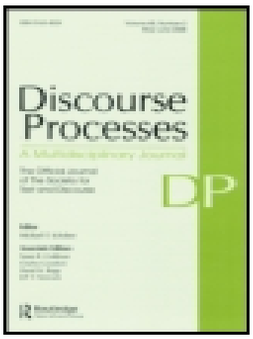

Discourse Processes

ISSN: 0163-853x (Print) 1532-6950 (Online) Journal homepage: www.tandfonline.com/journals/hdsp20

Language With Character: A Stratified Corpus Comparison of Individual Differences in E-Mail Communication

Jon Oberlander & Alastair J. Gill

To cite this article: Jon Oberlander & Alastair J. Gill (2006) Language With Character: A Stratified Corpus Comparison of Individual Differences in E-Mail Communication, Discourse Processes, 42:3, 239-270, DOI: 10.1207/s15326950dp4203_1

To link to this article: https://doi.org/10.1207/s15326950dp4203_1

Published online: 08 Jun 2010.

Submit your article to this journal

Article views: 671

View related articles

Citing articles: 12 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hdsp20

DISCOURSE PROCESSES, *42*(3), 239–270
Copyright © 2006, Lawrence Erlbaum Associates, Inc.

Language With Character: A Stratified Corpus Comparison of Individual Differences in E-Mail Communication

Jon Oberlander and Alastair J. Gill
School of Informatics
University of Edinburgh

To what extent does the wording and syntactic form of people's writing reflect their personalities? Using a bottom-up stratified corpus comparison, rather than the top-down content analysis techniques that have been used before, we examine a corpus of e-mail messages elicited from individuals of known personality, as measured by the Eysenck Personality Questionnaire–Revised (S. Eysenck, Eysenck, & Barrett, 1985). This method allowed us to isolate linguistic features associated with different personality types, via both word and part-of-speech n-gram analysis. We investigated the extent to which extraversion is associated with linguistic features involving positivity, sociability, complexity, and implicitness and neuroticism is associated with negativity, self-concern, emphasis, and implicitness. Numerous interesting features were uncovered. For instance, higher levels of extraversion involved a preference for adjectives, whereas lower levels of neuroticism involved a preference for adverbs. However, neither positivity nor negativity was as prominent as expected, and there was little evidence for implicitness.

Give two people a communication task—such as e-mailing a friend about recent activities—and they are likely to accomplish it in different ways. Some differences depend on their recent experiences or on what they think interests the recipient. Others might depend on character or personality. For example, the following items are initial excerpts from e-mail messages by different authors:

Correspondence should be addressed to Jon Oberlander, School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland. E-mail: J.Oberlander@ed.ac.uk

1. Hi, I'm just back at uni since yesterday, I'm finishing my project proposals. It's going okay, but I really don't want to be slogging over it at the weekend.
2. Hi. This has been my first full week of work in my new job. Actually, not much has changed because I have the same office and the same computer and am doing much the same work.
3. Hi, how's your week been? Mine has been okay but not very interesting. This weekend everyone is going skiing except me, which will be great fun … NOT!
4. Hi again. Okay, so next week I have a few things planned. As I said before I need to start revising—so work is going to be a focus point. I think if I try to do some studying every day, then I can still have fun at night.

The topics are similar, but the excerpts are rather different in style, even at first glance. In the corpus we discuss later, the author of Item 1 is an extravert, as measured by a standard self-report personality measure, whereas the author of Item 2 is an introvert. The author of Item 3 is relatively high in neuroticism, whereas that of Item 4 is low on that scale. However, do such personality differences help account for the linguistic differences?

Our primary aim was to learn more about language production capacities by using a comparative technique involving the study of individual differences among adults. There are, of course, many dimensions along which adults vary—such as working memory capacity, cognitive style, age, gender, and dialect—but we chose to investigate systematic differences in character or personality. We did not presuppose that personality differences are the most important differences in determining discourse style—merely that they are worth investigating. We focused on the medium of e-mail for both substantive and methodological reasons. On the one hand, how people express themselves in e-mail has not been as widely studied as text and speech, but it is a ubiquitous means of written communication, and, unlike most writing, it is regarded as having much of the spontaneity of speech (Bälter, 1998; Baron, 1998; Colley & Todd, 2002). On the other hand, e-mail certainly differs from speech; it is more verbose yet less emotional (Whittaker, 2003). Either way, as a relatively unplanned form of writing, and one in which conventions are quite fluid, e-mail is a genre in which we may expect to find real differences in how diverse individuals express themselves.

This article is structured as follows. First we introduce trait theories of personality, summarize some previous findings on language and personality, and select hypotheses for investigation. Some of these previous results have been obtained using content-analysis techniques. We argue that problems arise when applying such techniques to a corpus of elicited e-mail data. A solution is to exploit bottom-up techniques from computational corpus linguistics as developed by Gill (2004). Regularities are uncovered and related to our hypotheses.

BACKGROUND

Personality Traits

One view of personality sees it in terms of essential traits or factors. Cattell's (1946) pioneering work eventually led to the isolation of 16 primary personality factors. Much subsequent work on traits sought higher order secondary factors, and there are now two main models for these: Eysenck's three-factor model (H. Eysenck & Eysenck, 1991; S. Eysenck, Eysenck, & Barrett, 1985) and Costa and McCrae's (1992) five-factor model, closely related to the Big Five models that emerged from lexical research (Digman, 1990; Goldberg, 1993; Wiggins & Pincus, 1992). Each factor gives a continuous, orthogonal scale ranging from low to high. In practice, there may be some relation among traits, especially for extreme scorers (cf. Buckingham, Charles, & Beh, 2001; H. Eysenck, 1970; Matthews, Deary, & Whiteman, 2003). Two core traits are shared by the main models: extraversion (or extraversion–introversion) and neuroticism (emotionality–stability; Matthews et al., 2003). These are the focus of this article.

In their NEO–Personality Inventory–Revised model, Costa and McCrae (1992) divided extraversion into six facets: Warmth, Gregariousness, Assertiveness, Activity, Excitement-Seeking, and Positive Emotions. H. Eysenck and Eysenck (1975) described extraversion as follows:

> The typical extravert is sociable, likes parties, has many friends, needs to have people to talk to. … 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. (p. 9)

Costa and McCrae (1992) gave neuroticism six facets: Anxiety, Angry Hostility, Depression, Self-Consciousness, Impulsiveness, and Vulnerability. Costa and McCrae (1984) showed that all these are related to psychological well-being, negative affect, and lower life satisfaction. For H. Eysenck and Eysenck (1975), the high neuroticism scorer is "an anxious, worrying individual, moody and frequently depressed. … The stable individual, on the other hand, … is usually calm, even-tempered, controlled and unworried" (pp. 9–10).

From Traits to Linguistic Behavior

Working from these facets, we might expect that extraverts would use more positive emotional language (Warmth, Assertiveness, Positive Emotions), use more social language (Gregariousness), and produce more complex or extended utterances, reflecting their tendency to dominate interactions (Assertiveness, Excitement-Seeking). These predictions are consistent with those of Furnham

(1990). He suggested that extravert language has a more restricted rather than elaborated code and uses vocabulary less correctly. Most interesting, Furnham stated the extravert language is generally less formal than introvert language. This notion of formality can be understood in terms of explicitness (formality or context-independence) versus implicitness (informality or context-dependence), which Heylighen and Dewaele (2002) explored in greater detail. Their discussion assumed a notion of deixis following Levelt (1989), and they demarcated a group of expressions that must be anchored to some part of the spatiotemporal context of utterance if they are to be properly interpreted. Greater use of these expressions leads to greater implicitness (contextuality), whereas greater use of nondeictic expressions leads to greater explicitness (formality). They proposed that certain parts of speech (POS), such as verbs, are generally (although not invariably) deictic in nature, whereas others (such as nouns) are generally nondeictic. Implicitness can then be understood as a preference for pronouns, adverbs, and verbs, as opposed to nouns, adjectives, and prepositions. Heylighen and Dewaele argued that extraverts might produce implicit language because it requires less formulation effort, while relying more on the context for interpretation. Dewaele (2002a) suggested that implicitness arises because extraverts can take advantage of greater capacity in visual working memory and thus exploit extralinguistic context to a greater extent than other language producers. For now, we note that a tendency toward implicitness is consistent with extraverts using more verb-oriented language due to the personality facet of Activity.

