presumably) sentence initial constructions ([. *i have*], [. *it was*], [. *i think*], [. *i am*], and [. *i was*]), and appear to be favoured by the Extraverts. Note that [*! ! !*] is considered to be collocation, rather than colligation.

This use of colligation trigrams by Extraverts is perhaps unexpected. However, while the relative ratio suggests they are more characteristic of Extraverts, raw counts suggest they are used frequently by both Introverts and Extraverts. Indeed it may be the case that Introverts and Extraverts are using the same constructions differently. For example, examining the trigram data which is unique to the personality groups shows that whilst trigrams with the first element being a full stop are likely to indicate the end of a sentence for Introverts, for Extraverts this is more likely to be the last element of an elliptical [. . .].

Other patterns from the unique trigram data are that Extraverts show some use of colligation ([*next week .*], [*next week ,*], [< *END* > *next week*]). This seems to be largely topic specific, resulting from the extraposing of the author's current concern.

When this is contrasted with the colligation trigrams used uniquely by the Introverts, it can be seen that these contain a great deal more information about the relative focus and the syntactic constructions favoured by the Introvert authors. In choosing to write about their past or forthcoming week, rather than extraposing that time period, as in the case of the Extraverts, the colligations show that instead Introverts focus on themselves. Therefore a large proportion of their trigrams demonstrate a sentence initial first-personal singular pronoun, *I* ([. *i don't*], [. *i got*], [. *i played*], [. *i will*], [. *i wasn't*]). Furthermore, the colligation data of Introverts also demonstrate use of co-ordination, particularly *but* ([, *but i*], [, *but it*], [, *and then*]).

This data shows then, that Introverts do in fact show greater proportional use of colligation. We now turn to the collocation trigrams to examine the evidence for the frequent usage of chunks of text.

Both personality groups share the use of phrases such as *a bit* ([*a bit of*], [*for a bit*]) and *am going* ([*i am going*], [*i'm going to*]), although Extraverts prefer the former constructions and Introverts the latter. When the unique data for these personality groups is consulted, this pattern is borne out with Introverts' extensive use of collocations which include *going to* ([*going to the*], [*going to be*], [*am going to*], [*going to see*], [*going to go*]), versus those of the Extraverts ([*going to get*], [*going to do*]). Conversely, the Extraverts use more of *a bit* ([*a bit worried*], [*was a bit*]) versus the Introvert [*is a bit*].

Although featuring punctuation, [*! ! !*], is regarded as a collocation, and is a feature preferred by Extraverts. Examination of the unique data shows that this non-standard use of punctuation, along with the elliptical (...) are key features of Extravert texts ([. . .], [. . *i*], [*! ! so*], [. . < *END* >]).

The co-occurrences unique to Extraverts show a larger number of collocations. Some of these refer to the future, such as *will be* ([*i will be*], [*will be able*]), whereas the evaluative [*it was cool*], [*it was nice*] and [*! it was*] refer to the past. As previously mentioned, reference to the topic of *next week* occurs frequently ([*next week .*], [*next week ,*], [*the next week*], [*for next week*], [< *END* > *next week*]), as does *looking forward* ([*looking forward to*], [*really looking forward*], [*i'm looking forward*], [*am looking forward*]). These trigram patterns feature again

in the Extravert 5-grams in [. . . *it was*], [*really looking forward to seeing*], [*i'm really looking forward to*] and [*my plans for next week*].

On the basis of this evidence, it appears that Extravert and Introvert use of co-occurrence is different, with the Extraverts tending to use larger chunks of word collocations, and the higher proportion of Introvert colligations suggesting characteristic syntactic constructions. The co-occurrences which were shared by both groups were also shown to be used in significantly distinct ways.

4.2 Personality language style

Although the previous findings presented in Figure 1 were based upon bigram data, we now address whether the current extension of the analysis using higher *n*-grams still supports these broad personality language features.

Potentially using larger windows of text allows the identification of larger-scale features from the data, in the current case, patterns of between 3 and 5 words or characters. However, this also means that collocations of two words which co-occur with a large variety of words on either side will not show up in the current extension of the analysis. This means that whilst the Surface Realisation features (. . .) and (*!!!*) are very apparent in the trigram analysis, others such as the message initial *hi* or *hello* are not, since the name—or lack of name—which tends to follow is not a stable feature. Similarly, the bigrams characteristic of the Social Devices category *catch up* and *take care* also did not occur in the present analysis.

In a similar way, the result of analysis using 3-word windows on Message Planning and Expression features, is that the co-ordinations (, *and*) and (, *but*) are isolated in patterns which are even more strongly characteristic of Introversion (the previous bigram analysis found them used by both but preferred by Introverts). However, the Extravert feature (, *which*) was not found to occur in the present trigram or 5-gram analysis.

This pattern is repeated in the other bigram feature categories: for Quantification, Introverts do not demonstrate the large variety of features found originally, instead they make less use of (*a bit*) which is a rather vague, shared term used primarily by Extraverts; evidence of Modality is only found for the strongly predictive Extraverts (*will be*), but not for Introverts; the timid Ability of Introverts is found in (*trying to*) and the shared form (*going to*), but confident Extravert forms are not found.

Features expressing Valence were still found characteristically in the Extravert and Introvert texts: The former used expressions such as (*looking forward*) and *nice* and *cool*, with the latter employing the contracted negation *don't*. In the case of Self/Other Reference, although the Extravert tendency to refer to others was not maintained, further evidence for the mainly Introvert self-reference was found. Indeed, the colligations revealed interesting difference in the occurrence of the first-person singular, with Introverts tending to use this in the sentence initial position, whereas Extraverts were more likely to use it positioned within a sentence, or following elliptical (. . .).

These findings therefore largely support the previous Extravert-Introvert language features derived from the bigram analysis. Although the use of larger windows for co-occurrence analysis can uncover larger-scale language patterns, this can also result in the loss of patterns which only stably occur in two-word windows. Furthermore, the use of larger windows can

result in data sparsity, especially when using smaller corpora, and this is especially relevant for the Introvert data.

4.3 The lazy Extravert

In this paper we proposed that Extravert discourse strategy is based upon a kind of *laziness*, which manifests itself in their recycling of formulaic chunks of words. In our *n*-gram analyses we have demonstrated differences between Extravert and Introvert language usage which suggest that this is in fact the case. Note that we do not exclude the possibility that *everyone* re-cycles formulaic language, at least to some extent. The point is that Extraverts do so more than Introverts.

But why should Extraverts be particularly lazy and prone to re-using language features? Such behaviour is not without good reason since it serves the drives of the Extravert well. Earlier we described the Extravert as someone who is sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. Furthermore they think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic.

Through these personality descriptions we see an Extravert who wants to be the centre of attention, and as a result wants to gain the floor by quickly formulating a comment, or to hold on to it by continuing to talk. In contrast, the Introvert is less concerned with talking for talking's sake, but instead will be more inclined to enter into the conversation with a carefully considered contribution when they feel this is warranted.

These different conversational stances therefore impose a different set of constraints upon the Extravert and Introvert speakers. Introverts can afford greater mental resources in the planning and preparation of an utterance and thereby risk losing a conversational turn if another speaker formulates and executes a contribution more quickly thereby making the Introvert's irrelevant. Extraverts, when they are not speaking, are under pressure to quickly make a comment, thereby entering the conversation and gaining control of the floor. This process itself forms part of the Extravert's stimulation feedback loop, with fighting for the floor providing the stimulation which Extraverts crave.

We therefore propose that such pressure upon Extraverts to more quickly produce linguistic contributions leads to the employment of distinctive discourse strategies. Indeed, Furnham (1990) suggests that the Extravert has a more restricted code, which could well be the result of such constraints and would fit in with our observation of the reduced concreteness of such utterances, and the tendency to recycle pre-formed chunks of language.

Although previous discussion has concentrated upon the *spoken* language of Extraverts, we suggest that similar patterns occur in all naturalistic language production settings, since Extraversion is a stable trait which consistently influences an individual's behaviour. Instances where this may not play such a large role would be in carefully constructed written texts and where several iterations of editing are likely to occur. Given that the style of e-mail is considered to be close to that of oral communication (Bälter, 1998), we would expect that laziness, typical of Extraverts, is found in e-mails and similar texts.

5. Conclusion

We have shown that Extraverts and Introverts use larger-scale co-occurrences of words in characteristically distinct ways through *n*-gram analysis. This has extended recent work which derived Extravert and Introvert linguistic behaviour using a combination of techniques from psycholinguistics and statistical natural language processing.

This differentiation between personality groups lends support to our hypothesis, based on previous findings, that Extraverts are "lazy" and use larger-scale collocations of words in order to spend less time planning discourse. A greater proportion of co-occurrence information for Introverts was colligational and related to the structure of their text. Our trigram and 5-gram analyses broadly support previous findings for bigrams. However, we note that in some cases bigrams may be more informative, and that care should be taken with regard to data-sparsity with larger *n*-gram analyses.

Further, more technically sophisticated analyses can be carried out: we envisage the use of machine learning techniques to automatically classify texts on the basis of the distinctive features we are isolating, along with further *n*-gram analysis exploiting 'parts of speech' tags.

6. Acknowledgements

Thanks to Elizabeth Austin, James Curran and our anonymous reviewers for helpful advice and comments. This work was supported by the Economic and Social Research Council (Award R00429934162), and the School of Informatics, University of Edinburgh.

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph. D. thesis, Royal Institute of Technology, Stockholm.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, **29**, 433–448.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**, 265–324.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**, 61–74.

Eysenck, H. and Eysenck, S. (1991). *The Eysenck Personality Questionnaire-Revised.* Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself, but Neuroticism is more of a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, Boston, MA, August 2003.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp363-368. Fairfax VA, August 2002.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

D.3 Gill and Oberlander (2003b)

Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry

Alastair J. Gill (agill@cogsci.ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We investigate the impact of computer-mediated interaction on person perception. In particular, we study how traits important for socialisation and collaboration—Extraversion and Neuroticism—can be detected from the text of an e-mail communication. We have previously shown how Extraversion influences people's language production in electronic communication, in broadly intuitive ways. Here, we briefly outline the ways in which Neuroticism is expressed more through the high-level properties of a text. By their nature, these properties are less accessible to intuition. In subjective ratings of the texts for personality, we demonstrate that author Extraversion can be accurately perceived, given the limited cues, and that judges also exhibit relatively high agreement with each other for this trait. Neuroticism, however, appears more difficult. This result is consistent with previous findings, but suggests that e-mail exacerbates this discrepancy.

Introduction

One view of human cognition is that it has been shaped by natural selection to enable individuals to interact effectively with members of relatively large groups of peers: to estimate the trustworthiness of strangers, to recognise individuals, and to recall our judgement of familiars.

Until relatively recently, interaction has been conducted entirely face-to-face, or at least synchronously. It is therefore unsurprising that in such contexts, we are highly effective at judging people's characteristics, such as familiarity, gender, emotion or temperament (eg. Cheng, O'Toole, and Abdi, 2001). But technology now mediates much communication. Phone, e-mail or video-conference: in each case, people must make do with impoverished cues to help them estimate other people's emotional states, dispositions and personalities. E-mail is especially popular: it is designed to allow asynchronous communication; and it is often the means by which people make first contact with one another (Baron, 1998). Given this, it seems reasonable to ask: How easily can the personality of an author be perceived from their e-mail message?

To address this question, we here focus on the personality dimensions of Extraversion and Neuroticism. The rest of the paper is therefore structured as follows. We first describe the notions of Extraversion and Neuroticism with which we are working. We then briefly survey previous findings on perception of personality, before noting particular findings concerning the effects of technological mediation on personality perception. We note the objective features of text in our e-mail corpus that vary with Extraversion and Neuroticism, and then describe the methods and results of our perception study. The discussion section focuses on why Extraversion may be easier to detect in e-mail than other personality characteristics, like Neuroticism.

Background

Two personality traits

Extraversion and Neuroticism are traits which are considered central to theories of personality. They are common to the two major theories: Eysenck's three factor model (Eysenck and Eysenck, 1991); and the five factor model developed by Costa and McCrae (Costa and McCrae, 1992) and others. Beyond these two traits there is greater dispute, with personality described either in terms of the single trait Psychoticism, or divided into Conscientiousness, Agreeableness and Openness.

Extraversion is a trait strongly related to interpersonal interaction and sociability. High Extraverts are said to: be sociable, take chances, be easy-going and optimistic. Low Extraverts (or Introverts) are said to: be quiet, reserved, plan ahead, and dislike excitement (Eysenck and Eysenck, 1991). Unsurprisingly, then, there is popular awareness of this trait, and its manifestations in behaviour.

Neuroticism is generally related to internal emotional states. High Neurotics are said to be: anxious, worrying, over-emotional, and frequently depressed. Low Neurotics are said to be: calm, even-tempered, and unworried (Eysenck and Eysenck, 1991). Although internal states are less directly perceived than interpersonal behaviour, there is also considerable popular awareness of this trait, and it makes a real difference to productivity, collaboration, and performance in jobs requiring interpersonal interaction (Mount, Barrick, and Stewart, 1998).

Perception of personality

How much about ourselves do we give away in interaction? How good are other people at picking it up? From a cognitive science point of view, we need to know what aspects of interactive behaviour can be informative, before we design models of the relevant information processing. In fact, we must turn to social and personality psychology for the appropriate empirical methods.

Personality judgement data can be gathered in several ways. On the one hand, subjects' self-reports of personality, together with ratings of subjects by peers (such as spouses or colleagues), have been compared with each other for agreement. On the other hand, strangers have been called upon to make personality judgements, after being exposed to various different kinds of information about the target individuals. Funder's (1995) Realistic Accuracy Model views accuracy of judgement as a function of the availability, detection, and utilisation of relevant behavioural cues. The first two categories he describes as 'good judge', and 'good target': some people are better able to judge—or rate—personality; and some individuals are more easily judged than others. Generally, these kinds of variation do not appear to occur systematically across groups. However, other variation, labelled 'good trait' and 'good information', is more systematic.

Good Traits Distinguishing between the different personality dimensions has shown that, even in judgements by close acquaintances, much greater agreement is found for ratings of Extraversion than for Neuroticism (or Psychoticism) for the EPQ (Gomà-i-Freixanet, 1997), and this pattern has been mirrored in the five factor model (McCrae and Costa, 1987). Additionally, self-ratings were shown to be more informative in predicting behaviour for Extraversion—but not Neuroticism. Funder (1995) proposes that this is due, in part, to the 'visibility' of Extraversion. It is realised in 'frequent positive social interaction', whereas Neuroticism is realised via internal states. Furthermore, Neuroticism is regarded as more 'evaluative', ie. affectively charged. It may thus lead to: the concealment of undesirable behaviour from observers; or a distortion of self-perception, leading to lower target-judge agreement; or a greater reluctance to pass judgement on such behaviours, leading to reduced inter-judge agreement. When less evaluative measures of Neuroticism are used, agreement increases (John and Robbins, 1993).

Good Information The amount and relevance of target information available to the judges influences their agreement. Close acquaintances agree better with each other and with the target, than do relative strangers, although both predict target behaviour equally well, when they know the target in a relevant context (Colvin and Funder, 1991). Judgements by close acquaintances (especially when taken as a composite measure) generally also better predict target behaviour (Kolar, Funder, and Colvin, 1996). At the other extreme, studies have investigated personality perception of strangers on the basis of minimal cues, at so-called *zero-acquaintance*. Here there appears to be interaction between the available information and the visibility of the trait being judged. Albright, Kenny, and Malloy (1988) found that, on the basis of physical appearance, Extraversion and Conscientiousness could be reliably rated, although the former appeared to be mediated by judgements of physical attractiveness. On the basis of transcribed interactions, self-other agreement has been found for ratings of Extraversion (and also its opposite Introversion) (Gifford and Hine, 1994).

Technology mediated communication

Whether or not subject and judge have prior knowledge of each other, technology has an impact on what information is available in a communicative situation. Zero-acquaintance judgements are perhaps particularly vulnerable to technological artifacts. For example, interviews conducted over the telephone were found to result in reduced self-interviewer and peer-interviewer agreement than face-to-face interviews (Blackman, 2002).

In a computer-mediated environment (CMC), the cues are reduced even further, and following one-on-one interactions in an internet chat room, consensus was found between judges for a target's Extraversion, Agreeableness, and Openness, whilst target-judge agreement was only found for Extraversion and Openness (Markey and Wells, 2002).

Impressions of personality formed following task-oriented synchronous computer-mediated communication found that they were less detailed but more intense compared with those from face-to-face communication. Specifically, in the CMC environment, judges seemed less able to rate their partners for Extraversion, Neuroticism, and Agreeableness. Across both environments, Conscientiousness, Agreeableness, and Extraversion were the most rateable (Hancock and Dunham, 2001).

Linguistic features of personality

By analysing our personality e-mail corpus, we have previously shown that Extraverts and Introverts produce characteristic language features (Gill and Oberlander, 2002). For a summary, see Figure 1.

These results are broadly consistent with—and in some cases more detailed than—the prior literature (eg. Nass, Moon, Fogg, and Reeves, 1995; Furnham, 1990; Berry, Pennebaker, Mueller, and Hiller, 1997). In particular, study of texts *written* about thoughts and feelings by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclu-

Surface Realisation Extraverts are more informal, use *hi*, and use looser punctuation (*!! or ...*). Introverts use *hello*.
Quantification Introverts show greater use of quantifiers (*for exaggeration?*); Extraverts are looser and less specific.
Social Devices Stylistic expressions such as *catch up* and *take care* indicate the Extravert's relaxed social style.
Self/Other Reference Introverts use more first-person singular (*i*), whereas Extraverts are more likely to use plural *we*.
Valence Introverts prominently use negations; Extraverts use words suggestive of positive affect.
Ability Extraverts are more confident and assertive (eg., *want-, able-, need-(to)*); Introverts are more tentative and timid (*trying-, going- (to)*).
Modality Extraverts are more strongly predictive than Introverts (eg., modal auxiliaries *will- vs. should- (be)*).
Message Planning/Expression Introverts prefer coordinating conjunctions (*and, but*), whereas only Extraverts use the subordinative *which* (*usually for evaluation?*).

Figure 1: Extravert and Introvert Language

Table 1: Summary of LIWC, MRC, and TTR multiple regression analyses.

Analysis	Independent Var.	β	R^2	p
LIWC	Inclusive Words	.28		
	Total First Person	.21	**.11**	**.0030**
MRC	Mean Concreteness	.33		
	Mean Brown Verbal Frequency	.27	**.14**	**.0004**
TTR	10 Word Measures	−.27	**.07**	**.0057**

Note: In each case, EPQ-R Neuroticism Score is the Dependent Variable. LIWC = Linguistic Inquiry and Word Count; MRC = Medical Research Council Psycholinguistic Database; TTR = Type-Token Ratio.

sive words, while using more social and positive emotion words (Pennebaker and King, 1999).