Turning to neuroticism, we might expect that more neurotic individuals would use more negative emotional language (Anxiety, Angry Hostility, Depression), use more self-oriented language (Self-Consciousness, Vulnerability), and produce more emphatic utterances (Angry Hostility, Impulsiveness). Furnham's (1990) notion of implicitness may again be relevant. Anxiety—or perceived anxiety—has generally been found to be associated with greater repetitiveness in speakers (Bradac, 1990; Howeler, 1972). This may be because anxiety diverts resources away from sophisticated language production; if so, it would also lead to more implicit language, because this requires less effort to generate.

Previous Findings

To date, the majority of work exploring links between personality and language has focused on speech rather than writing, and the emphasis has been mostly on extraversion. Given that our focus is on written e-mail, we do not consider features specific to speech (such as amplitude or speech rate) any further but discuss in turn previous findings on lexical content and grammar (see also Dewaele & Furnham, 1999; Furnham, 1990; Pennebaker & King, 1999; Scherer, 1979; Smith, 1992).

Focusing first on lexical content in written text, significant results have been obtained using the Linguistic Inquiry and Word Count (LIWC) text analysis program (Pennebaker & Francis, 1999; see also Pennebaker, Francis, & Booth, 2001). LIWC is primarily concerned with lexical content, counting (context-free) occurrences of words or word stems that fall within predefined semantic and syntactic categories. For instance, words such as *could*, *should,* and *would* fall into the category of discrepancies. Although LIWC counts some syntactic features, such as pronouns, and verbs of various tenses, these are not derived from a POS analysis of the data. We return to methodological issues concerning LIWC and approaches like it shortly.

Pennebaker and King (1999) applied LIWC analysis to texts written by authors for whom (five-factor) personality information was available. The studies included multiple writing samples produced by a large number of participants in three quite different writing and topic contexts: daily diaries by patients at an addiction center, daily class assignments by summer school students, and abstracts to journal articles written by social psychologists. In their factor analysis study, a small set of linguistic factors grouping the LIWC features was derived and correlated with writers' scores on personality dimensions. There are parallels between this factor-analytic method and that adopted by Biber (1995). However, Biber used a broader set of linguistic features and a dictionary derived from the Brown corpus, aiming to analyze preexisting corpora to locate factors associated with register variation across genres. Three main factors were derived in the LIWC study: Making Distinctions, Immediacy, and the Social Past. Extraverts used language associated with the Social Past and avoided language associated with Making Distinctions; neurotic individuals used language associated with the Immediacy factor. Moving beyond Pennebaker and King's language factors and examining in more detail the relations between the personality dimensions and individual LIWC variables reveals the following. High extraverts use more social process (such as *talk* or *friend*), positive emotion words (*happy*, *good*), and inclusives (*and*, *with*) and fewer negations (*no*, *never*), tentative words (*maybe*, *perhaps*), exclusives (*but*, *without*), causation words (*because*, *hence*), negative emotion words (*hate*, *worthless*), and articles (*a*, *the*). High neurotics use more first-person singular (*I*, *my*) and negative emotion words and fewer positive emotion words and articles.

Gill (2004) analyzed a corpus of anonymized e-mail data with factor analyses and multiple regression analyses using both the LIWC dictionary and a dictionary based on the MRC psycholinguistic database (Wilson, 1987). The latter is a machine-readable dictionary, compiled in the 1980s from a number of sources, primarily to support psycholinguists developing experimental materials. It contains tens of thousands of words, with up to 26 linguistic and psycholinguistic attributes for each word. Gill's elicited, rather than naturally occurring, data was originally

gathered to allow a level of control over the variability in topic and audience. Factor analysis located dimensions differing only in minor details from those found by Pennebaker and King (1999), and it was found that both the topic-controlled LIWC dictionary and the MRC dictionary accounted for some variance in the data. Using this version of the LIWC dictionary, Gill could explain 8% of the variance in extraversion score and 11% in neuroticism score; the MRC dictionary helped explain 5% and 14% of the variances, respectively. The topic-controlled version of the LIWC follows Pennebaker and King in removing words associated with personal concerns. However, Pennebaker and King also controlled for genre, requiring that any linguistic variable to be included in the analysis had a minimum frequency of 1% in the corpus. If we do this, the variance in scores that is explained falls to 0% for extraversion and 11% for neuroticism. In passing, we note that level of neuroticism correlated positively with use of inclusive words and first-person pronouns.

At a higher linguistic level, analysis of speech acts showed that extraverts initiate more individual and group laughter, use more self-referent statements, and talk more (Gifford & Hine, 1994). In a study of conversational dyads, coding of the speech acts found that introverts used more hedges and problem talk and extraverts expressed more pleasure talk, agreement, and compliments, with content focusing more on extracurricular activities. However, significant differences were not found between the groups for talk time or number of speech acts (Thorne, 1987). This is a little surprising, because other studies have shown that extraverts use a greater total number of words (Carment, Miles, & Cervin, 1965).

On the subject of implicitness and patterns of use of POS, it has been shown that extravert speech has higher counts of pronouns, adverbs, verbs, and total number of words (taking *zestful* to be a synonym for extravert, cf. Furnham, 1990; see also Dewaele & Furnham, 1999). These characteristics of extravert language are also found for nonnative speakers. Using factor analysis of syntactic tokens from second-language speakers, Dewaele and Furnham (2000) confirmed an extravert preference for implicit language and an introvert preference for explicit language. This finding held in both informal and formal situations and mirrored previous analyses of the individual linguistic categories (Dewaele, 2001). Additionally, Heylighen and Dewaele (2002) noted that introvert language features tend to be closely related to those of formal language. A further finding was that extraverts demonstrated lower lexical richness in formal situations (Dewaele & Furnham, 2000). Cope (1969) also noted a lower lexical diversity (measured as type–token ratio) for extravert native English speakers. However, this is less reliable, given that extraverts also use a greater total number of words and thus may be a length effect (cf. Dewaele & Furnham, 2000). Dewaele (2002b) found that in third-language English production, there was a positive relation between neuroticism and communicative anxiety. Hence, our previous suggestion, that neuroticism may also relate to implicitness, is worthy of further investigation.

CURRENT HYPOTHESES

We here indicate the particular hypotheses we tested on the corpus of e-mail text and how they were measured. We relate the traits to types of linguistic behavior and indicate what differences we expected to find. The hypotheses are framed in terms of preferences in text generated by authors at the high end of a given personality dimension, as compared with authors at the low end of that dimension. Inverting these expectations gives the predictions for the low end. Suppose we say that a type of author (say, a high extravert) "prefers" a type of term (say, a social process expression). Operationally, this means that we expect high extraverts to either use instances of the type (such as *meet*) with a higher relative frequency than do low extraverts or use collocations involving that term (such as *I met*) more frequently than low extraverts.

Extraversion Hypotheses

Positivity. Warmth, Positive Emotions: Extraverts will prefer terms indicating positive emotions and fewer negative emotions and negations. Assertiveness: They will disprefer tentative expressions, such as hedges like *possibly.*

Sociability. Gregariousness: Extraverts will prefer third-person pronouns and proper names to refer to other people. They will prefer social process terms.

Complexity. Assertiveness, Excitement-Seeking: Reflecting their desire to communicate at length, extraverts will prefer to produce more complex utterances that link together several concepts in a sequence. They will link clauses and constituents by preferring more conjunctions, more clausal connectives, and more informal or nonstandard punctuation (ellipsis, exclamation, hyphenation).

Implicitness. Activity: Extravert language will prefer more adverbs, pronouns, and verbs and will disprefer nouns, adjectives, and prepositions. Dispreference for nouns also leads to dispreference for articles.

Neuroticism Hypotheses

Negativity. Anxiety, Angry Hostility, Depression: High neurotics will prefer terms indicating negative emotion and negation and disprefer positive emotions.

Self-concern. Self-Consciousness: High neurotics will prefer first-person singular pronouns over other second- or third-person pronouns or proper names. Vulnerability: They will prefer inclusive words, supposing that these indicate a desire for emotional attachment.

Emphasis. Angry Hostility, Impulsiveness: High neurotics will prefer nonstandard and multiple punctuation, to underline their attitude statements.

Implicitness. Anxiety: High neurotic language will prefer more adverbs, pronouns, and verbs and dispref nouns, adjectives, and prepositions. Pronoun use will be preferred but differentiated as noted previously. Dispreference for nouns leads to dispreference for articles.

ANALYTIC METHOD: FROM TOP-DOWN TO BOTTOM-UP

Earlier, we mentioned results from Gill's (2004) e-mail corpus. The amount of variance in personality explained by linguistic features was rather limited. In particular, the conservative LIWC analysis left no explanation of variance for extraversion.

There are two obvious possibilities. One is that the e-mail corpus simply did not possess the normal features associated with extraversion; indeed, perhaps the fact that the e-mail was elicited in an experiment also meant that it is not like "real" e-mail. The other is that the dictionaries were missing some of the relevant linguistic indicators. The former option does not seem right. We have already mentioned that as part of the dictionary-based study, Gill (2004) replicated the factor structure uncovered by Pennebaker and King (1999), with some minor differences. So the e-mail genre is relatively similar to the range of texts previously studied. It is also true that text elicited under laboratory conditions may differ significantly from naturally occurring text. However, the e-mail corpus can be compared with the relevant section of the British National Corpus, comprised of postings on a sports e-mail list. This shows that, in terms of Heylighen and Dewaele's (2002) contextuality/formality, the two are very similar (Nowson, Oberlander, & Gill, 2005). In this respect at least, the elicited e-mail was a reasonable approximation of natural e-mail and differed in predictable ways from naturally occurring personal weblogs. It follows, then, that there instead were problems in the application of dictionary-based analysis techniques.

First, Ball (1994) noted 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 relevant to LIWC, due to the size of its dictionaries; despite the inclusion of words and word stems to broaden potential matches, there are only around 2,000 words, compared with the 40,000 of the MRC database. A corollary is that the incorporation of systematic nonstandard features (such as words or spellings) in the analysis was precluded. In response, more recent work has adopted latent semantic analysis as a bottom-up alternative method (Campbell & Pennebaker, 2003). Although this is a data-driven approach, it expresses its findings in terms of vector measures for the texts. Hence, the results are less easily in-

terpretable than those from multidimensional analyses, which allow examination of the linguistic features that compose the factors.

Another limitation of content-analysis techniques is directly acknowledged by Pennebaker and King (1999) in relation to LIWC. Because a word's context is not taken into account, LIWC cannot say *how* it is used. Hazards include "context, irony, sarcasm, or … multiple meanings of words" (p. 1297). Disambiguation of word senses is less of a problem for the MRC psycholinguistic analysis, because it uses POS information, but contextual information has still been ignored in these analyses.

Therefore, instead of top-down approaches, we followed Tribble (2000) in adopting data-driven techniques from computational corpus linguistics; specifically the analysis of n-grams. The set of n-grams contained in a text is the set of all distinct sequences of *n* words; a unigram is a word sequence of length one (hence, it is just a word), whereas a bigram is a word sequence of length two, and so on. For instance, ignoring punctuation for now, a seven-word text such as *the grey dog chased the grey cat* contains six bigrams: *the grey*, *grey dog*, *dog chased,* and so on. We see that the bigram (*the grey*) occurs twice, so, in this case, there are only five distinct bigrams. N-gram analysis has previously been put to a variety of uses. For our purposes, it is especially relevant that n-grams have been used to characterize multiword terms that distinguish specific types of texts (Damerau, 1993).

Because n-gram analysis is a data-driven approach, all expressions are potentially relevant—not just those in a predefined dictionary. Also, the problem of context insensitivity is at least partially alleviated, because in calculating the probability of groups of terms, or n-grams, occurring together, it captures some of the contextual information of language use. It thus provides us with potential insight into differences in language structuring and the use of formulaic language, as noted by Wray and Perkins (2000).

CORPUS COLLECTION

Participants

One hundred and five current or recently graduated university students participated in this experiment. All participants were recruited via an e-mail sent by the experimenter. They were not remunerated for their participation.

A sociobiographical questionnaire and the Eysenck Personality Questionnaire–Revised (short version; S. Eysenck et al., 1985) were administered to gain information about the participants' backgrounds and personalities. Thirty-seven were men, and 68 were women. The mean age of participants was 24.3 years. Fifty-three were studying (or had studied) at an undergraduate level and 52 at a postgraduate level; the mean number of years of higher education was 4.2. All

spoke English as their first language. Ninety-five were of United Kingdom/Irish origin, seven North American, and three Australasian. Scores on the personality dimensions were as follows: psychoticism, *M* score = 2.90, *SD* = 1.7 (normative score for men is 3.08, for women, 2.35); extraversion, *M* score = 7.91, *SD* = 3.3 (normative score for men is 6.36, for women, 7.60); neuroticism, *M* score = 5.51, *SD* = 3.2 (normative score for men is 4.95, for women, 5.90); and Lie scale, *M* score = 3.48, *SD* = 2.2 (normative score for men is 3.86, for women, 2.71).

Materials

The experiment was conducted online via an HTML form that participants filled in and then submitted over the Internet.

The Web page had a simple design. It first gave an introduction and an estimate of the time required to complete the form, along with contact details, and it indicated that all responses would be treated in confidence and suitably anonymized. The second part of the form was for the collection of sociobiographical and personality information, with the results just noted. The final part consisted of the two message-writing tasks. Participants were first instructed "If during either of the following writing tasks, you are 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 into which participants could type, 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.

The second writing task was similar; participants were instructed to write about their plans for the week ahead. On final submission of the form, the participant was thanked, and the form was processed to check for any missing obligatory information. On acceptance of the form, the participant was given the contact details of the experimenter for any follow-up.

Preparation

Under these conditions, 105 participants provided two e-mail texts each, producing around 65,000 words in total. Apart from anonymization, preediting of these elicited texts was kept to a minimum so as to retain as much individuality as possible (for example, nonstandard words and spellings to imitate sounds). Such informal linguistic strategies, along with a relaxed attitude toward typographical errors, are regarded as a feature of e-mail (Baron, 1998; Colley & Todd, 2002). However, a distinction was made between intentional nonstandard spellings for communicative effect (such as *ohhhh*, *auld,* or *poptastico*) and spelling errors (such as *hte*, *abotu,* or *celecbrating*). A basic spell-check was carried out (using the standard emacs spell-checker; Stallman, 1994), and the resulting texts were hand-corrected 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).