Extraversion is generally considered most relevant to communication, but Neuroticism also has implications for interaction (Mount *et al.*, 1998). Furthermore, Pennebaker and King (1999), using the Linguistic Inquiry and Word Count (LIWC) text analysis program, showed that broad psychological language categories are related to Neuroticism. For example, they found that when writing about thoughts and feelings, high Neurotics use more negative emotion words and fewer positive emotion words, along with other features in their factor 'Immediacy'.

Using multiple regression analysis, we have uncovered characteristic language usage patterns for Neuroticism in our e-mail corpus. Table 1 shows the results of these analyses, using LIWC data (Pennebaker and King, 1999), psycholinguistic properties from derived from the Medical Research Council (MRC) Psycholinguistic Database (Wilson, 1987), and a measure of lexical diversity, type-token ratio (TTR) (Bradac, 1990). (See also Gill and Oberlander, prep, for more details.)

We would expect a text characteristic of high Neuroticism to exhibit the following: In terms of LIWC features, we would expect words such as *with, and, include* (indicating inclusion) to be used, which are possibly indicative of the high Neurotic's desire for attachment or reassurance; first person pronouns, such as *I, me, we* again indicate a preoccupation with self, and may be related to our previous findings for low Extraverts (Introverts).

This relationship between Neurotics and Introverts again appears in an increased use of concrete words (for entities which can be sensed); for example, *table, spoon, girl*, rather than abstract words, like *thoughts, flavours, pains*. Given the relationship between Neuroticism and Brown Verbal Frequency, we suggest that high Neurotics show a preference for forms occurring frequently in speech, for example, *I, and, that*, rather than less common words such as *abject, suspicion, tether*. This preference for common words contributes towards the very low lexical density found in highly Neurotic texts, demonstrated by the high repetition over ten-word sections of text.

So, e-mail from Extraverts and Neurotics has characteristic linguistic features. Do judges with zero-acquaintance pick up on these features? We turn now to our rating experiment.

Method

Participants

The 30 judges were current students at the University of Edinburgh, or recent graduates (15 males, 15 females; mean age= 21.6 years, s.d.= 1.24). All were highly experienced e-mail users (rating themselves between 7 and 10 on a scale of 1–10; mean= 9.23, s.d.= 0.77), and naive raters of personality (none had previously taken part in personality rating experiments, although 3 had studied Psychology as part of their course). Participants received a nominal 'experimental expenses' payment for taking part.

Materials

Selection of Target Texts The target e-mail texts were selected from data previously collected (see Gill and Oberlander, 2002, for further details). These texts were composed 'to a good friend' to ensure they elicited a naturalistic expression of personality. Only the 105 'past' texts, detailing recent activities, were considered since these were generally slightly longer (each approximating 10 minutes of written communication; cf. Blackman, 2002). Six texts were chosen to represent a range of scores from the Extraversion, Neuroticism, and Psychoticism dimensions. Extreme high and low personality scores were deemed to be those greater than 1 standard deviation of the mean (Dewaele and Pavlenko, 2002), and two texts represented each of these. Additionally, two further texts were selected—one above and below the mean—to represent less extreme realisa-

tions of the trait (each between .5 and 1 s.d. of the mean). In each case, the scores for the other personality dimensions were controlled for, being $< \pm 1$ s.d. of the mean (in most cases $< \pm .5$ s.d.). This resulted in 6 texts for each dimension. Each e-mail text was anonymised by name substitution before use in the experiment.

Subjective Rating Methods Descriptions of the personality dimensions were presented to the participants before rating of the e-mail texts. These were taken from Eysenck and Eysenck (1991) (with minor re-wording to enhance general intelligibility), and participants were informed that they could refer back to them at any point during the experiment. Although it is more usual to rate personality using a standard set of personality questions, Sneed, McCrae, and Funder (1998) have found that 'most laypersons can easily grasp the nature of the factors and their behavioural manifestations and can spontaneously recognise their grouping when presented with clear exemplars'.

Each text was followed by a set of questions, with answers rated on a scale of 1–10, as follows. (i) How Extravert (or Emotionally Stable, or Tough-minded[1]) is the author of the e-mail? (ii) How easy was it to judge the author's personality? (iii) How informative were Topic, Vocabulary, and Style in judging personality? (iv) How similar is the author's personality to your own? Finally, subjects supplied 5 words describing the author's personality.

Procedure

Upon commencing the experiment, subjects were given a rating booklet prefixed with written instructions explaining that the experiment was investigating how author personality can be perceived through e-mail texts. It was emphasised that they should answer honestly and accurately, not spend too long thinking about each question, and instead concentrate on giving their initial response.

The target e-mail texts (described above) were then presented in random order within their representative dimension. Each set of dimension texts (P, E, or N) were presented using a Latin square technique to avoid ordering effects.

Following the rating of the texts, participants were asked to confirm that they are Native English Speakers, detail their experience of personality psychology, and rate their previous experience using e-mail. Participants were then asked to complete EPQ-R and NEO-PI personality questionnaires (both short forms), before being debriefed about the experiment.

[1]The terms 'Emotional Stability' and 'Tough-mindedness' have been used in preference to Neuroticism and Psychoticism when discussing these traits with participants (cf. Eysenck and Eysenck, 1991).

Table 2: Summary of inter-judge agreement.

Trait	Target-rater	Inter-rater	
	Aggr. r_s	Mean r_s	s.d.
Extraversion	.89*	.48	.17
Neuroticism	−.29	.31	.16

Note: Target-rater = correlation of target self-reports and rater judgement; Inter-rater = correlation of rater judgements with each other. Aggregate correlation is calculated from 30 raters. Mean r_s is the mean correlation across all raters. $^*p < .05$, two-tailed.

Table 3: Summary of similarity ratings.

Rater group	High trait texts			Low trait texts		
	n	U	p	n	U	p
High E	42			42		
vs **Low E**	47	647	.005	47	941	.702
High N	33			32		
vs **Low N**	56	813	.341	57	732.5	.121

Note: Observations (n) vary due to missing cases.

Results

For clarity here we discuss only the results for texts contrasted on the Extraversion and Neuroticism scales, and we focus on the subjective ratings and similarity ratings for these texts. Spearman correlation of target personality scores and subjective ratings aggregated across the 30 judges is shown in the second column of Table 2. Inter-rater agreement, and standard deviation, are shown in the following columns; these are calculated from the mean of each rater's mean Spearman correlation with each of the other raters. Since this is a mean correlation, no significance value is shown. For a description and discussion of further results, see Gill and Oberlander (2003).

Table 3 shows the Mann-Whitney U-tests (two-tailed) calculated from the similarity ratings for the judges grouped by High and Low Neuroticism and Extraversion for the texts grouped by these categories. Examination of the means for the High and Low Extraverts rating High Extravert texts, shows that the High Extraverts do indeed rate themselves as significantly more similar (5.71; s.d.= 1.92 vs 4.65; s.d.= 1.89).

Although not significant, the next strongest difference is found between the High and Low Neurotic similarity ratings of Low Neurotic texts. Comparison of the means shows that it is the High—rather than Low—Neurotic raters who see themselves as most similar to the Low Neurotic e-mail authors (5.00; $s.d. = 2.17$ and 4.34; $s.d. = 2.09$, respectively).

Discussion

Before discussing the subjective ratings in detail, it should be noted that there is a much greater level of target-rater agreement, than inter-rater agreement, for judgements of Extraversion.

Part of this increased agreement for target-rater judgements can be explained by the use of aggregated scores across raters. This is because they may 'reflect more accurately the consensus of how an individual is viewed'. The high number of raters (30) for each target apparently contributes towards the good agreement (cf. McCrae and Costa, 1987). In fact, even without aggregation of judgements before correlation (ie. calculating the mean across each rater's correlation with the target), the same pattern is still preserved (mean r_s E= .64; N= −.02).

Subjective Ratings

In the case of inter-rater judgements, both Extraversion and Neuroticism show a level of agreement greater than .3, which is regarded as the lower level of acceptability within personality research (McCrae and Costa, 1987). In Neuroticism's case, this level is only just reached; by contrast, Extraversion shows much greater agreement between judges.

In the case of target-rater judgements, however, there is a greater discrepancy between the traits. Extraversion shows a strong, significant, positive correlation, while Neuroticism shows a relatively weak, non-significant, negative relationship.

Similarity Judgements

The similarity ratings show that only the High and Low Extraverts rate their similarity to the High Extravert texts significantly differently. This confirms the observability of Extravert behaviour—even in an asynchronous CMC environment. Furthermore, this also lends some support to Funder's (1995) claim that Extraverts may make more accurate judges of personality—at least for Extraversion.

The tendency of High Neurotic judges to rate themselves similar to Low Neurotic authors contributes further towards the confused picture that exists for ratings of Neuroticism. Indeed, it may well be the High Neurotic raters who are clouding the picture for ratings of Neuroticism as a whole.

Interpretation

Taking these results together, the picture for Extraversion seems relatively clear. There is a high level of agreement between judges, and the judges tend to agree with the targets themselves. It seems safe to conclude that writers of e-mail messages do betray their level of Extraversion through their linguistic choices; and readers of e-mail messages can reliably infer the author's level of Extraversion from the text alone. This supports previous findings from the literature for well-acquainted raters (Gomà-i-Freixanet, 1997; McCrae and Costa, 1987), zero-acquaintance raters (Gifford and Hine, 1994; Albright *et al.*, 1988), and in computer-mediated communication (eg. Markey and Wells, 2002).

In the rating of Neuroticism, there was a low but evident level of agreement between judges, but not between judges and targets. This follows a trend of lower agreement for Neuroticism than for Extraversion found more generally (Colvin and Funder, 1991; Kolar *et al.*, 1996). However, this lack of perception ability appears particularly acute for zero-acquaintance (Gifford and Hine, 1994; Albright *et al.*, 1988) or CMC (Markey and Wells, 2002; Hancock and Dunham, 2001). Indeed, the fact that raters agreed amongst themselves for the ratings of Neuroticism appears to mirror Markey and Wells's (2002) findings for Agreeableness, since despite inter-rater agreement, they were unable to find target-rater agreement. Since raters were in a cue-impoverished environment, this may have resulted in their relying upon cues—apparently stereotypical of the trait—but inappropriate (Scherer, 1972).

The similarity ratings confirm the observability of particularly high Extravert authored texts, and also point to the expertise of Extraverted raters. Both subjective and similarity results for Neuroticism point to confusion on the whole, and possible distortion of this trait on the part of high Neurotic raters (Funder, 1995). Given the findings of John and Robbins (1993) regarding the role of evaluativeness in the assessment of this trait, caution may be advisable in the subjective personality rating of Neuroticism (note, however this effect was not present for another highly evaluative trait, Psychoticism; cf. Sneed *et al.*, 1998).

To summarise the position on Neuroticism, we return to Funder's Realistic Accuracy Model. There is no reason to consider that we had bad targets on this dimension; deception would have been revealed in the EPQ-R Lie Scale. In general, we do not have bad judges; they agreed with targets and each other when rating Extraversion. It is however, possible that highly Neurotic authors linguistically conceal the full extent of their Neuroticism and this would tend to lower its visibility. This may have led to the confusion of highly Neurotic judges in ratings of similarity. So, in fact the main difficulty seems to be that Neuroticism is a bad trait. It is held to be high in evaluativeness, and low in visibility. Our study has provided evidence that the trait affects the form of the e-mail texts. But the evidence is in terms of the concreteness of language, or in repetitiveness. While these may cause unconscious reactions in judges, the latter appear unable—or unwilling—to recruit them in their judgements.

Conclusion

We have shown that at zero-acquaintance, people are able to take asynchronous communication, and are still able to subjectively rate the degree of Extraversion of the author. There is also a relatively high level of agreement between judges in rating the target. In the case of Neuroticism, raters show a reasonable level of agreement with each other, but their perceptions of Neuroticism do not appear to

match up with the targets' self reports. So the asynchronous nature of e-mail seems to exacerbate the differences in the perception of personality traits.

Acknowledgements

Thanks to Elizabeth Austin and our anonymous reviewers for comments. This work was supported by the School of Informatics, University of Edinburgh, and the Economic and Social Research Council, UK.

References

Albright, L., Kenny, D., and Malloy, T. (1988). Consensus in personality judgements at zero acquaintance. *Journal of Personality and Social Psychology*, **55**(3), 387–395.

Baron, N. (1998). Letters by phone or speech by other means: the linguistics of email. *Language and Communication*, **18**, 133–170.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Blackman, M. (2002). The employment interview via the telephone: Are we sacrificing accurate personality judgements for cost efficiency. *Journal of Research in Personality*, **36**, 208–223.

Bradac, J. (1990). Language attitudes and impression formation. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 387–412. Wiley, Chichester.

Cheng, Y., O'Toole, A., and Abdi, H. (2001). Classifying adults' and children's faces by sex: Computational investigations of subcategorial feature encoding. *Cognitive Science*, **25**, 819–838.

Colvin, C. and Funder, D. (1991). Predicting personality and behaviour: A boundary on the acquaintance effect. *Journal of Personality and Social Psychology*, **60**, 884–894.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**, 265–324.

Eysenck, H. and Eysenck, S. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Funder, D. C. (1995). On the accuracy of personality judgement: A realistic approach. *Psychological Review*, **102**, 652–670.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gifford, R. and 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. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003). Rating e-mail extraversion at zero acquaintance. In *Proceedings of the 11th Biennial Meeting of the International Society for the Study of Individual Differences, Graz, Austria July 13-17, 2003*.

Gill, A. and Oberlander, J. (in prep.). Dictionary approaches to personality language. *in prep.*

Gomà-i-Freixanet, M. (1997). Consensus validity of the EPQ: Self-reports and spouse-reports. *European Journal of Psychological Assessment*, **13**(3), 179–185.

Hancock, J. and Dunham, P. (2001). Impression formation in computer-mediated communication. *Communication Research*, **28**, 325–347.

John, O. and Robbins, R. (1993). Determinants of interjudge agreement: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *J. Personality*, **61**, 521–551.

Kolar, D., Funder, D., and Colvin, C. (1996). Comparing the accuracy of personality judgements by the self and knowledgeable others. *J. Personality*, **64**, 311–337.

Markey, P. and Wells, S. (2002). Interpersonal perception in internet chat rooms. *Journal of Research in Personality*, **36**, 134–146.

McCrae, R. and Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, **52**, 81–90.

Mount, M., Barrick, M., and Stewart, G. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance*, **11**, 145–165.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *Int J Human-Computer Studies*, **43**, 223–239.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Scherer, K. (1972). Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *J. Personality*, **40**, 191–210.

Sneed, C., McCrae, R., and Funder, D. (1998). Lay conceptions of the Five-Factor Model and its indicators. *Personality and Social Psychology Bulletin*, **24**, 115–126.

Wilson, M. (1987). MRC psycholinguistic database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.

D.4 Oberlander and Gill (2004)

Language generation and personality: two dimensions, two stages, two hemispheres?

Jon Oberlander and **Alastair Gill**
School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW UK
{J.Oberlander|A.Gill}@ed.ac.uk

Abstract

We are interested in generating text in a way which helps convey the writer's personality. This has led us to consider the relationship between language production and personality from a Marrian perspective. We already have data to be covered at the computational level (comparative corpus analysis). We consider that findings at the implementation level (cognitive neuroscience) will help guide architectural explorations at the algorithmic level (computational linguistics). This position statement indicates the data and processing hypotheses which we have arrived at, and suggests that neurocognitive results concerning hemispheric asymmetry may be particularly relevant.

Personality and language production

Personality traits lie at the more temporally-stable and less intense end of scale of affective states and processes. There are a number of approaches to personality (Matthews and Deary, 1998). Two of the most prominent trait theories are the five factor model (McCrae and Costa, 1987), and Eysenck's three-factor PEN model (Eysenck and Eysenck 1991, Eysenck et al. 1985). These agree that two main factors are Extraversion (sociability) and Neuroticism (emotional stability). The Five Factor Model sees three further dimensions: Conscientiousness, Agreeableness and Openness; PEN arguably conflates these into one dimension, Psychoticism (tough mindedness). In what follows, we focus on the first two dimensions, common to both models.

In the past, simple approaches to our generation task have involved two steps. First, checking the literature on individual differences and language production (Pennebaker and King 1999, Berry et al. 1997, Groom and Pennebaker 2003, Campbell and Pennebaker 2003, Furnham 1990, Dewaele and Furnham 1999, Dewaele and Furnham 2000, Dewaele and Pavlenko 2002, Scherer 1979). Secondly, picking a number of features associated with a personality trait, and then ensuring that they either always or never appear in a language generation system's output. For instance, Nass et al. (1995) manipulated dominance (a facet related to Extraversion) by avoiding hedge-expressions such as *perhaps*, and ensuring that the system initiated pairs of turns.

However, we are interested in text generation, and the great majority of work on language in personality psychology has focussed on spoken language, and where it has considered written text, it has usually confined itself to counting occurrences in a text of words listed in a pre-defined dictionary (eg. Pennebaker and Francis 1999, Pennebaker and King 1999). (Although Dewaele has gone further, and analysed part-of-speech and lemmatised word frequencies.)

Yet there's clearly more to language generation than lexical choice. So one obvious way of improving systems designed to convey personality is to use more sensitive techniques to detect subtle yet pervasive language-personality patterns. That is exactly what we have been doing (while also testing how good judges are at perceiving personality from texts sampled from these corpora; cf. Gill and Oberlander 2003b). We have exploited more sensitive data-driven techniques from corpus linguistics, and compared n-grams (of words and punctuation, and additional meta-linguistic information) of various lengths, as well as part of speech and semantic analysis and psycholinguistic measures on word use (Rayson 2003, Argamon et al. 2003, Aarts and Granger 1998, Milton 1998, Thomas and Wilson 1996, Rayson et al. 1997, Damerau 1993, Coltheart 1981). We have applied these techniques to corpora collected from subjects whose personality is measured via Eysenck's EPQ instrument (cf. Dewaele and Furnham 2000, Dewaele and Pavlenko 2002). This has allowed us to gather and analyse a corpus of email messages (amounting to 65,000 words from 105 subjects). The techniques and tools from computational corpus linguistics have allowed us to uncover more subtle relations between personality and language than has hitherto been possible (Gill and Oberlander 2002, 2003a, 2003b, 2003c; Gill 2003).

Language data to be explained

We have uncovered numerous surface cues to Extraversion, Neuroticism and Psychoticism; to indicate the kind of work we have been carrying out, this section briefly rehearses some of the features which appear to vary by these dimensions. On the one hand, we carried out dictionary-based top-down regression analysis of our corpus of e-mail texts. On the other, we carried out bottom-up comparative analysis of sub-corpora, to isolate patterns of words (or parts of speech) that were distinctive of personality types. We will

touch briefly on the first type of analysis, and go into more detail on the second.

Results of top-down analyses

A series of multiple regression analyses were carried out on the corpus, relating personality scores to prevalence of terms in either Pennebaker and Francis' LIWC dictionary, or in the MRC Psycholinguistic database (Coltheart 1981).