STRATIFIED CORPUS COMPARISON

To analyze the prepared corpus, we used techniques from comparative corpus linguistics and defined a "reference corpus" from authors with a personality profile that is not extreme on extraversion or neuroticism. We then compared authors from each end (high or low) of each personality dimension with this neutral (or "mid") group. To control for individuals who may be extreme on more than one dimension, we also ensured that authors representative of the extreme groups were neutral on the other dimension.

To gain sufficient material for corpus comparison, it was necessary to group together texts by individuals who differed from one another in their precise personality scores. Hence, the analysis effectively considered features associated with groups rather than individuals. Nonetheless, by maintaining the reference corpus (as an intermediate point on a dimension), graded and potentially nonlinear effects could be detected. There is some evidence that such effects may arise in personality language studies (Gill, Harrison, & Oberlander, 2004), so a three-way stratified corpus comparison allowed a check on the behavior of linguistic features along a dimension. By contrast, other studies have usually assumed a binary division of the data, with categories such as native–nonnative, young–old, or higher–lower class language users (Aarts & Granger, 1998; Granger & Rayson, 1998; Milton, 1998; Rayson, Leech, & Hodges, 1997).

The goal in this analysis was to identify words (unigrams) or strings of words (n-grams) that formed reliable collocations for one group but not for another; these can then be considered *distinctive* collocations. This gives a new way to explore the link between personality and lexical content. We also used stratified

comparison, but on POS rather than words, to help explore the link between personality and grammar.

Method

Procedure. The full corpus of elicited texts was stratified into subcorpora as follows. High and low personality group samples were created by splitting them at greater than 1 *SD* above and below the Eysenck Personality Questionnaire–Revised score for each dimension. Authors had to be within 1 *SD* on the dimension other than the one for which they were extremely high or low. Furthermore, all texts that were within 1 *SD* across both personality dimensions were assigned to the personality–neutral mid subcorpus.

The resulting sizes of the subcorpora were as follows. There were 9,428 words (17 authors) for the high extraverts and 6,475 words (16 authors) for the low extraverts. There were 5,621 words (10 authors) for the high neurotics and around 7,073 words (12 authors) for the low neurotics. The mid group contained more than 13,304 words (30 authors). Stratification thus left us with 65% of the words from the original corpus, and no subgroup contained fewer than 10 authors.

Analysis. First we used a version of the corpus tokenized using the CLAWS tagger (available via the Wmatrix tool; Rayson, 2003) and lemmatized. The parser treated "fused forms" as composed of separate words; for instance, *can't* was taken to be composed from *ca* and *n't,* and in subsequent analysis *ca* was assimilated to *can*. Equally, some multiword expressions, such as *of course,* were treated as single orthographic units. The process also provided some basic annotation, for instance marking sentence boundaries (represented as <NC>) and ellipsis (<E>). The latter included cases of multiple full stops, as in the *well maybe …* example in Table 4. By lemmatizing (or stemming), minor variants of words could be collapsed together, increasing the power of the analysis. In such a processed corpus, words such as *play*, *plays*, *played*, or *playing* were all realized in the base form of the verb *play*; sentence boundary indicators and punctuation markers (such as <NC> and <E>) were collapsed into <p>.[1] More important, in our data there were instances of proper nouns, such as names of places (*Edinburgh*), names of people (*Dave*), or days of the week (*Saturday*). These tended to be too specific to allow broader patterns of language usage to emerge or for the results to be easily generalized; therefore, a further script was used to collapse proper names into NP1, except for names of days, which were collapsed into NPD1.

[1]It is worth noting in passing that the total number of <p> in the analyzed corpus was 4,738, of which 2,438 were <NC>. Because most, but not all, sentence boundaries were also marked with explicit punctuation, this indicates that there was relatively little midsentence punctuation in the corpus.

Second, to identify robust collocations in the tagged subcorpora, we calculated one- to five-word n-grams and did not use rank or frequency cutoffs during calculation, but initially selected features with a frequency ≥ 5. This enabled an accurate log-likelihood statistic (G^2) comparing the features' rates of occurrence between groups to be calculated (cf. Rayson, 2003). We used n-gram software (Banerjee & Pedersen, 2003) to compute significance (also G^2) for two- and three-gram collocations.

Finally, to identify those robust collocations that distinguish one group from another, we made a three-way comparison of the linguistic features across the high–mid–low corpora for each group. We calculated the relations among the three groups; for each feature in each corpus we identified its frequency and relative frequency and then, where relevant, the relative frequency ratios and log-likelihood between high–low, high–mid, and low–mid groups. An author could contribute to only one group; for instance, an individual's text could not contribute to counts of words or word sequences for both the high extravert and the low neurotic groups. Hence, the independence of one (personality) group from another can be assumed, as in other studies (e.g., Rayson et al., 1997). This allowed us to compare the relative usage and statistical significance of the difference in the use of features between groups.

Results

In this section we present the results from the three-way stratified analysis. Because we examined only expected frequencies of five or more—which compare more reliably with the chi-square distribution—we included results with a critical value of 10.83 or greater. We took this to be equivalent to reaching $p \leq .001$ significance, and those results with a critical value of 15.13 or greater were taken to be equivalent to reaching $p \leq .0001$ significance (cf. Rayson, 2003).

At least two kinds of features could be associated with, say, high neuroticism: n-grams that were overused by high neurotics and n-grams that were underused by low neurotics. Tables 1 and 2 list, for each dimension and each extreme subgroup, the features that were associated with that group either via their overuse of the feature or an opposite group's underuse.

There were reasonable numbers of distinctive collocations. Sixty-three n-gram features reached the critical value of 10.83 ($p < .001$) for extraversion and 59 for neuroticism. Of these, a substantial proportion also reached the 15.13 critical value ($p < .0001$): 32 and 21, respectively. The latter set includes several n-grams that represent repeated punctuation (most of which correspond to multiple exclamation marks). It is notable that the vast majority of distinctive collocations (almost 90%) involved more than one word or punctuation mark; these would not be found by single-word dictionaries. Furthermore, a substantial proportion were predicted neither by theory nor by prior research.

The tables group the collocations by drawing together those that involve terms classified by the LIWC or some other common factor, such as the presence of punctuation. Only a small number of collocations ended up being over- or underused by the mid group (three and two reached the 15.13 critical value for extraversion and neuroticism, respectively). This demonstrates (rather than simply assumes) that linguistic behavior is linear. For this reason, mid collocations were omitted from the tables and are not discussed further in this article.

Obviously, if it contains more than one word, a particular collocation can appear in more than one place in a table. Equally, a collocation may be distinctive on more than one dimension. A notable example of this involves the verb *get*. *Get to* is a collocation preferred by high extraverts and low neurotics. By contrast, *get on* is only relevant to the neuroticism dimension, where it is preferred by high neurotics. We now consider the two personality dimensions in turn in terms of the various collocation types.

Extraversion. Consider first the features we expected to be related to extraversion in participants. Note that the table is divided into five subsections. Most of the language features listed, except for those involving punctuation, are taken from the LIWC. The table has annotations indicating which subgroup was predicted to make greater use of a given feature. To help visualize the results, Table 3 contains example texts, from high-extraversion and low-extraversion participants.

Collocations relating to positivity appear in the attitude section of Table 1. High extraverts have one collocation involving positive emotion, which contains the word *cool*. In fact, this item is not in the original LIWC lexicon for positive emotions but seems a good candidate for a positive emotion term. Of course, the word can be used with different meaning in other contexts, such as weather discourse (as the first example in Table 3 makes clear). Hence, including it in an LIWC category runs the risk of treating it context insensitively. Whenever we assimilated a collocation to an LIWC category, it is indicated by a mark in Tables 1 and 2. Low extraverts have two collocations, both involving *play*. What the meaning of such specific collocations might be is an issue that we return to in the Discussion section. Both groups had one collocation involving negation. High extraverts had two collocations that can be associated with certainty, whereas low extraverts had eight associated with tentativity.

Most collocations relating to sociability fell within the nominals section of the table. High extraverts had no collocations involving third-person pronouns, whereas low extraverts had three. Notably, high extraverts had six first-person pronoun collocations, compared with four for low extraverts. However, high extraverts did also have seven collocations involving proper names, whereas low extraverts had four. Other collocations relating to sociability include social pro-

TABLE 1
Summary of Tokenized, Lemmatized Analysis for Extraversion Preference

Feature Type	*High Extravert*	*Low Extravert*
Attitude		
Positive emotion	⊙ [*cool* <*p*>]$^{\oplus}$	[*i play*]′ [*play*]′
Negative emotion	–	⊙ –
Negation	[*will not*]	⊙ [*not really*]
Tentative/certain	[*be really*]$^{\oplus}$′ [*be so*]$^{\oplus}$′	⊙ [*be supposed*]′ [*be supposed to be*] [*be supposed to*] [*supposed to be*] [*supposed to*] [*supposed*] [*fairly*]$^{\oplus}$ [*not really*]$^{\oplus}$
Social processes	⊙ [*i meet*]	–
Conjunction/connectives		
Inclusives	⊙ [*and*]′ [*and* NP1]′ [*with* NP1]′ [NP1 *and* NP1] [NPD1 *and*] [*and see*] [*work and*]	[*with a*]
Exclusives	[*that be*]	⊙ [<p> *although*]′ [*although i*] [*although*] [*there* <*p*>]′ [*off*]′
Others	⊙ [<p> *then*]′ [*then i*]′ [<p> *well*] [<p> *which*]	–
Nominals		
1st person pronouns	[NPD1 *i*]′ [*then i*]′ [*to my*]′ [*we be*]′ [*i meet*] [*what i*]	[*i play*]′ [*i get*]′ [*i know*] [*although i*]
3rd person pronouns	⊙ –	[*because it*] [*of they*] [*they*]
Proper names	⊙ [*and* NP1]′ [*with* NP1]′ [NP1 *and* NP1] [NPD1 *and*] [NP1 NP1] [NPD1 *i*] [NPD1 *will*]	[*the* NPD1]′ [<p><p> NPD1 <p>]′ [<*p*> NPD1 <p>]′ [<p> NPD1]′
Articles	[*from the*] [*the week*]	⊙ [*the* NPD1]′ [*with a*]
Punctuation		
Multiple	⊙ –	[<p><p> NPD1 <p>]′
Punctuation-word	[<p> *take*]′ [<p> *then*]′ [<p> well]′ [<p> *which*]′ [<p> *what*]	[<p> *although*]′ [<p><p> NPD1 <p>]′ [<p> NPD1 <p>]′ [<p> NPD1]′
Word-punctuation	[*year* <p>]′	[<p><p> NPD1 <p>]′ [<p>NPD1 <p>]′ [*there* <*p*>]′ [*day* <p>]′ [*weekend* <p>]′
Other	[*to go*]′ [*get to*]′ [*of it*] [*what be*]	[*know*]′ [*of work*]′ [*know that*] [*essay*]

Note. All n-grams reached the 10.83 critical level ($p \leq .001$). ′ designates those n-grams that reached the 15.13 level ($p \leq .0001$). Where LIWC categories are relevant, $^{\oplus}$ designates items containing words that could be listed in the official LIWC dictionary but are not. ⊙ indicates, for a category, which group was predicted to possess more items in that category.

TABLE 2
Summary of Tokenized, Lemmatized Analysis for Neuroticism Preference

Feature Type	*High Neurotic*	*Low Neurotic*
Attitude		
Positive emotion	[*like to*]′ [*would like*]	⊙ [*a good*] [*nice* <p>]⊕
Negative emotion	⊙ –	–
Negation	⊙ –	–
Tentative/certain	–	–
Social processes	–	–
Conjunction/connectives		
Inclusives	⊙ [<p> *and*]′ [*and*] [*and see*]	–
Exclusives	[<p> *or*]′ [*though* <p>]⊕	[*but it*]
Others	[<p><p> *well*]′ [<p> *well*]′	[<p> *so*]′ [<p> *then*]′ [<p> which]′ [<p> *anyway*]
Nominals		
1st person pronouns	⊙ [<p><p> *we*]′ [<p> *we*]′ [*well i*]′ [<p> *well i*]	[*which i*]′ [*i get*]′
3rd person pronouns	[*its*] [*of they*]	⊙ [<p> *he*]′ [*about it*]′ [*about it* <p><p>] [*about it* <p>] [*but it*] [*it do*]
Proper names	–	⊙ [NPD1 <p>]′ [NP1 *come*]
Articles	[*all the*]′ [*see the*] [*the film be*] [*the film*]	⊙ [*a good*] [<p> *the*]
Punctuation		
Multiple	⊙ [<p><p><p><p><p><p>]′ [<p><p><p>]′ [<p><p><p><p><p>]′ [<p><p> *we*]′ [<p><p> *well*]′	[*about it* <p><p>]
Punctuation-word	[<p> *and*]′ [<p> *or*]′ [<p> *we*]′ [<p><p> *we*]′ [<p> *well*]′ [<p><p> *well*]′ [<p> *how*]	[<p> *he*]′ [<p> *so*]′ [<p> *then*]′ [<p> *which*]′ [<p> *anyway*] [<p> *the*]
Word-punctuation	[*time* <p>]′ [*soon* <p>] [*though* <p>]	[NPD1 <p>]′ [*about it* <p><p>] [*about it* <p>] [*nice* <p>] [*night* <p>] [*party* <p>] [*well* <p>]
Other	[*to work*]′ [*be write*] [*film be*] [*film*] [*get on*] [*have to go*] [*to spend*] [*up to*]	[*be in*] [*get to*] [*go on*] [*of time*] [*still have*] [*which*]

Note. All n-grams reached the 10.83 critical level ($p \leq .001$). ′ designates those n-grams that reached the 15.13 level ($p \leq .0001$). Where LIWC categories are relevant, ⊕ designates items containing words that could be listed in the official LIWC dictionary but are not. ⊙ indicates, for a category, which group was predicted to possess more items in that category.

TABLE 3
Extracts From Texts Written by High and Low Extravert Participants

High extravert	
he1	I spent New Year in X which was really cool! Well, actually a bit cold […] [*be really*], [*cool* <p>], [<p> *well*] [ADJ <p>] (*cool!*), [ADV O] (*actually a*)
he2	Hopefully I can persuade Y to come which would be really exciting! [*be really*] [ADJ <p>] (*exciting!*)
he3	It was a really cool night, and the guys we were with were so friendly. [*we be*], [*be so*] [ADJ <p>] (*friendly.*)
he4	[…] fluid on his lungs. Well I better go. Take care and speak soon […] [<p> *well*], [<p> *take*] [ADV VBN] (*better go*)
he5	It's really funny and Ben Stiller is just delightful. [*and* NP1] [ADJ <p>] (*delightful.*)
Low extravert	
le1	I am in the mood for just hanging out and not really doing much. [*not really*]
le2	I was supposed to be in work but I just couldn't get out of bed […] [*be supposed to be*] [PRN ADV] (*i just*)
le3	I'm done here pretty much, although I really have lots of work that I should be getting on with now […] [<p> *although*], [*although i*] [PRN ADV] (*i really*), [VPP ADV] (*done here*)
le4	I like it because it's a change from looking at a computer all day. [*because it*]
le5	We got another few shows in and lots of nice lunches and a big walk on the Sunday. [*the* NPD1]

Note. The extracts give examples of some of the word and parts of speech n-grams that are distinctive for the given groups. Most of the proper names have been replaced here by letters.

cesses, listed under attitude. High extraverts had one collocation in this class; low extraverts had none.