Taking Extraversion and the LIWC dictionary first, comparing higher with lower Extraversion, we found fewer number expressions and more words overall ($R^2 = .08, p < 0.05$). With the MRC dictionary, we found lower concreteness overall ($R^2 = .05, p < 0.05$). The former result fits with the general finding that Extraverts speak more, and are generally less precise. The latter finding suggests that they also prefer less specific, more abstract language. This would fit the idea that the need to seize or maintain the conversational floor leads to high Extraverts putting less effort into precise lexical choice. See Gill and Oberlander (2002) for more details.

Turning to Neuroticism, using LIWC again and comparing higher with lower Neuroticism, we found more 'Inclusive' words and more first person references ($R^2 = .11, p < 0.01$). The use of inclusives like *with, and* and *include* is arguably consistent with a desire for attachment, and the use of first person with a preoccupation with the self. With the MRC dictionary, we found higher concreteness overall, and higher mean verbal frequency ($R^2 = .14, p < 0.001$). This suggests fairly down-to-earth lexical choices, and language that is more speech-like or immediate, overall. The latter feature is consistent with another of our findings, to the effect that higher Neuroticism is associated with lower lexical density (and hence, repetitiveness). See Gill and Oberlander (2003b) for more details.

Results of bottom-up analyses

The original e-mail corpus of texts was divided into stratified sub-corpora. High and Low personality group samples were created by splitting them at greater than 1 standard deviation above and below the EPQ-R score for each dimension. The additional requirement was made that authors had to be *within* 1 standard deviation on the dimensions other than the one for which they were extremely high or low. Additionally, all texts which were within 1 standard deviation across *all* personality dimensions were assigned to the personality 'neutral' Mid sub-corpus. Thus, on any dimension, we have three groups to compare (High, Mid, and Low).

The primary goal is to identify words (unigrams) or strings of words (n-grams) which form reliable collocations for one group, but not for another; these can then be considered *distinctive* collocations. Here we present the results from the three-way *lemmatised* analysis for Extraversion and Neuroticism, in Tables 1 and 2. By lemmatising (or stemming), minor variants of words can be collapsed together, increasing the power of the analysis. In such a processed corpus words such as *play*, *plays*, *played*, or *playing*, are all realised in the base form of the verb: *play*. More importantly, in our data there are instances of proper nouns being used, for example, names of places (*Edinburgh*), days of the week (*Saturday*), or names of people (*Dave*), with these providing too much specificity to allow broader patterns of language usage to emerge, or for the results to be easily generalised. The corpora were pre-processed using the CLAWS tagger (Rayson 2003) to give vertical-output lemmatised words and part-of-speech (POS) tags. Additional scripts were then used to convert this into the form of lemmas, and in the case of the features being a proper noun, this was replaced by the POS tag.

To identify robust collocations in the sub-corpora, and then to identify those which distinguish one group from another, we start by specifying that a feature should exhibit a frequency in one of the three groups of at least 5 occurrences, and ordering the features by log-likelihood (G^2) value. Because we only examine expected frequencies of 5 or more—which compare more reliably with the χ^2 distribution—we can here present results with a critical value of 10.83 or greater, taking this to be equivalent to reaching $p \leq 0.001$ significance, and those results with a critical value of 15.13 or greater are taken to be equivalent to reaching $p \leq 0.0001$ significance (cf. Rayson 2003 on adjustments which have to made if frequencies of less than 5 are to be considered). Note that if a feature is overused by the Mid group, we do not report the G^2 for this, and in cases where the relative-frequency ratio or G^2 is not available, we replace this by '-'.

Tables 1 and 2 contain a lot of low-level data. Note that a feature (such as the collocation [*will be*]) may be underused by one sub-group, compared to the two other groups, or over-used by one group compared to the others. To help characterise the linguistic habits of a group at one or other end of a personality dimension, we can consider both which n-grams they over-use, and also which n-grams are underused by the group at the other end of the dimension. Figure 1 presents just such a digest, for Extraversion and Neuroticism.

Putting the content of Figure 1 into other words, we can say that there are a number of reliable collocations which appear to be distinctive of the personality groups under discussion.

Punctuation is surprisingly differentiated. Multiple punctuation (exclamation in particular, but also the multiple dots of ellipsis) is particularly associated with High-N, and also with High-E. Single hyphens are associated with Low-E; commas with Low-N.

Several collocations involving the first person singular are apparent for High-E, and a couple for Low-E ([*i play*], [*that i*]); for High-N, we find [*well i*] and [*i ca*], where the latter lemmatised bigram represents the initial subpart of *I can't* or *I couldn't*. There are none for Low-N. Interestingly, both High- and Low-N use first person *plural* less than the Mid reference group.

Expressions concerning ability or modality appear in different patterns for the groups. High-E have [*i will*] and [*will have*]; Low-E have [*be supposed to be*]; High-N have [*i ca*] and [*have to*]; Low-N have [*will be*], [*have be*], and [*have not*].

NPs appear in distinctive collocations for some groups.

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
play	1	3	0.0004	2	0.0002	14	0.0052	2.42	30.31	0.08	0.97	35.63****	22.53****			+
get a	2	15	0.0021	0	0	5	0.0019	-	-	1.12	28.86****	16.74****	0.05		−	
be so	3	14	0.0020	0	0	0	0	-	-	-	26.93****	-	8.88**	+		
i play	4	0	0	0	0	7	0.0026	-	-	-	-	23.43****	18.24****			+
christmas ⟨p⟩	5	10	0.0014	0	0	3	0.0011	-	-	1.24	19.24****	10.04**	0.11		−	
year ⟨p⟩	5	10	0.0014	0	0	0	0	-	-	-	19.24****	-	6.34*	+		
week ⟨p⟩	6	0	0	42	0.0036	7	0.0026	-	0.72	-	-	0.69	18.24****	−		
i will	7	28	0.0039	35	0.0030	0	0	1.29	-	-	1.02	-	17.76****			−
⟨p⟩ take	8	9	0.0013	0	0	0	0	-	-	-	17.31****	-	5.71*	+		
with i	8	9	0.0013	0	0	0	0	-	-	-	17.31****	-	5.71*	+		
be supposed	9	0	0	0	0	5	0.0019	-	-	-	-	16.74****	13.03***			+
that be	9	0	0	0	0	5	0.0019	-	-	-	-	16.74****	13.03***			+
bread	10	0	0	3	0.0003	6	0.0022	-	8.66	-	-	9.87**	15.63****			+
NP1 for	11	8	0.0011	0	0	0	0	-	-	-	15.39****	-	5.07*	+		
i really	11	8	0.0011	0	0	0	0	-	-	-	15.39****	-	5.07*	+		
then i	12	7	0.0010	0	0	3	0.0011	-	-	0.87	13.47***	10.04**	0.04		−	
day ⟨p⟩	12	7	0.0010	0	0	0	0	-	-	-	13.47***	-	4.44*	+		
will have	12	7	0.0010	0	0	0	0	-	-	-	13.47***	-	4.44*	+		
NP1 and	13	21	0.0029	22	0.0019	0	0	1.54	-	-	2.00	-	13.32***			−
be supposed to	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.75***	13.03***			+
be supposed to be	14	0	0	1	0.0001	5	0.0019	-	21.67	-	-	11.75***	13.03***			+
supposed to be	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.75***	13.03***			+
supposed	14	0	0	1	0.0001	5	0.0019	-	21.65	-	-	11.74***	13.03***			+
supposed to	14	0	0	1	0.0001	5	0.0019	-	21.66	-	-	11.74***	13.03***			+
fairly	14	0	0	4	0.0003	5	0.0019	-	5.41	-	-	6.03*	13.03***			+
⟨p⟩ although	14	0	0	7	0.0006	5	0.0019	-	3.09	-	-	3.34	13.03***	−		
that i	14	0	0	27	0.0023	5	0.0019	-	0.80	-	-	0.22	13.03***	−		
and i	15	20	0.0028	44	0.0038	0	0	0.73	-	-	1.35	-	12.69***			−
and NP1	16	19	0.0027	13	0.0011	0	0	2.36	-	-	5.84*	-	12.05***	+		
take	17	25	0.0035	13	0.0011	7	0.0026	3.11	2.33	1.33	11.79***	2.93	0.48		−	
cool ⟨p⟩	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
from the	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
of it	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
today ⟨p⟩	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		
what i	18	6	0.0008	0	0	0	0	-	-	-	11.54***	-	3.81	+		

Table 1: Lemmatised n-gram analysis, Extraversion.

Note. $* = p < .05$, $** = p < .01$, $*** = p < .001$, $**** = p < .0001$, $df = 1$. In Use columns, + indicates over-use, − indicates under-use.

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⟩ it	1	0	0	45	0.0039	38	0.0053	-	1.37	-	-	2.00	34.37****	−		
⟨p⟩ ⟨p⟩ ⟨p⟩ ⟨p⟩	2	16	0.0039	2	0.0002	0	0	22.67	-	-	31.66****	-	32.37****	+		
NP1 ⟨p⟩	3	16	0.0039	57	0.0049	0	0	0.80	-	-	0.68	-	32.36****			−
⟨p⟩ ⟨p⟩ ⟨p⟩ ⟨p⟩	4	21	0.0051	10	0.0009	2	0.0003	5.95	0.32	18.37	23.50****	2.65	30.70****	+		
⟨p⟩ as	5	0	0	0	0	14	0.0020	-	-	-	-	26.97****	12.66***			+
⟨p⟩ ⟨p⟩ ⟨p⟩	6	43	0.0105	56	0.0048	23	0.0032	2.18	0.67	3.27	13.89***	2.85	22.43****	+		
film	7	11	0.0027	2	0.0002	0	0	15.58	-	-	19.60****	-	22.25****	+		
i go	8	8	0.0020	0	0	8	0.0011	-	-	1.75	21.50****	15.41****	1.23		−	
that be	9	7	0.0017	0	0	11	0.0015	-	-	1.11	18.81****	21.19****	0.05		−	
⟨p⟩ he	9	0	0	0	0	11	0.0015	-	-	-	-	21.19****	9.95**			+
⟨p⟩ well	10	21	0.0051	12	0.0010	10	0.0014	4.96	1.35	3.67	20.43****	0.48	12.53***	+		
will be	11	0	0	39	0.0034	21	0.0029	-	0.87	-	-	0.26	18.99****	−		
have be	11	0	0	37	0.0032	21	0.0029	-	0.92	-	-	0.10	18.99****	−		
all the	12	9	0.0022	9	0.0008	0	0	2.83	-	-	4.67*	-	18.20****	+		
⟨p⟩ so	13	0	0	34	0.0029	20	0.0028	-	0.95	-	-	0.03	18.09****	−		
be in	14	0	0	0	0	9	0.0013	-	-	-	-	17.34****	8.14**			+
⟨p⟩ ⟨p⟩ well	15	12	0.0029	4	0.0003	6	0.0008	8.50	2.43	3.50	16.67****	1.94	6.78**	+		
film be	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
the film	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
well i	16	6	0.0015	0	0	0	0	-	-	-	16.12****	-	12.13***	+		
year ⟨p⟩	17	0	0	0	0	8	0.0011	-	-	-	-	15.41****	7.24**			+
to do	18	0	0	11	0.0009	17	0.0024	-	2.50	-	-	5.80*	15.37****			+
to the	19	0	0	24	0.0021	16	0.0022	-	1.08	-	-	0.06	14.47***	−		
though ⟨p⟩	20	7	0.0017	12	0.0010	0	0	1.65	-	-	1.06	-	14.16***			−
to NP1	20	7	0.0017	25	0.0022	0	0	0.79	-	-	0.31	-	14.16***			−
we	21	18	0.0044	119	0.0103	47	0.0066	0.43	0.64	0.67	13.74***	7.13**	2.21		+	
still	22	0	0	30	0.0026	15	0.0021	-	0.81	-	-	0.45	13.57***	−		
about it	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
it do	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
rowing	23	0	0	0	0	7	0.0010	-	-	-	-	13.48***	6.33*			+
and she	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
the film be	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
the time	24	5	0.0012	0	0	0	0	-	-	-	13.44***	-	10.11**	+		
experiment	24	5	0.0012	0	0	2	0.0003	-	-	4.37	13.44***	3.85*	3.54		−	
⟨p⟩ which	25	0	0	15	0.0013	14	0.0020	-	1.51	-	-	1.22	12.66***	−		
have not	25	0	0	32	0.0028	14	0.0020	-	0.71	-	-	1.20	12.66***	−		
NP1 and	25	0	0	22	0.0019	14	0.0020	-	1.03	-	-	0.01	12.66***	−		
stuff	26	3	0.0007	3	0.0003	13	0.0018	2.83	7.02	0.40	1.56	12.48***	2.38			+
⟨p⟩ ⟨p⟩ we	27	4	0.0010	34	0.0029	5	0.0007	0.33	0.24	1.40	5.73*	12.45***	0.25		+	
i ca	28	6	0.0015	10	0.0009	0	0	1.70	-	-	1.00	-	12.13***			−
of time	29	3	0.0007	0	0	6	0.0008	-	-	0.87	8.06**	11.56***	0.04		−	
get a	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
go on	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
party ⟨p⟩	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
stuff ⟨p⟩	29	0	0	0	0	6	0.0008	-	-	-	-	11.56***	5.43*			+
have to	30	21	0.0051	30	0.0026	11	0.0015	1.98	0.59	3.34	5.47*	2.35	11.24***	+		
thesis	31	6	0.0015	1	0.0001	3	0.0004	16.99	4.86	3.50	10.99***	2.24	3.39	+		
⟨p⟩ the	32	0	0	16	0.0014	12	0.0017	-	1.21	-	-	0.26	10.85***	−		
he be	32	0	0	22	0.0019	12	0.0017	-	0.88	-	-	0.12	10.85***	−		
well ⟨p⟩	32	0	0	18	0.0016	12	0.0017	-	1.08	-	-	0.04	10.85***	−		

Table 2: Lemmatised n-gram analysis, Neuroticism.

Note. $* = p < .05$, $** = p < .01$, $*** = p < .001$, $**** = p < .0001$, $df = 1$. In Use columns, + indicates over-use, − indicates under-use.

High Extraverts

[with i], [i really], [what i], [i will], [will have]; [NP for], [and NP], [NP and], [and i]; [be so], [from the], [of it]; [today ⟨p⟩], [day ⟨p⟩], [year ⟨p⟩]; [⟨p⟩ take], [cool ⟨p⟩]. Also (not apparent in the lemmatised analysis): use of ellipsis; double exclamation.

Low Extraverts

[play] and [i play]; the n-grams composing [be supposed to be]; [fairly], [⟨p⟩ although]; [that be], [that i]; [week ⟨p⟩]; [bread]. Also (not apparent in the lemmatised analysis): use of hyphens; use of new clauses.

High Neurotics

[⟨p⟩ well], [⟨p⟩⟨p⟩ well], [well i]; the n-grams composing [the film be]; [though ⟨p⟩], [i ca], [have to]; [NP ⟨p⟩], [to NP]; [the time], [thesis], [all the], [and she].

Mid Neurotics

Characteristically using n-grams involving [we] more than either High or Low.

Low Neurotics

[⟨p⟩ as], [⟨p⟩ he], [⟨p⟩ it], [⟨p⟩ so], [⟨p⟩ which], [⟨p⟩ the]; [well ⟨p⟩]; [be in], [get a], [to do], [it do], [go on], [will be], [have be], [have not], [he be]; [NP and]; [about it], [to the]; [year ⟨p⟩]; [party ⟨p⟩], [stuff ⟨p⟩], [stuff], [still], [rowing]. Also (not apparent in the lemmatised analysis): use of commas; use of new clause followed by [also].

Figure 1: Summary of tokenised, lemmatised n-gram analysis: characteristic language

High-E have [*NP for*], [*and NP*] and [*NP and*], showing use of conjoined NPs; High-N have preposition-phrase forming [*to NP*], as well as clause-final NPs (where we interpret ⟨p⟩ as indicating (at least) clause-level punctuation); Low-N have one of the High-E NP+*and* patterns.

Temporal expressions are also distinctive: clause-final *today, day, year* for High-E; clause-final *week* for Low-E; Low-N also have clause-final *year*, as well as *party* (which can be considered an event).

N-grams indicating that a word typically occurs clause-initially are a special characteristic of Low-N, and cover *as, he, it, so, which* and *the*. Low-N also have one bigram involving clause-final *well*; none of High-N's three collocations for *well* are clause-final.

Finally, considering the phenomenon of hedging, it is notable that low-E use [*fairly*] and clause-initial *although*; the only other such connective collocation is clause-final *though*, used by High-N.

In passing, we note that we also carried out stratified comparisons of n-grams of parts-of-speech; for those analyses, we considered the sequences of POSs, and once more found that there were robust associations between personality scores and the over- or under-use of particular patterns of POSs. There is no room to report these here, but we touch on one aspect of the POS results in the final section of this paper.

The issue now is: how can we use these results to guide language production in an automatic natural language generation system?

Marr's levels

But before trying to use these results to control more effectively personality projection in generation systems, it is useful to step back and take a cognitive science perspective on these results. Marr (1982) distinguished three main levels of investigation: computational, algorithmic, and implementational. Roughly, these correspond to determining: what is being computed; how it is computed; and where it is computed.

Building a system which produces the right behaviour can, of course, be achieved by establishing the computation-level specification, and meeting it. A personality-oriented generator could be built which has nothing to do with human personality, so long as it gets the surface behaviour right.

But we are interested in architectural possibilities at the algorithmic level. What mechanisms underlie the surface-level productions? If High-Neurotic and Low-Neurotic text differ systematically, the question is: at which stage in a natural language generation system do the representations or processes in a high-neurotic generator differ from those in a Low-Neurotic generator?

On the basis of our findings to date, we have hypotheses about the the algorithmic level; they can be framed in terms of Levelt's (1989) architecture for human language production, and also in terms of fairly standard natural language generation systems.

First: Extraversion finds its effects at the stages of formulation (surface realisation). That is, the process and representations used in realisation differ between high and low Extraverts. Table 1 furnishes some examples supporting this, and as noted above, we have also found more generally that High-Extraverts' tendency to use more words may be counter-balanced by a tendency for those words to be less lexically specific (Gill and Oberlander 2002).

Secondly, Neuroticism finds its effects at the stage of conceptualisation (content selection). That is, the process and representations used in content selection differ between high and low Neurotics. Some of the evidence for this lies in linguistic patterns in Table 2, and in work suggesting that Neurotics tend to select more negative content, as well as more self-involving content (cf. Gill 2003).

To determine whether these hypotheses are correct, we need to actually specify the differences in detail, and parametrise our generators to produce differing linguistic behaviours. But before doing this, we should again consider Marr's levels. Whatever we claim is going on at an algorithmic level, it must at least be consistent with what is happening at an implementational level.

So, the critical question is: is what is now known about human implementation consistent with the hypotheses that Extraversion primarily affects surface realisation, and Neuroticism primarily affects content selection?

Implementational evidence

Eysenck's PEN model explicitly makes biological claims: there, Extraversion is related to levels of cortical arousal, Neuroticism to activation thresholds in the limbic system, and Psychoticism to mechanisms underlying aggression. But these claims are not prima facie consistent with our hypotheses. However, by drawing on two kinds of work in recent cognitive neuroscience, we would claim that what is now known about human implementation is at least consistent with our hypotheses.