A number of collocations relevant to complexity fell within the connective section of the table. In total, high extraverts had 12 collocations in this group, whereas low extraverts had 5. It is notable that high extraverts collocations mostly involved inclusives, whereas low extraverts collocations were nearly all exclusives. The rest of the complexity collocations fall under the punctuation section of the table. High extraverts had no multiple punctuation collocations, whereas low extraverts had one. As noted previously, most sentence boundary markers in the corpus were accompanied by an explicit punctuation mark; hence, pairs of <p> are not remark-

able. High extraverts had five initial punctuation collocations, where a word or words follows at least one punctuation mark, and one final punctuation collocation, where at least one word precedes at least one collocation. Low extraverts had four initial collocations (three of which were closely related) and five final collocations (two of which were also counted initial, so that a word was found between at least two punctuation marks).

Implicitness results are given in more detail when we consider results on POS. However, we note here that in the nominal section of the table, both high and low extraverts had two collocations involving articles.

Neuroticism. The same approach to collocation types is useful for describing neuroticism. Again, to help visualize the results, Table 4 contains example texts from high neuroticism and low neuroticism participants.

Collocations relating to negativity appear in the attitude section of Table 2. Neither the high neurotics nor the low neurotics had any collocations involving either negative emotions or negations. Both had two collocations involving positive emotions.

Collocations relating to self-concern fell in both the nominal and connective sections of the table. High neurotics had four collocations involving first-person pronouns, as compared to two involving third persons and none involving proper names. Low neurotics had two collocations involving first-person pronouns, as compared to six involving third persons and two involving proper names. Regarding inclusive words, high neurotics had three collocations that involved them, whereas low neurotics had none.

Collocations relating to emphasis are in the punctuation section of the table. High neurotics had five collocations involving multiple punctuation, whereas low neurotics had one. Unlike the low extraverts, high neurotics' collocations included sequences of more than two <p>.

Again, implicitness results are given in more detail when we consider results on POS. However, we note here that in the nominal section of the table, high neurotics had four collocations involving articles, and low neurotics had two.

SYNTACTIC ANALYSIS OF THE CORPUS

Method

The personality corpus was tagged using the Penn POS tagset (Marcus, Santorini, & Marcinkiewicz, 1994) and the MXPOST tagger (Ratnaparkhi, 1996). Further processing replaced words with POS tags corresponding to 10 broad categories, as implemented in the electronic version of the *Shorter Oxford English Dictionary* that is incorporated into the MRC Psycholinguistic Database (Wilson, 1987). These are:

TABLE 4
Extracts From Texts Written by High and Low Neurotic Participants

High neurotic	
hn1	There are a few people I haven't seen for a while that I would like to catch up with.. well maybe.. [*would like*], [*like to*], [<p><p> *well*] [PRN O VBN O VBN] (*i would like to catch*)
hn2	I'm so rubbish at it though, and it's taking ages to do each one. [*though* <p>] [<p> CONJ] (*, and*)
hn3	We got back from W X last Wednesday night. We stayed over at Y and drove back to Z on Thursday. [<p><p> *we*] [VBN PRP] (*stayed over*)
hn4	I have nothing else to say!!!!!!!!!!!!!! [<p><p><p><p><p>] [<p><p><p><p><p>] (*!!!!!*), [<p><p><p><p><p>] (*!!!!*), [<p><p><p><p>] (*!!!*)
hn5	[…] I'm finally getting a chance to make the most of things and live how I want to live without having to think about someone else all the time. [*all the*] [ADV VBN] (*finally getting*), [VBN PRP] (*live without*), (*think about*)
Low neurotic	
ln1	One weekend two of my friends got married which was really nice - [*which*], [*nice* <p>]
ln2	She's hired out a club and is getting really excited about it. [*about it*], [*about it* <p>]
ln3	That's the plan, anyway. [<p> *anyway*] [<p> ADV] (*, anyway*),
ln4	I suppose I should do some of my dissertation as well, but right now all I really want to do is go to sleep. [*well* <p>] [CONJ ADV] (*but right*), [PRN ADV] (*i really*), [ADV ADV] (*as well*), (*right now*), [ADV O] (*now all*)
ln5	Not that I'd completely go on his opinion but he seemed a good person to ask. [*go on*], [*a good*]

Note. The extracts give examples of some of the word and parts of speech n-grams that are distinctive for the given groups. Proper names have been replaced here by letters.

noun (NN), adjective (ADJ), verb (VBN), adverb (ADV), preposition (PRP), conjunction (CONJ), pronoun (PRN), interjection (INT), past participle (VPP), and other (O). In addition to these categories, we also made use of <p> indicating punctuation; <ends>, which indicated the end of the e-mail texts; and NA, which indicated that a feature was not recognized as belonging to any of the previous categories.[2]

[2]Thus, the categories are a superset of those in the MRC database, but it should be noted that the labels for the categories differed in places from those used in both the Penn tagset and the MRC database; for instance, we used PRP whereas the MRC uses R.

The resulting general syntactic version of the corpus was then divided into the high–mid–low stratified corpus groups and analyzed, as in the previous section. We first discuss the results of the unigram analysis for each dimension. We then display the results of the overall n-gram analyses (one through five item sequences) together, as in the previous lemmatized word n-gram analysis. Once again, the small number of collocations found for the mid group are not reported.

Unigram Syntactic Analysis Results

In Table 5, we list, for each dimension and each subgroup, features associated with that group either via their overuse of the feature or an opposite group's underuse.

For extraversion, conjunction (CONJ) and adjectives (ADJ) are characteristic of high extraversion, but there are no POS characteristics of low extraversion. For neuroticism, conjunction (CONJ) and punctuation (<p>) are characteristic of high neuroticism and adverbs (ADV) and nouns (NN) of low neuroticism.

For these POS tag unigram results, we note the generally modest levels of significant differences among groups relative to the previous n-gram analyses. We may take this to indicate that the personality groups generally use quite similar proportions of the relevant POS. However, the POS may also occur in different contexts or sequences, thus indicating differences in the way they are used. We therefore turn to the results of the n-gram analysis of the syntactic tag data.

N-Gram Syntactic Analysis Results

Results using the reduced syntactic category tags reached higher levels of significance than the unigrams, so we considered only those that reached the critical value of 10.83 ($p \leq .001$). Twenty-four n-gram features reached this value for extraversion and 34 for neuroticism. Of these, the majority also reach the 15.13 critical value ($p \leq .0001$): 15 and 20, respectively. The stronger features are predominantly bigrams, exceptions being the longer n-grams for punctuation found

TABLE 5
Summary of Unigram
Parts of Speech Analysis

Personality Type	*Distinctive Unigrams*
High extraverts	[CONJ] [ADJ]
Low extraverts	—
High neurotics	[CONJ] []
Low neurotics	[NN] [ADV]

Note. Unigrams reached the 3.84 critical value ($p \leq .05$).

for neuroticism. A concise view is given in Tables 6 and 7; examples can be found in Tables 3 and 4.

Extraversion. From the unigram analysis, we are particularly interested in collocations involving conjunctions and adjectives (for high extraversion).

As far as conjunctions are concerned, high extraverts had two distinctive collocations ([CONJ ADV] and [PRN CONJ]); low extraverts had none. Regarding adjectives, we found two collocations for high extraverts ([<p> ADJ] and [ADJ <p>]). There were also two collocations for low extraverts ([NN <p> ADJ NN] and [NN NN <p> ADJ NN]).

TABLE 6
Summary of Parts of Speech Analysis for Extraversion Preference

Part of Speech Type	*High Extravert*	*Low Extravert*
Punctuation	⊙ [<p> NN]′ [<p> ADJ]′ [ADJ <p>]′	[O VBN <p> PRN]′ [NN <p> ADJ NN] [NN NN <p> ADJ NN] [NN VBN O <p>] [VBN <p> PRN] [VBN <p> PRN VBN]
Conjunction	[CONJ ADV]′ [PRN CONJ]	–
Noun	[<p> NN]′	⊙ [NN <p> ADJ NN] [NN NN <p> ADJ NN] [NN VBN O <p>] [ADV VBN PRN NN]
Adjective	[<p> ADJ]′ [ADJ <p>]′	⊙ [NN <p> ADJ NN] [NN NN <p> ADJ NN]
Preposition	[ADV VBN PRP]′	⊙ [ADV PRP]′
Pronoun	⊙ [ADV VBN PRN]′ [PRN CONJ] [PRN O VBN O]	[O VBN <p> PRN]′ [PRN ADV]′ [VBN <p> PRN] [VBN <p> PRN VBN] [ADV VBN PRN NN]
Verb	⊙ [ADV VBN]′ [ADV VBN PRN]′ [ADV VBN PRP]′ [PRN O VBN O]	[O VBN <p> PRN]′ [NN VBN O <p>] [ADV VBN PRN NN] [VBN <p> PRN VBN] [VBN <p> PRN] [VBN ADV O VBN O]
Past participle	–	[VPP ADV]′
Adverb	⊙ [<ENDS> ADV]′ [ADV ADV]′ [ADV O]′ [ADV VBN PRN]′ [ADV VBN PRP]′ [ADV VBN]′ [CONJ ADV]′	[ADV PRP]′ [PRN ADV]′ [VPP ADV]′ [ADV VBN PRN NN] [VBN ADV O VBN O]
Other	[ADV O]′ [O NA]′ [PRN O VBN O]	[O VBN <p> PRN]′ [NN VBN O <p>] [VBN ADV O VBN O]′
NA	[<ENDS> ADV]′ [O NA]′	–

Note. All n-grams reached the 10.83 critical level ($p \leq .001$). ′ designates those n-grams that reached the 15.13 level ($p \leq .0001$). ⊙ indicates, for a category, which group was predicted to possess more items in that category.

TABLE 7
Summary of Parts of Speech Analysis for Neuroticism

Part of Speech Type	*High Neurotic*	*Low Neurotic*
Punctuation	⊙ [<p><p><p><p><p>]′ [<p><p><p><p>]′ [<p><p><p>]′ [<p> O]′ [<p> ADV PRN VBN PRN] [<p> CONJ] [ADV <p> PRN O VBN]	[<p><p>]′ [<p> ADV ADV]′ [<p> ADV]′
Conjunction	[<p> CONJ] [CONJ VBN PRP] [VBN ADJ CONJ]	[CONJ ADV]′ [CONJ O]′ [CONJ VBN PRN]
Noun	–	⊙ [<ENDS> NN] [NN NN]
Adjective	[ADJ ADJ]′ [ADJ PRN VBN]′ [<ENDS> ADJ] [ADJ PRN] [ADV VBN ADJ]	⊙ [PRN ADJ]′ [PRP O ADJ ADJ]
Preposition	[VBN PRP]′ [CONJ VBN PRP]	⊙ [ADV ADV PRP]′ [PRP O ADJ ADJ] [PRP O PRP O]
Pronoun	⊙ [ADJ PRN VBN]′ [<p> ADV PRN VBN PRN] [ADJ PRN] [ADV <p> PRN O VBN] [ADV PRN VBN PRN] [PRN O VBN O VBN]	[ADV PRN]′ [PRN ADV]′ [PRN ADJ]′ [CONJ VBN PRN]
Verb	⊙ [ADJ PRN VBN]′ [ADV VBN]′ [VBN PRP]′ [<p> ADV PRN VBN PRN] [ADV <p> PRN O VBN] [ADV PRN VBN PRN] [ADV VBN ADJ] [CONJ VBN PRP] [PRN O VBN O VBN]	[CONJ VBN PRN]
Past Participle	–	[VPP ADV]′
Adverb	⊙ [ADV VBN]′ [<p> ADV PRN VBN PRN] [ADV <p> PRN O VBN] [ADV PRN VBN PRN] [ADV VBN ADJ]	[<p> ADV ADV]′ [<p> ADV]′ [ADV ADV PRP]′ [ADV ADV]′ [ADV O]′ [ADV PRN]′ [CONJ ADV]′ [VPP ADV]′ [PRN ADV]′
Other	[<p> O] [ADV <p> PRN O VBN] [PRN O VBN O VBN]	[ADV O]′ [CONJ O]′ [PRP O ADJ ADJ] [PRP O PRP O]
NA	[<ENDS> ADJ]	[<ENDS> NN]

Note. All n-grams reached the 10.83 critical level ($p \leq .001$). ′ designates those n-grams that reached the 15.13 level ($p \leq .0001$). ⊙ indicates, for a category, which group was predicted to possess more items in that category.

Neuroticism. Here we are most interested in collocations involving punctuation and conjunctions (for high neuroticism) and nouns and adverbs (for low neuroticism).

Regarding punctuation, we found high neurotics had a total of seven collocations involving punctuation (several involving multiple punctuation). Low neurot-

ics had three collocations (one involving double punctuation and two involving a mark followed by an adverb). Regarding conjunctions, we found that high and low neurotics had three collocations each.

Noun collocations were more sparse: There are none for high neurotics and two for low neurotics (of which one involved multiple nouns). Adverb collocations, on the other hand, were common. High neurotics had 5, including [ADV VBN], but low neurotics had a total of 9, and in fact 9 of their 12 most significant collocations involved adverbs. This is the most vivid result to emerge from the n-gram syntactic analysis, although we discuss the ramifications for implicitness in the next section.

DISCUSSION

Using data-driven techniques we have been able to investigate linguistic features that characterize the expression of personality in e-mail communication, without being restricted by predefined dictionaries. Recalling the predictions in the Hypotheses section, we discuss findings for the two dimensions in turn.

Extraversion

The original predictions involved positivity, sociability, complexity, and implicitness.

First, there is the issue of positivity. Both groups had positive emotion collocations, and each group also had their own distinctive negation collocation. However, tentativity did emerge among low-extravert collocations: *although*, *be supposed* (and its relatives), and *fairly.* Assuming that the high extraverts' *be really* and *be so* can be associated with certainty, and the low extraverts' *not really* with tentativity, there does seem to be a difference between the two groups in certainty and tentativity.

We expected sociability for high extraverts, but their pronoun biases seemed to work against it: High extraverts tended toward word collocations involving first-person pronouns, whereas low extraverts' pronoun collocations included a more even mix of first and third person. On the other hand, patterns of noun use suggested that high extraverts did refer to other people by name, particularly in conjoined noun phrases. In addition, high extraverts, but not low extraverts, had a collocation involving social processes.

We now turn to the complexity hypothesis. Nonstandard, multiple punctuation for high extraverts, such as ellipsis, are here rendered as single punctuation tags, so they do not appear in the distinctive word or POS collocations. The word collocations showed many more conjunction constructions for high extraverts than for low extraverts, and high extraverts' distinctive use of conjunctions was confirmed by the n-grams for POS. It was notable, also, that high extraverts' conjunction col-

locations included what LIWC would term inclusive words, such as *and* and *with*, whereas, the low extroverts had collocations involving *although*, an exclusive word. There is some evidence that high extroverts use a broader range of distinctive collocations involving initial punctuation (they had five as compared with four for low extroverts, but the latter included three related collocations to do with names of days of the week). These can be seen as introducing clause-initial connectives, such as *then*, *which*, and *what*.