On the one hand, there is work on hemispheric asymmetry and emotion. There is much to discuss here. We can only scratch the surface by pointing to the work of Davidson and colleagues (for instance, Davidson 1992, 2001; Davidson and Irwin 1999, Davidson and Rickman 1999). It has been held that the left cerebral hemisphere is responsible for approach behaviours, and the right for withdrawal behaviours. Evidence comes from unilateral lesion studies and imaging studies with normals. For instance, with prefrontal cortex lesions, it appears that while left lesions leave subjects with excessive withdrawal, right lesions leave subjects with a tendency to excessive approach. Or children with separation anxiety tend to show higher levels of right hemisphere prefrontal activity. Converging evidence suggests that circuits connecting the prefrontal cortex with the amygdala are associated with positive affect (in the left hemisphere), and negative affect (in the right hemisphere).

On the other hand, there is work on hemispheric asymmetry and language processing. Again, there is much to discuss; again, we only scratch the surface by pointing to the work of Chiarello and colleagues (for instance, Chiarello and Richards 1992; Chiarello et al. 2001). While areas within the left hemisphere (particularly Broca's area) have long been acknowledged to play a role in language processing—and sequencing behaviour more generally—the right hemisphere's role is less well-understood. By presenting word stimuli to the left or right visual fields, it is possible to probe the differential contributions of the left and right hemispheres to language interpretation and generation. For instance, in semantic interpretation, the right hemisphere appears to hold broad/inferential associations, while the left hemisphere narrows down the field. Or in generation, it has been argued that the right hemisphere may help "activation of multiple responses". There is thus some evidence to suggest that the left hemisphere is most important for timing and surface ordering, while the right hemisphere is important for intonation and deeper semantic processing.

Pulling these two strands of work together, we can draw the following speculative conclusions. Language production processes involving semantic associations are linked to the right hemisphere, which is responsible for withdrawal behaviours, and negative emotionality. Language production processes involving surface sequencing are linked to the left hemisphere, which is responsible for approach behaviours, and positive emotionality. Thus, if levels of Extraversion are associated with left hemisphere activity, it is reasonable to assume that this affects surface realisation; and if levels of Neuroticism are associated with right hemisphere activity, it is reasonable to assume that this affects content selection.

Consequences and challenges

This is all rather speculative. Even so, there are some obvious wrinkles that need to be ironed out.

First, there is a question as to how to fit language variation associated with other personality dimensions. For instance, we have found that High-Psychoticism is associated with high levels of use of otherwise low-frequency (unusual) words. Where in language generation—and the brain—does Psychoticism have its effects? According to Eysenck, Psychoticism is associated with trait aggression. Evidence from behavioural genetics suggests the situation may be complex, but that levels of stress hormones and aggression may indeed implicated. Still, this suggests that localisation (hemispheric or otherwise) of the effects of Psychoticism might not be feasible or desirable. Indeed, behaviour genetics may generally push us towards accounts in which all the traits are associated with overall efficiency of production or uptake of neurotransmitters—rather than particular brain areas.

Secondly, there is the fact that some of the language effects of the first two dimensions do not fit quite so neatly into the hemispheres. High-Neuroticism, for instance, does not lead to greater frequency of adverb use in our corpus; but it does lead to adverbs being placed more "promiscuously". That looks like a surface realisation behavioural variation, not a content selection variation.

We aim to explore a two-dimension, two-stage, two-hemisphere architecture in sufficient detail to address these problems. In the first instance, we can treat the computational model as a simulation, and determine to what extent changes in the values of parameters at a given generation stage produce behavioural variation of the kind that has been observed in human subjects. However, thoroughly testing the hypotheses requires more work on the direct relations between asymmetry and affect, and between asymmetry and language production—and on the perhaps less direct relation between affect and production. Methods for probing the affect–production link go well beyond corpus studies, and could involve a range of techniques short of imaging or lesion studies. With Annabel Harrison, we have carried out some initial psycholinguistic work on interpersonal priming, and while the results are not yet ready for publication, they are promising.

Thus, the Marrian perspective is helpful: implementation-level findings (from cognitive neuroscience) give us confidence that a possible algorithmic-level solution (from computational linguistics) is worth pursuing in greater depth. That solution is not necessary to meeting the computational-level specification (from comparative corpus analysis), but it is both sufficient and stimulating. We have no doubt that computational linguistics has much to learn from cognitive neuroscience.

Acknowledgements

Our thanks to the organisers, and our anonymous reviewer for helpful suggestions on presenting our approach to the Spring Symposium audience. The second author gratefully acknowledges studentship support from the UK Economic and Social Research Council and the School of Informatics.

References

Aarts, J. and Granger, S. 1998. Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In S. Granger ed. *Learner English on Computer*. Longman, London, pp. 132–141.

Argamon, S., Koppel, M., Fine, J. and Shimoni, A. 2003. Gender, genre, and writing style in formal written texts, *Text*, to appear.

Berry, D.S., Pennebaker, J.W., Mueller, J.S, Hiller, W.S. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Campbell, R.S. and Pennebaker, J.W. 2003. The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, **14**, 60–65.

Chiarello, C. and Richards, L. 1992. Another look at categorical priming in the cerebral hemispheres. *Neuropsychologica*, **30**, 381–392.

Chiarello, C, Liu, S., Kacinik, N. and Shears, C. 2001. Cerebral asymmetries for verb generation. Presented at the 8th Annual Meeting of the Cognitive Neuroscience Society, New York, 2001.

Coltheart, M. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**, 497–505.

Costa, P. and McCrae, R. R. 1992. *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Eysenck, H. and Eysenck, S. 1991. *The Eysenck Personality Questionnaire-Revised.* Hodder and Stoughton, Sevenoaks.

Damerau, F. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, **29**, 433 448.

Davidson, R.J. 1992. Emotion and affective style: hemispheric substrates. *Psychological Science*, **3**, 39–43.

Davidson, R.J. 2001. Towards a biology of personality and emotion. In Damasio, Harrington, Kagan, McEwen, Moss and Shaikh (eds.) *Unity of Knowledge: The Convergence of Natural and Human Science*, pp191–207. New York: New York Academy of Sciences

Davidson, R.J and Irwin, W. 1999. The functional neuroanatomy of emotion and affective style. em Trends in Cognitive Science, **3**, 11-21.

Davidson, R.J. and Rickman, M.D. 1999. Behavioral inhibition and the emotional circuitry of the brain: stability and plasticity during the early childhood years. In Schmidt and Schulin (eds.) *Extreme Fear, Shyness and Social Phobia*, pp67–87. New York: OUP.

Dewaele, J.-M., and Furnham, A., 1999. Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49(3)**, 509-544.

Dewaele, J.-M., and Furnham, A., 2000. Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences*, **28**, 355-365.

Dewaele, J.-M., and Pavlenko, A., 2002. Emotion vocabulary in interlanguage. *Language Learning*, **52(2)**, 265-324.

Eysenck, S., Eysenck, H., and Barrett, P. 1985. A revised version of the psychoticism scale. *P*ersonality and Individual Differences, **6**, 21–29.

Gill, A. 2003. *Personality and Language: The projection and perception of personality in computer-mediated communication*. PhD Thesis, School of Informatics, University of Edinburgh.

Gill, A. and Oberlander, J. 2002. Taking care of the linguistic features of Extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp363–368. Fairfax, VA, August 2002.

Gill, A. and Oberlander, J. 2003a. Expression and perception of emotionality in e-mail discourse. Paper presented at the *13th Annual Meeting of the Society for Text and Discourse*. Madrid, Spain, June 26–28, 2003.

Gill, A. and Oberlander, J. 2003b. Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself, but Neuroticism is more of a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*. Boston, MA, July 31–August 2, 2003.

Gill, A. and Oberlander, J. 2003c. Looking forward to more Extraversion with N-grams. In L. Lagerwerf, W. Spooren and L. Degand (eds.) *Determination of Information and Tenor in Texts: Multidisciplinary Approaches to Discourse 2003*, pp125–137. Amsterdam: Stichting Neerlandistiek VU Amsterdam.

Groom, C.J. and Pennebaker, J.W. 2003. Words. *Journal of Research in Personality*, **36**, 615-621.

Levelt, W. 1989. *Speaking: From Intention to Articulation.* Cambridge, MA.: MIT Press.

Matthews, G. and Deary, I.J. 1998. *Personality Traits.* Cambridge: CUP.

Milton, J. 1998. Exploring L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (ed.) *Learner English on Computer*. Longman, London, pp. 186–198.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. 1995. Can computer personalities be human personalities? *Int. J Human-Computer Studies*, **43**, 223–239.

Pennebaker, J.W. and Francis, M.E. 1999. *Linguistic Inquiry and Word Count (LIWC).* LEA Software and Alternative Media, Inc. Lawrence Erlbaum Associates, Inc. New Jersey, USA.

Pennebaker, J.W. and King, L.A. 1999. Linguistic styles: language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

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

Scherer, K.R., 1979. Personality markers in speech. In K.R. Scherer H. Giles, editors, *Social markers in speech*, pages 147-209. Cambridge University Press, Cambridge.

Thomas, J. and Wilson, A. 1996. Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short (eds.) *Using corpora for language research.* Longman, London, pp. 92–109.

D.5 Gill, et al. (to appear)

Interpersonality: Individual differences and interpersonal priming

Alastair J. Gill (A.Gill@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Annabel J. Harrison (annabelh@cogsci.ed.ac.uk)
School of Philosophy, Psychology, and Language Sciences, University of Edinburgh
7 George Square, Edinburgh, EH8 9JZ UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We study how Extraversion and Neuroticism influence people's language production in interpersonal interactive situations. A priming study used confederate priming methodology to investigate syntactic priming behaviour. We expected that Extravert sociability would be related to the strength of priming effects, although Neurotic emotionality might also have an effect. Results indicate that Extraversion has no effect, but Neuroticism does have an effect. We discuss possible reasons and suggest further experimentation to investigate this finding. Implications and applications of this work are outlined.

Personality and interaction

Individuals differ in the way they speak and write. Some of those differences are systematic, and can be attributed to apparently deeper differences, such as personality traits, like Extraversion and Neuroticism (or Emotional Stability). Level of Extraversion is intuitively related to sociability and communication, and this is expressed through interpersonal behaviour. However, level of Neuroticism appears to be more related to anxiety and inward focus, and thus having greater influence on solo behavior. In the past, it has been found that both these personality traits do significantly influence an individual's language production behaviour in a variety of contexts (Pennebaker and King, 1999; Dewaele and Furnham, 1999). Recent work has investigated e-mail text, and suggested that even in that genre, there are characteristic sequences of words associated with each end (High or Low) of both dimensions (Extravert or Neurotic) (Gill and Oberlander, 2002, 2003b).

The majority of work on the relations between personality and language production has studied monologue only. Yet most everyday language occurs in the context of interpersonal interaction. So here, we aim to investigate the role of personality upon language use in a dialogue setting.

Studies of conversational behaviour have demonstrated that individuals align with their interlocutors on a number of levels (Pickering and Garrod, in press). The phenomena have been examined from both social and cognitive perspectives. On the social side, a key focus of interest is cooperation and audience design. On the cognitive side, a key focus is coordination and interpersonal priming.

For example, sociolinguistic studies have shown that speakers adopt accent or dialectal variation or a level of lexical density appropriate to their audience. This variation operates at phonological, lexical, and syntactic levels (Labov, 1972; Coupland, 1980; Bell, 1984; Bradac and Wisegarver, 1984). Audience design is regarded as a relatively conscious process over which the speaker has a certain amount of control. It may be a result of co-operativity, affiliation, or willingness to take another's perspective (Haywood, Pickering, and Branigan, 2003).

By contrast, from a cognitive perspective, coordination is viewed as an artifact of the underlying language production mechanisms. For example, it has been argued that references from the comprehension system are recycled to provide output for the production system (Pickering and Garrod, in press). Alignment is found at the lexical level (Brennan and Clark, 1996; Branigan, Pickering, and Cleland, 2000), the conceptual level (Garrod and Doherty, 1987), and the syntactic level (Pickering and Branigan, 1998). Unlike cooperation, such coordination is considered to be largely subconscious.

Coordination therefore provides a more direct insight into underlying processing abilities, and is less prone to outside influence. In approaching the study of personality in dialogue, we therefore use an interpersonal priming paradigm. At the outset, our question is very general: Can differences in interpersonal priming be attributed to personality?

To make this question more specific—and to attempt to answer it—the rest of this paper is structured as follows. First, we introduce a little more background on personality theory. Then, we frame a possible explanation of recent findings on the relations between Extraversion, Neuroticism and language production; this leads to two hypotheses concerning the possible relation between personality and interpersonal priming. We then present the priming experiment which tested these hypotheses. The results were somewhat unexpected, and we conclude by discussing their implications.

Overview

There are a number of approaches to personality (Matthews and Deary, 1998). Two of the most prominent trait theories are the five factor model (Costa and McCrae, 1992), and Eysenck's three-factor PEN model (Eysenck, Eysenck, and Barrett, 1985; Eysenck and Eysenck, 1991). These agree that two main factors are Extraversion (sociability) and Neuroticism (emotional stability). The Five Factor Model sees three further dimensions: Conscientiousness, Agreeableness and Openness; PEN arguably conflates these into one dimension, Psychoticism (tough mindedness). In what follows, we focus on the first two dimensions, common to both models.

The traits can be summarised thus: A typical Extravert tends to be sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert (Low Extravert) is quiet, retiring, reserved, plans ahead, and dislikes excitement; A typical High Neurotic tends to be an anxious, worrying, moody individual. A typical Low Neurotic tends be calm, even-tempered and relaxed (Eysenck and Eysenck, 1991).

Personality and language

Work on personality and language behaviour has studied a range of features. For instance, Extraverts are regarded as talking louder (Scherer, 1978), demonstrating a higher speech rate (Siegman, 1987), and they show less hesitation, but make a higher proportion of semantic errors (Dewaele and Furnham, 2000). At a grammatical level, Extraverts use greater proportions of pronouns, adverbs, verbs (Cope, 1969), which contrasts with the more explicit language of the Introverts and their increased use of nouns, modifiers and prepositions (Dewaele and Furnham, 2000). Additionally, Extraverts demonstrate lower lexical richness in formal situations (Dewaele and Furnham, 2000), whilst analysis of informal e-mail communication has shown highly Neurotic language to be more repetitious (Gill, 2003; Gill and Oberlander, 2003b). At a more content-oriented level, Pennebaker and King (1999), using the Linguistic Inquiry and Word Count text analysis program, showed that broad psychological language categories are related to dimensions of personality variation. For example, they found that when writing about thoughts and feelings, high Neurotics use more negative emotion words and fewer positive emotion words.

However, our interest here is on interaction: dialogue and conversation. Studies using speech act coding have found that Introverts used more hedges and problem talk, namely expressing qualification, and dissatisfaction with one's own activities, while Extraverts expressed more pleasure talk, agreement, and compliments, with content focusing more on extracurricular activities (Thorne, 1987). Extraverts have also been shown to use more self-referent statements, and initiate more laughter (Gifford and Hine, 1994). Gifford and Hine also found that Extraverts talk more, with other studies finding that they use a greater total number of words (Campbell and Rushton, 1978; Carment, Miles, and Cervin, 1965). As would be expected, Extraverts show greater desire to initiate interactions (McCroskey and Richmond, 1990), even in computer-mediated communication (Yellen, Winniford, and Sanford, 1995). Also, Dewaele (2002) finds that in L3 English production, Extraversion (and also Psychoticism) showed a strong negative relationship to communicative anxiety, whilst Neuroticism showed a positive relationship.

Studies investigating hemispheric asymmetry provide a further perspective on this area, for example, Davidson (2001) proposes the relationship between Extraversion and positive affect with approach behaviours, and Neuroticism and negative affect and withrawal behaviours. In the following hypotheses, we explore the implications of personality, affect and approach/withdrawal on priming behaviour.

Hypotheses for interpersonal priming

The likelihood of priming may be affected by the tendency to approach or the tendency to withdraw—or by both.

If Extraversion is associated with approach behaviours, it is natural to expect that higher Extraversion will lead to "more approach", and that this might mean that an individual will coordinate more with their interlocutor. Furthermore, the Extravert's higher drive to gain or retain the conversational floor will mean that less effort can be directed towards detailed language planning. Hence, if their partner has made a lexical or syntactic choice, the High Extravert is likely to re-use that choice, rather than explicitly planning a new one (cf. Gill and Oberlander, 2003a).

If Neuroticism is associated with withdrawal behaviours, it could well be that high levels of this trait result in "more withdrawal" and lower engagement with the interlocutor. Furthermore, the inward (worrying) focus of a High Neurotic might mean that more resources are devoted to inner thought, and fewer to interaction with the environment. Thus, we might expect that such an individual will coordinate less with their interlocutor.

Thus, there is a clear prediction for Extraversion, and a slightly more complex picture for Neuroticism. Of course, it could be that neither Extraversion nor Neuroticism have any effect on coordination or priming.

Method

In syntactic priming, a particular syntactic structure is more likely to be produced given prior exposure to the same structure (Schenkein, 1980). This

phenomenon has been replicated under experimental conditions when speakers say, hear, or read sentences (e.g., Bock, 1986; Pickering and Branigan, 1998; Corley and Scheepers, 2002). Bock and colleagues found that people tended to repeat the active or passive form of a sentence they had just read in describing an unrelated picture (Bock, 1986; Bock, Loebell, and Morey, 1992). In this study we employ the confederate priming method (Pickering and Branigan, 1998): The subject of the experiment takes part in a dialogue game along with a confederate of the experimenter. The game involves matching and describing pictures. Both participants apparently have the same two tasks: to describe a set of pictures so that the other participant can match them, and to verify whether the descriptions that they hear match the picture that they see. However, the confederate's descriptions are scripted.

Participants

Forty University of Edinburgh students who were self-declared native speakers of English were paid to participate in this study. Personality information derived from the NEO-PI questionnaire is as follows: Extraversion M = 51.75 (SD = 12.82), and Neuroticism M = 54.18 (SD = 12.72).

Materials and Design

We prepared two sets of pictures depicting actions. Each set included 12 pictures depicting transitive actions involving an agent and a patient. The entities depicted were chosen to be easily recognisable and nameable. There were two pictures for each of 12 transitive verbs (*bite, chase, dust, hit, kick, lift, poke, pull, push, shoot, touch, weigh*). These 24 pictures comprised the set of targets. The remaining 120 pictures in each set depicted intransitive actions. There were several pictures for each of 20 intransitive verbs. These comprised the filler pictures.

The appropriate verb was printed under each action. Each set of pictures depicted the same range of entities and actions. However, the pairing of entities with actions was different.

We term one set the Subject's Description Set and the other set the Confederate's Description Set. We created ordered pairs of prime and target pictures by pairing each description of a transitive action from the Confederate's Description Set (the prime) with a picture depicting a transitive action from the Subject's Description Set (the target picture).

Half of the prime sentences were assigned active descriptions of the form 'the X verbing the Y', and half were assigned passive descriptions of the form 'the Y being verbed by the X'. An experimental item was defined as the confederate's scripted description of a prime picture plus the subject's target picture paired with it. There were thus two versions of each item: active confederate description and passive confederate description.

We constructed four lists containing 24 experimental items and 120 subject fillers. The confederate fillers were randomly distributed in the remaining gaps. The entities depicted in the target picture were not present in the immediately preceding block (prime plus subject fillers and confederate fillers). The verb also differed between prime and target. Each picture was assigned to either the match or the mismatch condition for the matching task. For the latter, we assigned another picture depicting a different entity doing the same action (thus using the same verb) was assigned. Each list contained 12 experimental items with active prime descriptions and 12 with passive prime descriptions. Exactly one version of each item appeared in each list. Hence, Prime Type (active vs. passive) was manipulated within subjects and items. The dependent measure was the proportion of descriptions of target pictures produced with a passive structure.

Procedure

The Subject's Description Set was presented to the subject via a computer program.The order of the pictures was randomised for each subject, with between four and eight filler items intervening between each experimental item. A divider prevented the subject from seeing the confederate or his computer screen. The experimenter told the subject and the confederate that the experiment was investigating how well people communicate when they cannot see each other. Their tasks were alternately to describe the pictures to the other participant, and to match their picture to the other participant's descriptions. When it was the subject's turn to match, the confederate would see a sentence appear on his screen which he would read aloud and then press space bar, at which point a picture would appear on the subject's screen. The subject was instructed to say "yes" or "no" (or ask for repetition) and to press the Z key for "no" and the M key for "yes" according to whether the picture matched or mismatched the description. When it was the subject's turn to describe, a picture would appear on the subject's screen and the confederate would say "yes" or "no" (or ask for repetition) and press the Z key or the M key according to whether the picture on his screen matched or mismatched the description. Throughout the session, the experimenter and confederate acted as if the confederate was a genuine subject (e.g., the confederate asked questions about the task). Before the experiment, there was a practice session with two filler items each, after which the subject could ask for clarification if necessary. The confederate also gave the first description. Hence the confederate's description of a prime always immediately preceded the subject's description of a target. Both dialogue participants wore a lapel microphone. The experimental session was recorded on audio tape and subsequently transcribed.

Table 1: Proportion of Passive target responses after active and passive primes and degree of priming

Group	Nos.	PP	AP	Priming
Low E	8	.1363	.0300	10.6
Mid E	27	.2015	.0270	17.5
High E	5	.1500	.0480	10.0
Low N	5	.1160	.0480	6.8
Mid N	28	.2271	.0261	20.1
High N	7	.0486	.0343	1.4
Total	40	.1820	.0302	15.2

We coded the first response that the subject produced; 3 target responses that described the agent as the patient and the patient as the agent were excluded. We coded the remaining target 957 responses as passive if the patient was described as being verbed by the agent and as active if the agent of the action was described as verbing the patient.

An analysis of variance (ANOVA) was conducted, with prime type (active vs. passive) as a within subjects factor and Neuroticism (Low [> -1 s.d. of the mean], Mid [< 1 s.d. of the mean], High [$> +1$ s.d. of the mean]) as a between subjects factor.

Results

Proportions of passive target responses following passive and active primes are reported in Table 1; these are described by personality type of participant, and also for the group overall. Here we can see that in both cases the Mid groups appear to show greater priming. However the High and Low Neurotic groups appear to show even lower levels of priming than for Extraversion.

Turning now to our analysis of variance, and here the ANOVA revealed a significant effect of prime type (active vs. passive) on the proportion of passive forms used (F_1 (1,37) = 6.63; $p < 0.05$; F_2 (1,23) = 97.01; $p < 0.05$).

A significant interaction was found between Neuroticism (Low, Mid or High) and prime type (F_1 (1,37) = 3.68; $p < 0.05$). Post-hoc Tukey tests revealed that both the High N and Low N groups primed significantly less than the Mid N group ($p < 0.05$). No interaction was found between Extraversion and prime (F_1 (1,37) = 0.60; $p > 0.1$).

Discussion

We found a reliable effect of syntactic priming of active and passive structures in a dialogue task. This confirms our expectations and replicates previous syntactic priming found in dialogue (e.g., Pickering and Branigan, 1998) and with active vs. passive forms (e.g., Bock, 1986).

Additionally, our results demonstrate that Neuroticism is related to the degree of syntactic priming for passive constructions; Extraversion is not. We now relate these results to our hypotheses. For Extraversion, we proposed that higher levels of Extraversion would lead to an increase in priming. Here we found that the Mid group primed more, however this result was not significantly different to that of the Low and High groups. In this case we therefore accept the null hypothesis that Extraversion is not related to levels of priming. For Neuroticism, we find that the Low and High groups primed significantly less than the Mid group. Comparing this result directly with our Neuroticism hypothesis creates a tension: We proposed that the High group would less likely to prime due to an inward focus and thus withdrawal from their partner. To address these findings, we therefore reframe our Neuroticism hypothesis as follows: as before, we claim that the High group are less likely to prime due to inward focus, but that the Low group are also less likely to prime, since they are less concerned with monitoring themselves in relation to their interlocutor. In this case—as in our results—the extreme High and Low levels of the trait have an inhibitory effect on priming, and the Mid trait levels represent a facilitating effect.

We acknowledge that such explanation is relatively speculative, and further experimentation will be required to test this hypothesis. For example, the NEO-PI questionnaire divides Neuroticism into 6 facets: anxiety, angry hostility, depression, self-consciousness, impulsivity, vulnerability. It may be that these may relate more specifically to withdrawal or threat-monitoring, in which case these could be related to the priming information. However, we expect that a larger experimental population would be required for such work. For Extraversion, no significant pattern emerges, however we propose that the extremes are similarly inhibited by over- or under-other-directedness.

Turning now the significance of our findings, and they have several important implications. At a theoretical level, they provide more data about personality behaviour in dialogue contexts, which extend previous research using monologue data. Additionally this can better inform our understanding of personality in relation to models of language production.

Our results also contribute to the dialogue and priming literature which, for example, acknowledge that individuals often behave differently, but that systematic variation has mainly been examined in sociological terms. Here we have presented data which shows real and important differences between individuals in conversational behaviour, and highlights the potential role of personality in priming experimentation, more generally.

Finally, our findings can be used to directly inform dynamic computer interface technology, which could allow linguistic alignment in a realistic way. For example, Nass, Moon, Fogg, and Reeves (1995) have shown that computer users viewed their ma-

chine more favourably when it mirrored their personality. On the basis of work reported here, we are closer to being able to represent personality at the conversational, interactive level. We therefore anticipate that this will lead to more convincing artificial agents and intelligent dynamic computer interfaces.

These findings also nicely complement those presented by Branigan, Pickering, Pearson, McLean, and Nass (2003), in which computer users syntactic align with a pre-programmed computer interface, whether they believed this to be another person or an 'unintelligent computer'. Therefore, if such an 'unintelligent computer' was to project personality, we may expect it to vary its degree of priming—in addition to its lexicon—depending upon the sort of personality *it* may wish to project.

Conclusion

We have used experimental priming data to investigate the influence of personality on interpersonal language behaviour. Proposing hypotheses which suggested both Extraversion and Neuroticism influence linguistic coordination, here we found that the less interpersonal trait—Neuroticism—surprisingly influenced priming, whilst Extraversion did not. Given our finding that priming is facilitated by moderate Neuroticism, but inhibited by more extreme levels, we explain this in terms of withdrawal by building upon a previously proposed model of personality and language production. Issues regarding the significance and potential implications of this study are also discussed.

Acknowledgements

The first and second authors gratefully acknowledge support from the UK Economic and Social Research Council and the School of Informatics. We also express our gratitude to Holly Branigan, Sarah Haywood, Alan Marshall, Janet McLean, Martin Pickering and Matt Watson for help and advice with the study.

References

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

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, **18**, 355–387.

Bock, J. K., Loebell, H., and Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, **99**, 150–171.

Bradac, J. and Wisegarver, R. (1984). Ascribed status lexical diversity, and accent: Determinants of perceived status, solidarity, and control of speech style. *Journal of Language and Social Psychology*, **3**, 239–255.

Branigan, H., Pickering, M., and Cleland, A. (2000). Syntactic coordination in dialogue. *Cognition*, **75**, B13–B25.

Branigan, H., Pickering, M., Pearson, J., McLean, J., and Nass, C. (2003). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 186–191.

Brennan, S. E. and Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Memory and Cognition*, **22**, 1482–1493.

Campbell, A. and Rushton, J. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology*, **17**, 31–36.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to Intelligence and Extraversion. *British Journal of Social and Clinical Psychology*, **4**, 1–7.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, **16**, 1–19.

Corley, M. and Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, **9**, 126–131.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Coupland, N. (1980). Style-shifting in a Cardiff work-setting. *Language in Society*, **9**, 1–12.

Davidson, R. J. (2001). Toward a biology of personality and emotion. *Annals of the NY Academy of Sciences*, **935**, 191–207.

Dewaele, J.-M. (2002). Psychological and sociodemographic correlates of communication anxiety in L2 and L3 production. *International Journal of Bilingualism*, **6**, 23–28.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**, 21–29.

Garrod, S. and Doherty, G. (1987). Saying what you mean in dialogue. *Cognition*, **27**, 181–218.

Gifford, R. and 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. (2003). *Personality and Language: The projection and perception of personality in computer-mediated communication*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003a). Looking forward to more extraversion with n-grams. In L. Lagerwerf, W. Spooren, and L. Degand, editors, *Determination of Information and Tenor in Texts: Multiple Approaches to Discourse 2003*, pages 125–137. Stichting Neederlandistiek & Nodus Publikationen, Amsterdam & Münster.

Gill, A. and Oberlander, J. (2003b). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Haywood, S., Pickering, M., and Branigan, H. (2003). Co-operation and co-ordination in the production of noun phrases. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 533–538.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennysylvania Press, Philadelphia.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

McCroskey, J. and Richmond, V. (1990). Wilingness to communicate: A cognitive view. *Journal of Social Behaviour and Personality*, **5**, 19–37.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, **43**, 223–239.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Pickering, M. and Branigan, H. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, **39**, 633–651.

Pickering, M. and Garrod, S. (in press). Towards a mechanistic psychology of dialogue. *Behavioural and Brain Sciences*.

Schenkein, J. (1980). A taxonomy for repeating action sequences in natural conversation. In B. Butterworth, editor, *Language production*, volume 1, pages 21–47. Academic Press, London.

Scherer, K. R. (1978). Inference rules in personality attribution from voice quality: The loud voice of extraversion. *European Journal of Social Psychology*, **8**, 467–487.

Siegman, A. W. (1987). The tell-tale voice: Nonverbal messages of verbal communication. In A. Siegman and S. Feldstein, editors, *Nonverbal behaviour and communication*, pages 642–654. Erlbaum, Hillsdale, NJ.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, **53**, 718–726.

Yellen, R., Winniford, M., and Sanford, C. (1995). Extraversion and introversion in electronically-supported meetings. *Information & Management*, **28**, 63–74.

D.6 Oberlander and Gill (to appear)

Individual differences and implicit language: personality, parts-of-speech and pervasiveness

Jon Oberlander (J.Oberlander@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Alastair J. Gill (A.Gill@ed.ac.uk)
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

Dewaele and Furnham predict that in oral language Extraverts prefer to produce what they term implicit language. They use: more pronouns, adverbs and verbs; and fewer nouns, adjectives and prepositions. However, communication in a computer-mediated environment, such as e-mail, might disrupt these preferences. Also, other personality dimensions, such as Neuroticism, may be related to implicitness. The study exploited an existing corpus of e-mail texts written by native English speakers of known personality. Stratified corpus comparison used n-gram-based techniques from statistical natural language processing, to compare relative frequencies of use of (sequences of) parts-of-speech. Implicitness effects were found, and Neuroticism appeared to have a clearer impact than Extraversion.

Personality and language

Individuals differ in the way they speak and write. Some of those differences are systematic, and can be attributed to apparently deeper differences, such as personality traits, like Extraversion and Neuroticism. Extraversion is a trait strongly related to interpersonal interaction and sociability, whereas, Neuroticism, or Emotional Stability, is related to internal emotional states, rather than interaction. In the past, it has been found that both these personality traits do significantly influence an individual's language production behaviour in a variety of contexts (Pennebaker and King, 1999; Dewaele and Furnham, 1999). Recent work has investigated e-mail text, and suggested that there are characteristic sequences of words and punctuation associated with each end of both dimensions (Extravert or Neurotic) (Gill and Oberlander, 2002, 2003).

However, Mehl and Pennebaker (2003) note that linguistic style is more consistently described by its syntactic component, than by content. So, it could be that the relative use of different parts-of-speech (POSs) is a more important indicator of personality than the relative use of words or strings of words.

The work by Dewaele and Furnham suggests that, at least for Extraversion, there are real effects to be found in spoken language, at the level of POSs. In their account, implicit language involves a preference for pronouns, adverbs and verbs, whereas explicit language involves a preference for nouns, adjectives and prepositions. Heylighen and Dewaele (2002) suggest that Extraversion leads to implicitness due to greater visual-spacial capacities, and this is part of an overall preference for informal language. However, this work leaves open whether or not implicitness effects will be found for Neuroticism. Gill and Oberlander's work suggests that formality may also be a factor in Neurotic language behaviour, because the reduced resources of high Neurotics do not enable detailed language planning. But that work did not investigate implicitness in patterns of POS use. It would therefore be interesting to know whether Dewaele and Furnham's 'Implicit-Extravert hypothesis' applies in the genre of e-mail text—a genre close to spoken language—and if so, how.

To address this question, the rest of this paper is structured as follows. First, we give some background to help frame implicitness hypotheses that gives POS predictions for both Extraversion and Neuroticism. We then present the stratified corpus comparison methods used in analysing POS use in the e-mail corpus. Results were somewhat unexpected, in that implicitness predictions appear to be confirmed for Neuroticism, but not for Extraversion. We discuss possible ways of resolving the issue.

Background

Two personality traits

Extraversion and Neuroticism are traits which are common to the two major trait theories of personality: Eysenck's three factor model (Eysenck and Eysenck, 1991); and the five factor model developed by Costa and McCrae (Costa and McCrae, 1992) and others.

They are described as follows: High Extraverts are said to be sociable, easy-going, and optimistic, and to take chances. Low Extraverts (or Introverts) are said to be quiet, and reserved, and to plan ahead, and dislike excitement. High Neurotics are said to be: anxious, worrying, over-emotional, and frequently depressed. Low Neurotics are said to be: calm, even-tempered, controlled, and unworried (Eysenck and Eysenck, 1991).

Dewaele and Furnham

Furnham (1990) has proposed the following features of Extravert and Introvert language. Extravert language: is less formal; has a more restricted (rather than elaborated) code; uses vocabulary more loosely, where this is defined in terms of how correctly words are used, and how unusual they are. And it uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions). This last tendency directly involves POSs. Using factor analysis of syntactic tokens produced by L2 speakers, Dewaele and Furnham (2000) describe implicit language as a preference for pronouns, adverbs and verbs, and they contrast it with explicit language, seen as a preference for nouns, modifiers and prepositions. So Extraverts prefer implicitness, and Introverts prefer explicitness. For the purposes of this paper, we shall term this the Implicit-Extravert Hypothesis. The hypothesis appears to hold in both informal and formal situations, and is consistent with previous analyses of the individual linguistic categories (Dewaele, 2001). Cope (1969) also notes a lower lexical diversity (measured as type-token ratio), for Extravert native French speakers, with this also the case for non-native speakers of English (Dewaele and Furnham, 2000).

However, although they have discussed varieties of anxiety and their effects on communication, Dewaele and Furnham have not attempted to predict which part-of-speech patterns might be characteristic of the related trait Neuroticism. What might we expect to find?

An extension: Implicit-Neuroticism

Previous work by Gill and Oberlander (2002, 2003) gathered a corpus of e-mail messages, and analysed it for characteristic words and sequences of words. The corpus comprised 210 texts produced by 105 University students or recent graduates (37 males, 68 females). Each participant composed two e-mails *to a good friend whom they hadn't seen for quite some time*, spending around 10 minutes on each message. The first e-mail concerned their activities in the past week, the second discussed their plans for the next week. The total corpus size is around 65,000 words.

Following analysis of occurrences of individual words, and sequences of words, it was reported that the corpus results on Extravert words were broadly consistent with previous findings, for instance using informal language, looser punctuation, vaguer quantification and more co-ordination. This therefore appears to fit the Implicit-Extravert hypothesis; however, no POS analysis was reported.

However, there were also results on Neurotic language use. Pennebaker and King (1999) previously argued that High Neuroticism was associated with a language factor for 'Immediacy'. Gill and Oberlander (2003) extended these results, suggesting that 'High Neurotics show a preference for forms occurring frequently in speech, for example, *I, and, that*, rather than less common words such as *abject, suspicion, tether*. This preference for common words contributes towards the very low lexical density found in highly Neurotic texts, demonstrated by the high level of repetition over ten-word sections of text.'

What is interesting about this is that it suggests that Dewaele and Furnham's ideas about formality and implicitness might be as relevant to the Neuroticism dimension as they are to the Extraversion dimension. If they are, then we would expect that—like High Extraverts—High Neurotics will use more verbs, adverbs and pronouns, while Low Neurotics will use more nouns, adjectives, and prepositions. We call this the Implicit-Neurotic Hypothesis (INH). It obviously raises the question of whether or not *both* dimensions are related to implicitness, and the relative strength of any connections.

To address this question, we here apply to the existing e-mail corpus a series of techniques to derive POS frequencies, and POS sequences.

Syntactic Analysis of the Corpus

Method

The personality corpus was acquired as described above. It was tagged using the Penn part-of-speech tagset, using the MXPOST tagger (Ratnaparkhi, 1996). Further processing removed the original words, leaving their associated POS tags. A subsequent stage of processing reduced the POS tags from the detailed Penn tagset to more general syntactic categories. The 45 Penn tags (see Marcus, Santorini, and Marcinkiewicz, 1994, for more details) were converted to 10 broader categories, as implemented in the electronic version of the Shorter Oxford English Dictionary which is incorporated into the MRC Psycholinguistic Database (Wilson, 1987). These are: Noun (NN), Adjective (ADJ), Verb (VBN), Adverb (ADV), Preposition (PRP), Conjunction (CONJ), Pronoun (PRN), Interjection (INT), Past Participle (VPP), and Other [syntactic categories] (O). In addition to these categories, we also make use of ⟨p⟩ indicating punctuation, and 'NA', which indicates that a feature does not belong to any of the above categories and generally represents the ⟨END⟩, end of text marker. Note that here we use a different set of labels to enhance intelligibility, and these do not co-incide exactly with those used in the MRC database: for instance, we use 'PRP' instead of 'R'.

The reduced-tag corpus—with the more general syntactic categories—was then divided into stratified sub-corpora. In stratifying, we isolate a 'reference corpus' of text from authors with a personality profile which is not extreme on any of the measured dimensions. We can then compare authors from each of the extreme personality groups with this 'neutral' (here termed 'mid') group. Thus, High

and Low personality group samples were created by splitting them at greater than 1 standard deviation above and below the EPQ-R score for each dimension. The additional requirement was made that authors had to be *within* 1 standard deviation on the dimensions other than the one for which they were extremely high or low. Additionally, all texts which were within 1 standard deviation across *all* personality dimensions were assigned to the personality 'neutral' Mid sub-corpus. Thus, on any dimension, we have three groups to compare (High, Mid, and Low).

The resulting sizes of the subcorpora are as follows: Around 6,000 words for the high Extraversion, and over 2,000 words for the low Extraversion groups (11 and 4 authors respectively); Over 3,000 words for the high Neurotic and around 6,000 words for the low Neurotic groups (6 and 9 authors). The Neutral group was around 10,000 words (23 authors).

To identify collocations in the tagged sub-corpora, we calculate 1–5 word n-grams, and do not use a rank or frequency cut-off during calculation, but only present features with a frequency ≥ 5. This enables an accurate log-likelihood statistic (G^2) of their occurrence between groups to be calculated (cf. Rayson, 2003). We use N-gram software (Banerjee and Pedersen, 2003) to compute G^2 for 2- and 3-grams. To identify those robust collocations which distinguish one group from another, we need to make a three-way comparison of the linguistic features across the high-mid-low corpora for each group. We calculate the relationships between the three groups, and for each feature in each corpus we identify its frequency and relative frequency, and then where relevant its relative-frequency ratio and log-likelihood between High-Low, High-Mid and Low-Mid groups. This allows us to compare the relative usage and statistical significance of the difference in the use of features between groups.

Results

We first report the results of the unigram analysis for Extraversion and Neuroticism dimensions, we then report the findings of the overall n-gram analyses (1–5 item sequences). Following this, the results for Extraversion and Neuroticism are outlined.

Unigram Syntactic Analysis

Results of the unigram analysis for the reduced set of syntactic tags can be found in Tables 1 and 2. We display the results for all tags present in our data; however G^2 values which achieve significance of $p \leq 0.05$ or $p \leq 0.01$ are noted by $*$ or $**$ respectively.

In this presentation of the results, we draw attention to features which are characteristic of the High or Low groups, compared with the usage of the feature more generally. In the tables, we distinguish whether a feature is under- or over-used by one of the three groups (High, Mid or Low), relative to the two other groups; this information is given in the final three columns of each table, with over-use indicated by + and under-use by −. However, a more concise view of the results can be gained in the following way. At least two kind of features can be associated with (say) High Neuroticism: unigrams which are over-used by High Neurotics; and unigrams which are under-used by Low Neurotics. Thus, Figure 1 lists, for each dimension and each sub-group, the features which are associated with that group *either* via their over-use of the feature, *or* an opposite group's underuse.

High Extraverts	[CONJ]
Mid Extraverts	–
Low Extraverts	[VPP]
High Neurotics	[CONJ] [PRN]
Mid Neurotics	–
Low Neurotics	[ADJ] [NN]

Figure 1: Summary of unigram POS analysis

For Extraversion, conjunction (CONJ) is characteristic of High Extraverts, and past participle verbs (VPP) of Low Extraverts. The Mid Extravert group shows no significant under- or over-use of the general tags. For Neuroticism, conjunction (CONJ) and pronouns (PRN) are characteristic of High Neurotics, and adjectives (ADJ) and nouns (NN) of Low Neurotics. The Mid Neurotic group shows no significant under- or over-use of the general tags.

For these results, we note the generally modest levels of significant differences we found between personality groups. We may take this to indicate that these groups generally use relatively similar proportions of the relevant parts of speech. However, the POSs may also occur in different contexts or sequences, thus indicating differences in they way they are used. We therefore turn to the results of the n-gram analysis of the syntactic tag data.

N-gram Syntactic Analysis

There is insufficient space to display the full results. A concise view is therefore given in Figure 2. Notice that for the Mid groups, we have to distinguish features labelled specifically as under-use, since this is of course relative to both the High and Low groups.

The features here reach much higher levels of significance than the unigrams, so here we only discuss those which reach the critical value of 10.83 (i.e., $p \leq 0.001$). 32 n-gram features reach this value for Neuroticism, and 25 for Extraversion. Of these, the majority in each case reach the 15.13 critical value ($p \leq 0.0001$): 23 and 17, respectively. The features reaching this higher value are predominantly bigrams, exceptions being the longer n-grams for

Feature	Rank	High Freq.	High R.Freq	Mid Freq.	Mid R.Freq	Low Freq.	Low R.Freq	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
VPP	1	118	0.0173	202	0.0185	66	0.0260	0.34	5.43*	6.73**			+
CONJ	2	258	0.0378	338	0.0310	88	0.0347	5.80*	0.88	0.50	+		
ADV	3	562	0.0824	963	0.0882	238	0.0938	1.67	0.71	2.76			
PRP	4	679	0.0995	1100	0.1008	231	0.0910	0.06	2.02	1.40			
O	5	1071	0.1570	1714	0.1570	369	0.1454	0.00	1.82	1.64			
VBN	6	1156	0.1695	1804	0.1652	449	0.1769	0.44	1.65	0.60			
⟨p⟩	7	667	0.0978	1048	0.0960	228	0.0898	0.14	0.84	1.23			
ADJ	8	404	0.0592	617	0.0565	136	0.0536	0.53	0.32	1.03			
NA	9	23	0.0034	47	0.0043	9	0.0035	0.95	0.30	0.02			
PRN	10	696	0.1020	1118	0.1024	277	0.1091	0.01	0.89	0.89			
NN	11	1177	0.1725	1945	0.1782	442	0.1742	0.76	0.19	0.03			
INT	12	11	0.0016	21	0.0019	5	0.0020	0.23	0.00	0.13			

Table 1: Reduced syntactic tag unigram analysis, Extraversion.
Note. $*p < .05$, $**p < .01$, $df = 1$.

Feature	Rank	High Freq.	High R.Freq	Mid Freq.	Mid R.Freq	Low Freq.	Low R.Freq	High-Mid G^2	Low-Mid G^2	High-Low G^2	High Use	Mid Use	Low Use
ADJ	1	193	0.0501	617	0.0565	447	0.0660	2.15	6.15*	10.50**			+
CONJ	2	155	0.0403	338	0.0310	210	0.0310	7.09**	0.00	6.01*	+		
NN	3	625	0.1624	1945	0.1782	1230	0.1815	4.13*	0.27	5.22*	−		
PRN	4	424	0.1102	1118	0.1024	648	0.0956	1.62	1.93	5.06*	+		
INT	5	9	0.0023	21	0.0019	6	0.0009	0.23	3.19	3.48			
VPP	6	63	0.0164	202	0.0185	146	0.0215	0.74	1.95	3.44			
VBN	7	688	0.1787	1804	0.1652	1132	0.1671	3.04	0.09	1.94			
NA	8	13	0.0034	47	0.0043	19	0.0028	0.63	2.63	0.26			
PRP	9	352	0.0915	1100	0.1008	650	0.0959	2.55	0.99	0.53			
O	10	627	0.1629	1714	0.1570	1035	0.1528	0.62	0.48	1.60			
ADV	11	318	0.0826	963	0.0882	595	0.0878	1.04	0.01	0.78			
⟨p⟩	12	382	0.0992	1048	0.0960	657	0.0970	0.31	0.04	0.13			

Table 2: Reduced syntactic tag unigram analysis, Neuroticism.
Note. $*p < .05$, $**p < .01$, $df = 1$.

punctuation found for Neuroticism. In interpreting this data, we seek distinctive POS collocations. Table 3 shows, for each sub-group, how many distinctive collocations involving each POS were found.

Extraversion From the unigram analysis, we are particularly interested in collocations involving conjunctions (for the High E group) and past participle verbs (for the Low E group). As far as conjunctions are concerned, High Extraverts are associated with the use of [CONJ VBN] and [CONJ ADV], while Low Extraverts are associated with the use of [CONJ VBN PRN]. The latter offers a particularly distinctive collocation, since the pronoun switches the preference from High to Low E. Turning to past participles, we find that High E prefer [VPP PRP], but there are no preferred collocations for Low Extraverts.

Given Table 3, the remaining discrepancies between the High and Low E groups are as follows. Allowing that there are substantially more distinctive collocations for the High E group overall, we find that the High E group has notably more collocations involving: punctuation, adjectives, nouns, and POSs in the Other category. The Low E group has notably more collocations involving verbs and pronouns.

Neuroticism Here, we are most interested in collocations involving pronouns and conjunctions (for the High N group) and adjectives and nouns (for the Low N group). Taking pronouns first, we find a High Neurotic preference for [ADJ PRN VBN], [ADJ PRN] and [VBN PRN O]. Turning to conjunctions, they also show a preference for [VBN ADJ CONJ]. Three of these collocations also involve adjectives, which are used overall more by Low Neurotics. However, the rest of High N preferences for collocations involving pronouns instead involve adverbs: [VBN PRN O ADV VBN], [VBN PRN O ADV], [PRN VBN PRN O ADV] and [ADV PRN VBN PRN]. While Low Neurotics have only one pronoun collocation involving an adjective—[PRN ADJ]—the other three of their preferred pronoun or conjunction collocations also involve adverbs: [PRN ADV], [ADV PRN] and [CONJ ADV].

Given Table 3, and allowing that there are rather more distinctive collocations for the High Neurotic group overall, we find that the High Ns have notably more collocations involving verbs, and POSs in the Other category. The Low Ns have notably more collocations involving: past participle verbs and adverbs.

Discussion

Dewaele and Furnham's original Implicit-Extravert Hypothesis predicted that in spontaneous speech High Extraverts will use more verbs, adverbs and pronouns, and that Low Extraverts will use more nouns, adjectives, and prepositions (see Heylighen and Dewaele, 2002, for a discussion as to why certain POSs are preferred by Extraverts). The unigram analysis did not support these predictions. It indicated that High E use more conjunctions, and that Low E use more past participle verbs. No other overall differences were found, although it is perhaps

High Extraverts [CONJ VBN] [NN NN] [ADV ⟨p⟩] [PRN NN] [⟨p⟩ O] [ADV O] [ADJ ⟨p⟩] [NN ADV] [CONJ ADV] [VPP PRP] [ADJ O] [⟨p⟩ ADJ] [PRN O ADV] [VBN O NN ⟨p⟩] [PRN O ADV VBN] [⟨p⟩ O VBN ADJ ⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩]

Mid Extraverts Underuse: [⟨p⟩ ADV] [⟨p⟩ NN]

Low Extraverts [ADV PRP] [PRN ADV] [VBN PRN O] [VBN PRN ADV] [CONJ VBN PRN] [VBN ⟨p⟩ PRN]

High Neurotics [VBN PRP] [⟨p⟩ O] [⟨p⟩⟨p⟩⟨p⟩⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩] [⟨p⟩⟨p⟩⟨p⟩] [VBN PRN O] [ADJ PRN VBN] [PRP ADJ] [VBN O VBN ADV] [PRN VBN PRN O ADV] [VBN ADJ CONJ] [ADJ PRN] [VBN PRN O ADV VBN] [VBN PRN O ADV] [ADV PRN VBN PRN]

Mid Neurotics Underuse: [PRN ⟨p⟩ ADV] [NN VBN O ADJ] [NN VBN O ADJ NN] [PRN O VBN ⟨p⟩]

Low Neurotics [⟨p⟩ ADV] [PRN ADV] [ADV ADV] [ADJ ⟨p⟩] [ADV O] [VPP ADV] [O ADV] [ADV PRN] [CONJ ADV] [ADV VPP] [PRN ADJ] [VPP PRP]

Figure 2: Summary of n-gram POS analysis

worth noting that since we have both past participles and general verbs, our categories are slightly more fine-grained, which may affect the result.

The new Implicit-Neurotic Hypothesis predicted that High Neurotics will use more verbs, adverbs and pronouns, and that Low Neurotics will use more nouns, adjectives, and prepositions. The unigram analysis partially supported these predictions. It found that High N use more pronouns (and conjunctions), and that Low N use more nouns and adjectives. However, no overall differences were found for verbs, adverbs or prepositions.

At first glance, then, it appears that the Neuroticism dimension is more closely related to implicitness than the Extraversion dimension, in this corpus

POS	Extraversion			Neuroticism			Total
	High	Mid	Low	High	Mid	Low	
⟨p⟩	7	2	1	5	2	2	19
ADJ	4	0	0	4	2	2	12
ADV	6	1	3	5	1	9	25
CONJ	2	0	1	1	0	1	5
NN	4	1	0	0	2	0	7
PRN	3	0	5	7	2	3	20
PRP	1	0	1	2	0	1	5
VBN	4	0	4	9	3	0	20
VPP	1	0	0	0	0	3	4
O	7	0	1	6	3	2	19
NA	0	0	0	0	0	0	0
Total	39	4	16	39	15	23	136

Table 3: Distinctive collocations involving a given POS.

of e-mail text. Two potential explanations emerge to explain the difference between this and Dewaele and Furnham's results: Firstly, they were studying spoken, rather than written, language; and secondly, that they were largely dealing with L2 speakers. Perhaps implicitness is more closely related to Neuroticism in written language, and for Extraversion in spoken language; likewise it may have different effects for native and non-native language users. However, before following this line of reasoning, we should also consider the results of the n-gram analysis. At least two gross patterns are interesting.

First, where a High and Low group do not differ overall in the relative frequency of use of a POS, one group may have rather more types of distinctive collocation involving that POS than the other group. If overall use does not differ, it means that one group is using the POS in many different contexts; the other may be using it in a narrower, or perhaps more stereotypical, range of contexts. Let us call the greater-range case 'pervasive' use. Secondly, where a High and Low group do differ in relative frequency of use of a POS, it is interesting to note whether higher frequency is associated with a greater set of collocations involving that POS, or a smaller set. Intuitions here are not firm; but we might expect that greater relative frequency is associated with a greater range of use—and hence, with perhaps fewer stereotypical collocations. If so, frequency may track pervasiveness.

So, consider again the original Implicit-Extravert Hypothesis: High Extraverts will use more verbs, adverbs and pronouns, and Low Extraverts will use more nouns, adjectives, and prepositions. We find that High E prefer conjunctions overall, but that it is the Low E who tend towards POS-collocations involving verbs and pronouns. So High E use of verbs and pronouns may not be not greater overall, but it is pervasive. Equally, Low E prefer past participle verbs overall, but it is the High E who tend towards POS-collocations involving nouns, adjectives, punctuation, and the Other category. Perhaps Low E use of adjectives and nouns is pervasive. And since Low Extraverts actually use proportionately more VPP, their complete lack of distinctive robust collocations suggests that they use VPP pervasively.

Now, let us turn to the new Implicit-Neurotic Hypothesis. High Neurotics will use more verbs, adverbs and pronouns, and Low Neurotics will use more nouns, adjectives, and prepositions. We find that High N prefer pronouns and conjunctions overall, but that it is the Low N who tend towards POS-collocations involving past participle verbs and adverbs. So perhaps High N use of past participle verbs and adverbs is pervasive. Equally, Low N prefer adjectives and nouns overall, but it is the High N who tend towards POS-collocations involving verbs and the Other category. And again, perhaps Low N use of verbs and Other is pervasive.

This pattern is not quite so simple as the Extravert case, and this may in part be because we have split the verb category in two, distinguishing past participle verbs from verbs in general. Putting this to one side, however, we do find High N use of adverbs to be pervasive; and this at least fits the picture of pervasiveness that seemed to be emerging with Extraversion.

Conclusion

This paper set out to establish whether Dewaele and Furnham's Implicit-Extravert Hypothesis for oral language applies in the genre of written e-mail text produced by native English speakers.

At the simple unigram level, it appears that Neuroticism rather than Extraversion fits the implicitness predictions concerning frequency of use of parts-of-speech. However, we can drill down to the collocations level, and we may assume that the pervasive use of a POS tends to reduce the likelihood of finding stereotypical collocations involving it. If we do, then Extraversion does involve implicitness after all. On this interpretation, a POS can be characteristic of some personality group not because they use it more frequently than other groups; rather, it is characteristic because they use it more pervasively.

Applications of this work include affective text categorisation, and therefore could contribute towards the rapidly expanding field of sentiment classification. In taking this work further, we need to give the idea of pervasiveness a more solid basis. But this is only worth pursuing if the idea is really needed to explain the data. And we will only know this once we have tested the hypotheses against larger corpora in other domains. The corpora could be brand new; but it would certainly be possible to apply the analytic techniques presented here to other previously gathered personality corpora.

Acknowledgements

Our thanks to Jean-Marc Dewaele for his comments and suggestions about this paper. The second author gratefully acknowledges studentship support from the UK Economic and Social Research Council and the School of Informatics.

References

Banerjee, S. and 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.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, **16**, 1–19.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Dewaele, J.-M. (2001). Interpreting the maxim of quantity: interindividual and situational variation in discourse styles of non-native speakers. In E. Nèmeth, editor, *Cognition in Language Use: Selected Papers from the 7th International Pragmatics Conference*, volume 1, pages 85–99. International Pragmatics Association, Antwerp.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**, 509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, **28**, 355–365.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Heylighen, F. and Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, **7**, 293–340.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, **19**, 313–330.

Mehl, M. and Pennebaker, J. (2003). The sounds of social life: A psychometric analysis of student's daily social interactions. *Journal of Personality and Social Psychology*, **84**, 857–870.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**, 1296–1312.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

Rayson, P. (2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Ph.D. thesis, Lancaster University.

Wilson, M. (1987). MRC Psycholinguistic Database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.

Bibliography

Aarts, J. and Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In Granger, S., editor, *Learner English on Computer*, Studies in Language and Linguistics, pages 132–141. Addison Wesley Longman, New York.

Aarts, J. and Rayson, P. (1998). Automatic profiling of learner texts. In Granger, S., editor, *Learner English on Computer*, Studies in Language and Linguistics, pages 119–131. Addison Wesley Longman, New York.

Akert, R. and Panter, A. (1988). Extraversion and the ability to decode nonverbal communication. *Personality and Individual Differences*, 9:965–972.

Albright, L., Kenny, D., and Malloy, T. (1988). Consensus in personality judgements at zero acquaintance. *Journal of Personality and Social Psychology*, 55(3):387–395.

Altenberg, B. (1989). Review of Douglas Biber, *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988. *Studia Linguistica*, 43:167–174.

Amichai-Hamburger, Y. (2002). Internet and personality. *Computers in Human Behaviour*, 18:1–10.

Andersen, S. (1984). Self-knowledge and social inference: II. The diagnosticity of cognitive/affective and behavioural data. *Journal of Personality and Social Psychology*, 46:294–307.

Archer, D., McEnery, T., Rayson, P., and Hardie, A. (2003). Developing an automated semantic analysis system for early modern english. In Archer, D., Rayson, P., Wilson, A., and McEnery, T., editors, *Proceedings of the Corpus Linguistics 2003 conference*, volume 16 of *UCREL technical paper*, pages 22–31, Lancaster. UCREL, University of Lancaster.

Archer, D., Wilson, A., and Rayson, P. (2002). Introduction to the USAS category system. Benedict project report, UCREL, University of Lancaster, Lancaster.

Argamon, S., Koppel, M., Fine, J., and Shimoni, R. (2003a). Gender, genre, and writing style in formal written texts. *Text*, 23:321–346.

Argamon, S., Šarić, M., and Stein, S. (2003b). Learning algorithms and features for multiple authorship discrimination. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico*, pages 27–34.

Baayen, R. (1997). Review of Douglas Biber, *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press, 1995. *Literary and Linguistic Computing*, 12:65–67.

Ball, C. (1994). Automated text analysis: Cautionary tales. *Literary and Linguistic Computing*, 9:295–302.

Ball, G. and Breese, J. (1998). Emotion and personality in a conversational character. In *Proceedings of the Workshop on Embodied Conversational Characters*, pages 83–86, Lake Tahoe, CA.

Bälter, O. (1998). *Electronic Mail in a Working Context*. PhD thesis, Royal Institute of Technology, Stockholm.

Banerjee, S. and 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. (1998). Letters by phone or speech by other means: the linguistics of email. *Language and Communication*, 18:133–170.

Baron, N. (2001). Commas and canaries: the role of punctuation in speech and writing. *Language Sciences*, 23:15–67.

Beeman, M. and Chiarello, C., editors (1998). *Right hemisphere language comprehension: Perspectives from cognitive neuroscience*. Lawrence Erlbaum Associates, Mahwah, NJ.

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

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, 23:526–537.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62:384–414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, 19:219–241.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge University Press, Cambridge.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics*. Cambridge University Press, Cambridge.

Blackman, M. (2002a). The employment interview via the telephone: Are we sacrificing accurate personality judgements for cost efficiency. *Journal of Research in Personality*, 36:208–223.

Blackman, M. (2002b). Personality judgement and the utility of the unstructured employment interview. *Basic and Applied Social Psychology*, 24:241–250.

Blass, T. (1984). Social psychology and personality: Toward a convergence. *Journal of Personality and Social Psychology*, 47:1013–1027.

Bloch, J. (2002). Student/teacher interaction via email: the social context of internet discourse. *Journal of Second Language Writing*, 11:117–134.

Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin*, 117:187–215.

Borkenau, P. and Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, 62:645–657.

Bradac, J. (1990). Language attitudes and impression formation. In Giles, H. and Robinson, W., editors, *Handbook of Language and Social Psychology*, pages 387–412. Wiley, Chichester.

Bradac, J., Bowers, J., and Courtright, J. (1980). Lexical variations in intensity, immediacy, and diversity: An axiomatic theory and causal model. In Clair, R. S. and Giles, H., editors, *Social and Psychological Contexts of Language*, pages 193–223. Lawrence Erlbaum Associates, Hillsdale, N.J.

Bradac, J., Desmond, R., and Murdock, J. (1977). Diversity and density: Lexically determined evaluative and informational consequences of linguistic complexity. *Communication Monographs*, 44:273–283.

Bradac, J., Konsky, C., and Davies, R. (1976). Two studies of the effects of linguistic diversity upon the judgements of communicator attributes and message effectiveness. *Communication Monographs*, 43:70–79.

Bradac, J. and Mulac, A. (1984). A molecular view of powerful and powerless speech styles. *Communication Monographs*, 51:307–319.

Bradac, J., Mulac, A., and House, A. (1988). Lexical diversity and magnitude of convergent versus divergent style shifting: Perceptual and evaluatory consequences. *Language and Communication*, 8:213–228.

Bradac, J. and Wisegarver, R. (1984). Ascribed status lexical diversity, and accent: Determinants of perceived status, solidarity, and control of speech style. *Journal of Language and Social Psychology*, 3:239–255.

Brown, G. and Yule, G. (1983). *Discourse analysis*. Cambridge University Press, Cambridge.

Brown, P. and Levinson, S. (1987). *Politeness: Some universals in Language Usage*. Cambridge University Press, Cambridge.

Buchanan, T. (2000). Potential of the internet for personality research. In Birnbaum, M. H., editor, *Psychological Experiments on the Internet*, pages 121–140. Academic Press, San Diego.

Buchanan, T. (2001). Online implementation of an IPIP five factor personality inventory. Technical report, University of Westminster, `http://www.wmin.ac.uk/~buchant/wwwffi/introduction.html`.

Buchanan, T. and Smith, J. (1999). Using the internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, 90:125–144.

Buckingham, R., Charles, M., and Beh, H. (2001). Extraversion and Neuroticism, partially independent dimensions? *Personality and Individual Differences*, 31:769–777.

Burgoon, J., Bonito, J., Bengtsson, B., Cederberg, C., Lundeberg, M., and Allspach, L. (2000). Interactivity in human-computer interaction: a study of credibility, understanding and influence. *Computers in Human Behavior*, 16:553–574.

Busch, D. (1982). Introversion-Extraversion and the EFL proficiency of Japanese students. *Language Learning*, 32:109–132.

Buss, A. and Finn, S. (1987). Classification of personality traits. *Personality and Social Psychology*, 52:432–444.

Butler, C. (1985). *Statistics in Linguistics*. Blackwell, Oxford.

Butler, C. (2001). A matter of GIVE and TAKE: Corpus linguistics and the predicate frame. *Revista Canaria de Estudios Ingleses*, 42:55–78.

Cañamero, D. (1998). Issues in the design of emotional agents. In *Emotional and Intelligent: The Tangled Knot of Cognition. Papers from the 1998 AAAI Fall Symposium. Technical Report FS-98-03*, pages 49–54, Menlo Park, CA.

Campbell, A. and Rushton, J. (1978). Bodily communication and personality. *British Journal of Social and Clinical Psychology*, 17:31–36.

Campbell, R. and Pennebaker, J. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, 14:60–65.

Carment, D. W., Miles, C. G., and Cervin, V. B. (1965). Persuasiveness and persuasibility as related to Intelligence and Extraversion. *British Journal of Social and Clinical Psychology*, 4:1–7.

Casciaro, T. (1998). Seeing things clearly: social structure, personality, and accuracy in social network perception. *Social Networks*, 20:331–351.

Chambers, J. and Trudgill, P. (1980). *Dialectology*. Cambridge University Press, Cambridge.

Cheng, Y., O'Toole, A., and Abdi, H. (2001). Classifying adults' and children's faces by sex: Computational investigations of subcategorial feature encoding. *Cognitive Science*, 25:819–838.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX 94: 3rd Conference on Computational Lexicography and Text Research (July 7-10 1994). Budapest, Hungary*, pages 23–32.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. Croom Helm, London.

Cochran, W. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10:417–451.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, 2:1–25.

Colley, A. and 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.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proc of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191, San Francisco. Morgan Kaufmann Publishers.

Collot, M. and Belmore, N. (1996). Electronic language: A new variety of English. In Herring, S., editor, *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pages 13–28. Benjamins, Amsterdam.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Colvin, C. and Funder, D. (1991). Predicting personality and behaviour: A boundary on the acquaintance effect. *Journal of Personality and Social Psychology*, 60:884–894.

Cooper, C. (1998). *Individual Differences*. Arnold, London.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counselling Psychology*, 16:1–19.

Corley, M. and Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an internet-based study. *Psychonomic Bulletin and Review*, 9:126–131.

Costa, P. and McCrae, R. (1984). Personality as a lifelong determinant of well-being. In Malatesta, C. and Izard, C., editors, *Affective processes in adult development and aging*, pages 141–157. Sage, Beverley Hills, CA.

Costa, P. and McCrae, R. (1986). Major contributions to the psychology of personality. In Modgil, S. and Modgil, C., editors, *Hans Eysenck: Consensus and controversy*, pages 63–72. Falmer Press, Philadelphia. to check ML .15eys.han.

Costa, P. and McCrae, R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, 13:653–665.

Costa, P. and McCrae, R. R. (1992b). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin*, 52:652–670.

Daille, B. (1995). Combined approach for terminology extraction : Lexical statistics and linguistic filtering. Technical Report 5, UCREL, University of Lancaster, Lancaster.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management*, 29:433–448.

Davidson, R. J. (2001). Toward a biology of personality and emotion. *Annals of the NY Academy of Sciences*, 935:191–207.

de Vicente, A. and Pain, H. (2002). Informing the detection of the students' motivational state: an empirical study. In *Proceedings of the Sixth International Conference on Intelligent Tutoring Systems, volume 2363 of Lecture Notes in Computer Science*, pages 933–943.

Deary, I. and Matthews, G. (1993). Personality traits are alive and well. *The Psychologist*, 6:299–311.

DePaulo, B., Kenny, D., Hoover, C., Webb, W., and Oliver, P. (1987). Accuracy of person perception: Do people know what kinds of impressions they convey? *Journal of Personality and Social Psychology*, 52:303–315.

Depue, R. and Collins, P. (1999). Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion. *Behavioral and Brain Sciences*, 22:491–569.

Dewaele, J.-M. (1993). Extraversion et richnesse lexicale dans deux styles d'interlangue française [Extraversion and lexical richness in 2 styles of French interlanguage]. *I.T.L Review of Applied Linguistics*, 22:87–105.

Dewaele, J.-M. (1995). Variation dans la longueur moyenne d'énoncés dans l'interlangue française [variation in the mean length of utterances in french interlanguage]. In Beheydt, L., editor, *linguistique appliquée dans les années 90 [Special Issue]*, volume 16 of *ALBA Papers*, pages 43–58. ALBA.

Dewaele, J.-M. (1996a). How to measure formality of speech? A model of synchronic variation. In Sajavaara, K. and Fairweather, C., editors, *Approaches to second language acquisition [Special issue]*, volume 17 of *Jyväskylä Cross Language Studies*, pages 119–133. Jyväskylä University.

Dewaele, J.-M. (1996b). Variation dans la composition lexicale de styles oraux [variation in the composition of the lexicon of oral styles]. *IRAL. International Journal of Applied Linguistics*, 34:261–282.

Dewaele, J.-M. (1998). Speech rate variation in 2 oral styles of advanced French interlanguage. In Regan, V., editor, *Contemporary approaches to second language acquisition in social context: Cross-linguistic perspectives*, pages 113–123. University College Academic Press, Dublin.

Dewaele, J.-M. (2002a). Individual differences in L2 fluency: the effect of neurobiological correlates. In Cook, V., editor, *Portraits of the L2 user*, pages 219–250. Multilingual Matters, Clevedon.

Dewaele, J.-M. (2002b). Psychological and sociodemographic correlates of communication anxiety in L2 and L3 production. *International Journal of Bilingualism*, 6:23–28.

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, 49:509–544.

Dewaele, J.-M. and Furnham, A. (2000). Personality and speech production: a pilot study of second language learners. *Personality and Individual Differences*, 28:355–365.

Dias, G. (2003). Multiword unit hybrid extraction. In *Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 12, 2003*.

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41:417–440.

DiMarco, C. and Hirst, G. (1994). A computational theory of goal-directed style in syntax. *Computational Linguistics*, 19:451–499.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74.

Eckman, P., Friesen, K., O'Sullivan, M., and Scherer, K. (1980). Relative importance of face, body, and speech in judgements of personality and affect. *Journal of Personality and Social Psychology*, 38:270–277.

Edelsky, C. (1981). Who's got the floor? *Language and Society*, 10(3):383–421.

Embick, D., Marantz, A., Miyashita, Y., O'Neill, W., and Sakai, K. (2000). A syntactic specialization for Broca's area. *Proceedings of the National Academy of Sciences USA*, 97:6150–6154.

Epstein, J. and Klinkenberg, W. (2001). From Eliza to Internet: a brief history of computerized assessment. *Computers in Human Behavior*, 17:295–314.

Epstein, J., Klinkenberg, W., Wiley, D., and McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior*, 17:339–346.

Evert, S. and Krenn, B. (2001). Methods for the quantitative evaluation of lexical association. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. Toulouse, France*, pages 188–195.

Eysenck, H. (1970). *The Biological Basis of Personality*. Thomas, Springfield, IL.

Eysenck, H. (1993). From DNA to social behaviour: conditions for a paradigm of personality research. In Hettema, J. and Deary, I., editors, *Foundations of personality*. Kluwer, Dortrect.

Eysenck, H. and Eysenck, S. B. G. (1975). *The Eysenck Personality Questionnaire*. Hodder and Stoughton, London.

Eysenck, H. and Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised*. Hodder and Stoughton, Sevenoaks.

Eysenck, H. J. (1991). Dimensions of personality: 16, 5, or 3?—Criteria for a taxonomic paradigm. *Personality and Individual Differences*, 12:773–90.

Eysenck, H. J. (1992). Four ways the five factors are not basic. *Personality and Individual Differences*, 13:667–73.

Eysenck, H. J. and Eysenck, S. B. G. (1964). *Manual of the Eysenck Personality Inventory*. University of London Press.

Eysenck, S., Barrett, P., and Barnes, G. (1993). A cross-cultural study of personality: Canada and England. *Personality and Individual Differences*, 14:1–9.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, 6:21–29.

Ferrando, P. (2003). The accuracy of E, N and P trait estimates: an empirical study using the EPQ-R. *Personality and Individual Differences*, 34:665–679.

Finn, A. and Kushmerick, N. (2003). Learning to classify documents according to genre. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico*, pages 35–45.

Fleischman, M. and Hovy, E. (2002). Towards emotional variation in speech-based natural language generation. In *International Natural Language Generation Conference*, New York, NY.

Fox, S. and Schwartz, D. (2002). Social desirability and controllability in computerized and paper-and-pencil personality questionnaires. *Computers in Human Behavior*, 18:389–410.

Funder, D. (1987). Errors and mistakes: Evaluating the accuracy of social judgement. *Psychological Bulletin*, 101:75–90.

Funder, D. (2001). Personality. *Annual Review of Psychology*, 52:197–221.

Funder, D. and Colvin, R. (1988). Friends and strangers: Aquaintanceship, agreement, and the accuracy of personality judgement. *Journal of Personality and Social Psychology*, 55:149–158.

Funder, D. and Colvin, R. (1991). Explorations in behavioural consistency: Properties of persons, situations, and behaviours. *Journal of Personality and Social Psychology*, 60:773–794.

Funder, D. and Dobroth, K. (1987). Differences between traits: Properties associated with inter-judge agreement. *Journal of Personality and Social Psychology*, 52:409–418.

Funder, D., Kolar, D., and Blackman, M. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology*, 69:656–672. check - cited in funder 95.

Funder, D. C. (1995). On the accuracy of personality judgement: A realistic approach. *Psychological Review*, 102:652–670.

Furnham, A. (1990). Language and personality. In Giles, H. and Robinson, W., editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Furnham, A. and Capon, M. (1983). Social skills and self-monitoring processes. *Personality and Individual Differences*, 4:171–178.

Gabrenya, W. and Arkin, R. (1980). Factor structure and correlates of the Self-Monitoring Scale. *Personality and Social Psychology Bulletin*, 6:13–22.

Gallois, C. and Callan, V. (1986). Decoding emotional messages: Influences of ethnicity, sex, message type, and channel. *Handbook of Language and Social Psychology*, 51:755–762.

Garside, R. and Rayson, P. (1997). Higher-level annotation tools. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 179–193. Longman, London.

Gifford, R. and 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.

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, 15:87–105.

Giles, H., Wilson, P., and Conway, T. (1981). Accent and lexical diversity as determinants of impression formation and employment selection. *Language Sciences*, 3:92–103.

Gill, A., Harrison, A., and Oberlander, J. (to appear). Interpersonality: Individual differences and interpersonal priming. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

Gill, A. and Oberlander, J. (2002). Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pages 363–368.

Gill, A. and Oberlander, J. (2003a). Looking forward to more extraversion with n-grams. In Lagerwerf, L., Spooren, W., and Degand, L., editors, *Determination of Information and Tenor in Texts: Multiple Approaches to Discourse 2003*, pages 125–137. Stichting Neederlandistiek & Nodus Publikationen, Amsterdam & Münster.

Gill, A. and Oberlander, J. (2003b). Perception of e-mail personality at zero-acquaintance: Extraversion takes care of itself; Neuroticism is a worry. In *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pages 456–461.

Gill, A., Oberlander, J., and Conway, S. (pat. pend.). Personality style checker. ("Text Processing Method and System"). Patent applied for/pending: UK/European filing.

Gill, A. J. (1998). Type-token ratio and the measurement of competence in task-oriented dialogue. Master's thesis, School of Cognitive Science, University of Edinburgh.

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist*, 48:26–34.

Gomà-i-Freixanet, M. (1997). Consensus validity of the EPQ: Self-reports and spouse-reports. *European Journal of Psychological Assessment*, 13(3):179–185.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528.

Granger, S. and Rayson, P. (1998). Automatic profiling of learner texts. In Granger, S., editor, *Learner English on Computer*, Studies in Language and Linguistics, pages 119–131. Addison Wesley Longman, New York.

Graybeal, A., Sexton, J., and Pennebaker, J. (2002). The role of story-making in disclosure writing: The psychometrics of narrative. *Psychology and Health*, 17:571–581.

Greenbaum, S., Nelson, G., and Weitzman, M. (1996). Complement clauses in English. In Thomas, J. and Short, M., editors, *Using Corpora for Language Research*, pages 76–91. Longman, London.

Hamburger, Y. and Ben-Artzi, E. (2000). The relationship between extraversion and neuroticism and the different uses of the internet. *Computers in Human behavior*, 16:441–449.

Hancock, J. and Dunham, P. (2001a). Impression formation in computer-mediated communication. *Communication Research*, 28:325–347.

Hancock, J. and Dunham, P. (2001b). Language use in computer-mediated communication: The role of coordination devices. *Discourse Processes*, 31:91–110.

Harrington, S. (2003). The application of intent to style. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico*, pages 55–59.

Herring, S. (1996). Two variants of an electronic message schema. In Herring, S., editor, *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pages 81–106. Benjamins, Amsterdam.

Hewson, C., Laurent, D., and Vogel, C. (1996). Proper methodologies for psychological and sociological studies conducted via the internet. *Behavior Research Methods, Instruments, & Computers*, 28:186–191.

Heylighen, F. and Dewaele, J.-M. (1999). Formality of language: definition, measurement and behavioral determinants. Internal Report: Center Leo Apostel, Free University of Brussels.

Hills, P. and Argyle, M. (2003). Uses of the Internet and their relationships with individual differences in personality. *Computers in Human Behavior*, 19:59–70.

Hinton, P. (1995). *Statistics Explained.* Routledge, London.

Hovy, E. (1996). *Generating Natural Language under pragmatic constraints.* Lawrence Erlbaum, Hillsdale, NJ.

Howeler, M. (1972). Diversity of word usage as a stress indicator in an interview situation. *Journal of Psychological Research*, 1:243–248.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge University Press, Cambridge.

Indefrey, P., Brown, C., Hellwig, F., Amunts, K., Herzog, H., and Seitz, R. (2001). A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences USA*, 98:5933–5936.

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53:251–267.

Johansson, S. and Oksefjell, S. (1996). Towards a unified account of the syntax and semantics of GET. In Thomas, J. and Short, M., editors, *Using Corpora for Language Research*, pages 57–75. Longman, London.

John, O. and Robbins, R. (1993). Determinants of interjudge agreement: The Big Five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, 61:521–551.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Keller, F. and Alexopoulou, T. (2001). Phonology competes with syntax: Experimental evidence for the interaction of word order and accent placement in the realization of information structure. *Cognition*, 79(3):301–372.

Keller, F., Corley, M., and Scheepers, C. (2002). Conducting psycholinguistic experiments over the World Wide Web. Technical report, University of Edinburgh: Human Communication Research Centre, Edinburgh, UK.

Kenny, D. (1994). *Interpersonal perception: A social relations analysis*. Guilford, New York.

Kenny, D. A. and Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, 102:390–402.

Kiesler, D. (1983). The 1982 Interpersonal Circle: A taxonomy for complementarity in human transactions. *Psychological Review*, 90:185–214.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6:231–245.

Kilgarriff, A. and Tugwell, D. (2001). Word sketch: Extraction and display of significant collocations for lexicography. In *COLLOCATION: Computational Extraction, Analysis and Exploitation*, pages 32–38. 39th ACL & 10th EACL, Toulouse, July 2001.

Kline, P. (1983). *Personality: Measurement and Theory*. Hutchinson, London.

Kline, P. (1993a). Comments on "personality traits are alive and well". *The Psychologist*, 6:304.

Kline, P. (1993b). *The Handbook of Psychological Testing*. Routledge, London.

Knapp, H. and Kirk, S. (2003). Using pencil and paper, internet and touch-tone phones for self-administered surveys: does methodology matter? *Computers in Human Behavior*, 19:117–134.

Kolar, D., Funder, D., and Colvin, C. (1996). Comparing the accuracy of personality judgements by the self and knowledgeable others. *Journal of Personality*, 64:311–337.

Koole, S., Jager, W., van den Berg, A., Vlek, C., and Hofstee, W. (2001). On the social nature of personality: Effects of extraversion, agreeableness, and feedback about collective resource use on cooperation in a resource dilemma. *Personality and Social Psychology Bulletin*, 27:289–301.

Koppel, M., Akiva, N., and Dagan, I. (2003a). A corpus-independent feature set for style based text categorization. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis, Acapulco, Mexico*, pages 61–67.

Koppel, M., Argamon, S., and Shimoni, A. (2003b). Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Krantz, J. and Dalal, R. (2000). Validity of web-based psychological research. In Birnbaum, M. H., editor, *Psychological Experiments on the Internet*, pages 35–60. Academic Press, San Diego.

Kshirsagar, S. and Magnenat-Thalmann, N. (2002). A multilayer personality model. In *Proceedings of the Symposium on Smart Graphics, June 11-13, 2002, Hawthorne, NY, USA*.

Kucera, H. and Francis, W. (1967). *Computational analysis of present day American English*. Brown University Press, Providence, RI.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennysylvania Press, Philadelphia.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Langkilde, I. and Knight, K. (1998a). Generation that exploits corpus-based statistical knowledge. In *COLING-ACL*, pages 704–710.

Langkilde, I. and Knight, K. (1998b). The practical value of N-grams in derivation. In Hovy, E., editor, *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 248–255. Association for Computational Linguistics, New Brunswick, New Jersey.

Larstone, R., Jang, K., Livesley, W., Vernon, P., and Wolf, H. (2002). The relationship between Eysenck's P-E-N model of personality, the five-factor model of personality, and traits delineating personality dysfunction. *Personality and Individual Differences*, 33:25–37.

Leech, G. (1997a). Grammatical tagging. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation: Linguistic information from computer text corpora*, pages 19–33. Longman, London.

Leech, G. (1997b). Introducing corpus annotation. In Garside, R., Leech, G., and McEnery, A., editors, *Corpus Annotation: Linguistic information from computer text corpora*, pages 1–18. Longman, London.

Lester, D. (1991). Accuracy of recognition of genuine versus simulated suicide notes. *Personality and Individual Differences*, 12:765–766.

Levelt, W. (1989). *Speaking: From Intention to Articulation*. MIT Press, Cambridge, MA.

Levy, J. and Bullinaria, J. (2001). Learning lexical properties from word usage patterns: which context words should be used. In French, R., editor, *Models of Evolution, Learning and Development*, pages 273–282. Springer-Verlag, Heidelberg.

Lippa, R. and Dietz, K. (2000). The relations of gender, personality, and intelligence to judges' accuracy in judging strangers' personality from brief video segments. *Journal of Nonverbal Behavior*, 24:25–43.

Luu, P., Collins, P., and Tucker, D. (2000). Mood, personality, and self-monitoring: negative affect and emotionality in relation to frontal lobe mechanisms of error monitoring. *Journal of Experimental Psychology: General*, 129:43–60.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M., Santorini, B., and Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19:313–330.

Markey, P. and Wells, S. (2002). Interpersonal perception in internet chat rooms. *Journal of Research in Personality*, 36:134–146.

Marr, D. (1982). *Vision*. MIT Press, Cambridge, MA.

Matthews, G. (1997). An introduction to the cognitive science of personality and emotion. In Matthews, G., editor, *Cognitive science perspectives on personality and emotion*, pages 3–30. Elsevier Science, Amsterdam.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

McCrae, R. and Costa, P. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52:81–90.

McCrae, R. and Costa, P. (1989). The structure of interpersonal traits: Wiggins's Circumplex and the five-factor model. *Journal of Personality and Social Psychology*, 56:586–595.

McCrae, R. and Costa, P. (1997). Personality trait structure as a human universal. *American Psychologist*, 52:509–516.

McCroskey, J. and Richmond, V. (1990). Wilingness to communicate: A cognitive view. *Journal of Social Behaviour and Personality*, 5:19–37.

McDonald, S. (2000). *Environmental determinants of lexical processing effort*. PhD thesis, University of Edinburgh, Edinburgh.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press, Edinburgh.

McGraw, K., Tew, M., and Williams, J. (2000). The integrity of web-delivered experiments: Can you trust the data? *Psychological Science*, 11:502–506.

Mehl, M. and Pennebaker, J. (2003). The sounds of social life: A psychometric analysis of student's daily social interactions. *Journal of Personality and Social Psychology*, 84:857–870.

Milic, L. (1966). Unconscious ordering in the prose of Swift. In Leed, J., editor, *The Computer and Literary Style*, pages 79–106. Kent State University Press, Kent, Ohio.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In Granger, S., editor, *Learner English on Computer*, Studies in Language and Linguistics, pages 186–198. Addison Wesley Longman, New York.

Moffat, D. (1991). Personality parameters and programs. In Trappl, R. and Petta, P., editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, volume 1195 of *Lecture Notes in Artificial Intelligence*, pages 120–165. Springer-Verlag, Berlin.

Monaghan, P., Shillcock, R., and McDonald, S. (2004). Hemispheric asymmetries in the split-fovea model of semantic processing. *Brain and Language*, 88:339–354.

Moon, Y. and Nass, C. (1996). How "real" are computer personalities? *Communication Research*, 23:651–674.

Morris, P., Gale, A., and Duffy, K. (2002). Can judges agree on the personality of horses? *Personality and Individual Differences*, 33:67–81.

Mount, M., Barrick, M., and Stewart, G. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance*, 11:145–165.

Nass, C. and Lee, K. M. (2000). Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In *Proceedings of CHI 2000, The Hague, Amsterdam, 2000*, pages 329–336.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies*, 43:223–239.

Newman, M., Pennebaker, J., Berry, D., and Richards, J. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29:665–675.

Norman, D., Ortony, A., and Russell, D. (2003). Affect and machine design: Lessons for the development of autonomous machines. *IBM Systems Journal*, 42:38–44.

Norušis, M. J. and SPSS Inc. (1994). *SPSS 6.1 Base System User's Guide*. SPSS Inc, Chicago.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh University Press, Edinburgh.

Oberlander, J. and Brew, C. (2000). Stochastic text generation. *Philosophical Transactions of the Royal Society of London, series A*, 358:1373–1385.

Oberlander, J. and Gill, A. (2004). Language generation and personality: two dimensions, two stages, two hemispheres? In *Papers from the AAAI Spring Symposium on Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, pages 104–111.

Oberlander, J. and Gill, A. (to appear). Individual differences in implicit language: personality, parts-of-speech, and pervasiveness. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*.

O'Donnell, M., Mellish, C., Oberlander, J., and Knott, A. (2001). ILEX: An architecture for a dynamic hypertext generation system. *Natural Language Engineering*, 7:225–250.

Orăsan, C. and Krishnamurthy, R. (2002). A corpus-based investigation of junk emails. In *Proceedings of The Third International Conference on Language Resources and Evaluation (LREC-2002)*, Las Palmas de Gran Canaria, Spain.

O'Sullivan, M., Ekman, P., Friesen, W., and Scherer, K. (1985). What you say and how you say it: The contribution of speech content and voice quality to judgements of others. *Journal of Personality and Social Psychology*, 48:54–62.

Oxman, T., Rosenberg, S., Schnurr, P., and Tucker, G. (1988). Diagnostic classification through content analysis of patients' speech. *American Journal of Psychiatry*, 145:464–468.

Panteli, N. (2002). Richness, power cues and email text. *Information & Management*, 40:75–86.

Paulhus, D. and Bruce, M. (1992). The effect of acquaintanceship on the validity of personality impressions: A longitudinal study. *Journal of Personality and Social Psychology*, 63:816–824.

Paunonen, S. (1989). Consensus in personality judgements: Moderating effects of target-rater acquaintanceship and behaviour observability. *Journal of Personality and Social Psychology*, 56:823–833.

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. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, Pittsburgh, PA.

Pedersen, T., Kayaalp, M., and Bruce, R. (1996). Significant lexical relationships. In *Proceedings of the 13th National Conference on Artificial Intelligence*, pages 455–460, Portland, OR.

Pennebaker, J., Mehl, M., and Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54:547–577.

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, 8:162–166.

Pennebaker, J. W. and Francis, M. (1999). *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC2001)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, J. W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77:1296–1312.

Pennebaker, J. W., Mayne, T., and Francis, M. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology*, 72:863–871.

Person, N. K., Graesser, A. C., Bautista, L., Mathews, E. C., and the Tutoring Research Group (2001). Evaluating student learning gains in two versions of autotutor. In Moore, J. D., Redfield, C. L., and Johnson, W. L., editors, *Artificial intelligence in education: AI-ED in the wired and wireless future*, pages 286–293. IOS Press, Amsterdam.

Piao, S. S. L., Rayson, P., Archer, D., Wilson, A., and McEnery, T. (2003). Extracting multiword expressions with a semantic tagger. In *Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, at ACL 2003, 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, July 12, 2003*, pages 49–56.

Picard, R. (2000). Towards computers that recognize and respond to user emotion. *IBM Systems Journal*, 34:705–719.

Pickering, A., Farmer, A., Harris, T., Redman, K., Mahmood, A., Sadler, S., and McGuffin, P. (2003). A sib-pair study of psychoticism, life events and depression. *Personality and Individual Differences*, 34:613–623.

Pickering, M. and Branigan, H. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39:633–651.

Pinsoneault, T. (1996). Equivalency of computer-assisted and paper-and-pencil administered versions of the Minnesota Multiphasic Personality Inventory-2. *Computers in Human Behavior*, 12:291–300.

Pytlik Zillig, L., Hemmenover, S., and Dienstbier, R. (2002). What do we assess when we assess a Big 5 trait? a content analysis of the affective, behavioural, and cognitive processes represented in Big 5 personality inventories. *Personality and Social Psychology Bulletin*, 28:847–858.

Ramsay, R. (1968). Speech patterns and personality. *Language and Speech*, 11:54–63.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

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

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*. PhD thesis, Lancaster University.

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In Birnbaum, M. H., editor, *Psychological Experiments on the Internet*, pages 89–117. Academic Press, San Diego.

Reiter, E. and Sripada, S. (2002a). Human variation in lexical choice. *Computational Linguistics*, 28:545–553.

Reiter, E. and Sripada, S. (2002b). Should corpora texts be gold standards for NLG? In *Proceedings of INLG-02*, pages 97–104.

Rusting, C. (1998). Personality, mood, and cognitive processing of emotional information: Three conceptual frameworks. *Psychological Bulletin*, 124:165–196.

Sanford, F. (1942). Speech and personality. *Psychological Bulletin*, 39:811–845.

Savicki, V., Kelley, M., and Oesterreich, E. (1999). Judgements of gender in computer-mediated communication. *Computers in Human Behavior*, 15:185–194.

Scherer, K. (1972). Judging personality from voice: A cross-cultural approach to an old issue in interpersonal perception. *Journal of Personality*, 40:191–210.

Scherer, K. (1979). Personality markers in speech. In Scherer, K. R. and Giles, H., editors, *Social Markers in Speech*, pages 147–209. Cambridge University Press, Cambridge.

Scherer, K. R. (1978). Inference rules in personality attribution from voice quality: The loud voice of extraversion. *European Journal of Social Psychology*, 8:467–487.

Schumacher, P. and Morahan-Martin, J. (2001). Gender, internet and computer attitudes and experiences. *Computers in Human Behavior*, 17:95–110.

Scott, M. (1997). PC analysis of key words - and key key words. *System*, 25:233–245.

Scott, M. (2001). Mapping key words to *problem* and *solution*. In Scott, M. and Thompson, G., editors, *Patterns of Text*, pages 109–127. John Benjamins, Amsterdam.

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 Siegman, A. and Feldstein, S., editors, *Nonverbal behaviour and communication*, pages 642–654. Erlbaum, Hillsdale, NJ.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.

Smith, C. (1992). Introduction: inferences from verbal material. In Smith, C., editor, *Motivation and personality: Handbook of thematic content analysis*, pages 1–17. Cambridge University Press, Cambridge.

Sneed, C., McCrae, R., and Funder, D. (1998). Lay conceptions of the Five-Factor Model and its indicators. *Personality and Social Psychology Bulletin*, 24:115–126.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, 30:526–537.

Snyder, M. (1987). *Public appearances, private realities: The psychology of self monitoring*. Freeman, New York.

Spain, J., Eaton, L., and Funder, D. (2000). Perspectives on personality: The relative accuracy of self versus others for the prediction of emotion and behaviour. *Journal of Personality*, 68:837–867.

Spain, J., Eaton, L., and Funder, D. (2003). perspectives on personality: The relative value of self versus others for the prediction of emotion and behavior. *Journal of Personality*, 68:837–867.

Stallman, R. (1994). *GNU Emacs Manual*. Free Software Foundation Press, Boston, MA, 10th edition.

Stephenson, G., Laszlo, J., Ehmann, B., Lefever, R., and Lefever, R. (1997). Diaries of significant events: Socio-linguistic correlates of therapeutic outcomes in patients with addiction problems. *Journal of Community and Applied Psychology*, 7:389–411.

Stirman, S. W. and Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63:517–522.

Stubbs, M. (1995). Collocations and semantic profiles: on the cause of the trouble with quantitative studies. *Functions of Language*, 2:23–55.

Stubbs, M. and Barth, I. (2003). Using recurrent phrases as text-type discriminators: a quantitative method and some findings. *Functions of Language*, 10:61–104.

Sussman, N. and Tyson, D. (2000). Sex and power: gender differences in computer-mediated interactions. *Computers in Human Behavior*, 16:381–394.

Swickert, R., Hittner, J., Harris, J., and Herring, J. (2002). Relationships among internet use, personality and social support. *Computers in Human Behavior*, 18:437–451.

Tapasak, R., Roodin, P., and Vaught, G. (1979). Effects of extraversion, anxiety, and sex on children's verbal fluency and coding task performance. *The Journal of Psychology*, 100:49–55.

Teiger, P. and Barron-Teiger, B. (1998). *The Art of SpeedReading People*. Little, Brown, Boston.

Thomas, J. and Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In Thomas, J. and Short, M., editors, *Using Corpora for Language Research*, pages 92–109. Longman, London.

Thomson, R. and Murachver, T. (2001). Predicting gender from electronic discourse. *British Journal of Social Psychology*, 40:193–208.

Thomson, R., Murachver, T., and Green, J. (2001). Where is the gender in gendered discourse? *Psychological Science*, 12:171–175.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality and Social Psychology*, 53:718–726.

Trapnell, P. D. and Wiggins, J. S. (1990). Extention of the interpersonal adjective scales to include the big five dimensions of personality. *Journal of Personality and Social Psychology*, 59:781–790.

Tribble, C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In Burnard, L. and McEnery, T., editors, *Rethinking language pedagogy from a corpus perspective*, pages 75–90. Peter Lang, Frankfurt.

Trott, K. (1994). *A study of lexical evidence for sex-appropriate language use in children between 4 and 9 years of age*. PhD thesis, University of Sheffield, Sheffield.

Trudgill, P. (1974). *The social differentiation of English in Norwich*. Cambridge University Press, Cambridge.

Varges, S. (2002). *Instance-based Natural Language Generation*. PhD thesis, School of Informatics, University of Edinburgh.

Walker, M., Cahn, J., and Whittacker, S. (1997). Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the 1st International Conference on Autonomous Agents (Agents'97)*, pages 96–105. ACM Press.

Weber, B., Schneider, B., Fritze, J., Gille, B., Hornung, S., Kühner, T., and Maurer, K. (2003). Acceptance of computerized compared to paper-and-paper assessment in psychiatric inpatients. *Computers in Human Behavior*, 19:81–93.

Werry, C. (1996). Linguistic and interactional features of internet relay chat. In Herring, S., editor, *Computer-mediated communication: Linguistic, social and cross-cultural perspectives*, pages 47–63. Benjamins, Amsterdam.

White, M. and Baldridge, J. (2003). Adapting chart realization to CCG. In *Proceedings of the 9th European Workshop on Natural Language Generation (EWNLG-03) at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 119–126, Budapest, Hungary.

Wiggins, J. and Pincus, A. (1992). Personality: Structure and assessment. *Annual Review of Psychology*, 43:473–504.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology*, 37:395–412.

Williams, R. (1983). *Keywords: A voacabulary of culture and society*. Fontana Press, London, second edition.

Wilson, A. and Rayson, P. (1993). Automatic content analysis of spoken discourse: a report on work in progress. In Souter, C. and Atwell, E., editors, *Corpus Based Computational Linguistics*, pages 215–226. Rodopi, Amsterdam.

Wilson, M. (1987). MRC Psycholinguistic Database: Machine usable dictionary. Technical report, Oxford Text Archive, Oxford.

Woods, A., Fletcher, P., and Hughes, A. (1986). *Statistics in Language Studies*. Cambridge University Press, Cambridge.

Wray, A. and Perkins, M. (2000). The functions of formulaic language: an integrated model. *Language and Communication*, 20:1–28.

Yellen, R., Winniford, M., and Sanford, C. (1995). Extraversion and introversion in electronically-supported meetings. *Information & Management*, 28:63–74.

Zelenski, J., Rusting, C., and Larsen, R. (2003). Consistency in the time of experiment participation and personality correlates: a methodological note. *Personality and Individual Differences*, 34:547–558.