Finally, following Dewaele and Furnham (2000), it was predicted that high extraverts would use more verbs, adverbs, and pronouns and low extraverts would use more nouns, adjectives, and prepositions. The unigram POS analysis did not support the overall predictions. It indicated that high extraverts use more conjunctions and adjectives. No other overall differences were found, although it is worth noting three points. First, as expected, high extraverts did have more pronoun collocations of at least one type: first person. Second, as expected, low extraverts had two collocations involving articles, but high extraverts also had two collocations involving articles. Finally, in the POS analyses, because we had split past participles from general verbs and added conjunctions, our categories were slightly finer-grained than Dewaele and Furnham's, which may affect the result. Another reason for the divergence from Dewaele and Furnham's results could be that they were largely dealing with speech, rather than computer-mediated writing, and with nonnative speakers. Perhaps implicitness is more closely related to, say, neuroticism, for native writers, and more closely related to extraversion only for nonnative speakers.

But before accepting this line of reasoning, we should also consider the results of the n-gram POS analysis. Elsewhere, we have suggested that at least two possibilities might be interesting (Oberlander & Gill, 2004). 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 the former group is using the POS in a more stereotypical range of contexts; the latter group is using the POS more flexibly. Second, when 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. As we noted, intuitions on this question are not firm, but we suggested that greater relative frequency might be associated with a greater range of use—and hence with perhaps fewer stereotypical collocations.

However, the results, which are based on a broader range of participant data than those of Oberlander and Gill (2004), suggest the opposite. In particular, high extraverts preferred conjunctions and adjectives overall and also tended toward more (and stronger) distinctive collocations involving these two POS.

One other relevant finding on implicitness can be derived. We totaled the number of collocations involving nouns, adjectives, and prepositions (eliminating du-

plicates) and did the same for verbs (including past participles), adverbs, and pronouns. We found that high extroverts had four distinct collocations on the explicit side (which includes the nouns) and nine collocations on the implicit side (which includes the verbs), but low extraverts also had fewer explicit collocations than implicit (five and eight, respectively). Because the balance in favor of implicit collocations is weakened but not reversed, we conclude that our results do not support the implicitness hypothesis for extraversion

Neuroticism

The original predictions involved negativity, self-concern, emphasis, and implicitness.

First, there is the issue of negativity. We found no collocations for negative emotion or negation; for positive emotion, we found high neurotics had as many as low neurotics.

Considering self-concern, from the results on word collocations, we found that high neurotics had no more collocations than low neurotics involving first-person singular. However, high neurotics did have first-person plural collocations, unlike the low neurotics. At the same time, low neurotics had eight collocations involving third-person pronouns and proper names, compared to just two for high neurotics. So, arguably, the smaller set of collocations involving reference to other persons for high neurotics reflects a less outward-looking discourse. The unigram POS results identify an overall high neurotics preference for conjunction (shared with the high extraverts). The word n-grams show that connectives for high neurotics include both inclusives, such as *and*, and exclusives, such as *though*. The former, at least, could fit with a drive for attachment. Low neurotics had no inclusives and only one exclusive. The POS collocation results add nothing further to this picture.

On the matter of emphasis, the high neurotics' preference for genuinely multiple punctuation is revealed in both word and POS collocations and reflects a particular use of exclamation marks. Although we did not frame the hypothesis earlier, it is also possible that an emphatic character might lead to greater use of both adjectives and adverbs. In fact, the POS unigram results did not indicate an overall high neurotics' preference for adjectives, and it was the low neurotics who proved to prefer adverbs.

Regarding implicitness, like high extraverts, high neurotics were predicted to prefer implicit language, using more verbs, adverbs, and pronouns and fewer nouns, adjectives, and prepositions. The unigram analysis did not support these predictions. It found that high neurotics used more punctuation (and conjunctions) and that low neurotics used more nouns and adverbs. Low neurotics were expected to be less implicit, and the preference for nouns fits the hypothesis, but the preference for adverbs does not.

In this case, considering the POS collocations does add something to the picture. Once again, we totaled the number of collocations involving nouns, adjectives, and prepositions (eliminating duplicates) and did the same for verbs (including past participles), adverbs, and pronouns. We found that high neurotics had 7 distinct collocations on the explicit side and 10 collocations on the implicit side. Low neurotics had 6 and 11, respectively. But by the implicitness hypothesis, we would have expected high neurotics to have had more bias toward implicit collocations than low neurotics. It is true that high neurotics have many more collocations involving verbs than low neurotics (nine to one). But overall high neurotics' collocations did not show a greater bias toward implicitness for two main reasons, both of which confound expectations. First, high neurotics had six collocations involving adjectives (compared to two for low neurotics). Second, low neurotics had nine collocations for adverbs (compared to five for high neurotics). Indeed, as well as their using more adverbs overall, a high proportion of low neurotics' strongest distinctive collocations involved those adverbs. It appears that low neurotics use them in a range of contexts and do so robustly. High neurotics use fewer adverbs and use them in less stereotypical contexts.

Questions

With the overall results in mind, a number of more general questions can be addressed.

First, one might ask to what extent the results derived via this bottom-up n-gram approach differ from those that can be found via top-down dictionary-based analyses. One response is that the vast majority of word and POS collocations involve more than element. This suggests that there are relatively few single-word "shibboleths" that might distinguish authors who score at different ends of one personality dimension. But top-down approaches do not claim that these exist: They merely assume that relative frequencies of use of some words may correlate with personality scores. Another response is that the n-gram approach helpfully shows that a word such as *really* may occur in different collocations for high and low scorers on a personality dimension. Recall that *be really* characterizes high extraversion whereas *not really* characterizes low extraversion. So our approach captures more context than does single word counting, although it still does not take context or semantics fully into account, as we noted when discussing the fact that *cool* has more than one meaning. At the same time, there is nothing to stop dictionary-based approaches from using multiword expressions, and that would certainly be one recommendation flowing from the results here. A final response is that the bottom-up approach reveals significant linguistic behavior not captured in existing dictionaries. The case of high neurotics' multiple punctuation shows that this is surely true.

In light of this, one might ask what all the collocations actually mean. In particular, whereas some appear to relate to the hypotheses we framed at the start, others are rather obscure. In response, we would concede that some of the distinctive collocations are intriguing but hard to explain, even post hoc. For instance, the low extraverts' preference for *i play* or the high neurotics' preference for *the film be* do not fit into any obvious patterns and may just be idiosyncrasies of the corpus. On the other hand, some of the other unexpected findings can perhaps be explained and are certainly worthy of further investigation. In this connection, we note the presence of high neurotics' collocations involving clause-initial *well*; this is a filler expression with little independent meaning but with connotations of concession. The low neurotics' use of clause-initial *anyway* (not included in the current LIWC dictionary) provides an interesting contrast. This discourse cue phrase is not concessive in the same way and is associated with "popping" up from embedded discourses. It could be that low neurotics are more likely to pop, instead of continuing an embedded segment, or it could be that they are more likely to generate embeddings in the first place. To follow up this study, subsequent work should aim to analyze larger bodies of naturally occurring text. With appropriate discourse-level annotation, this would allow proper investigation of the ways in which discourse structure is affected by differences between language producers.

A third question relates to the specifics of the stratification method we have chosen. The point is that members of our extreme subgroups on one dimension (such as high extraverts) are neutral scorers on the other dimension (they cannot be high neurotics or low neurotics). This has some advantages, but it does mean that we have deferred the study of interactions between dimensions. For instance, some of the predictions pointed to potential interactions. For example, if implicitness is relevant to both extraversion and neuroticism, then someone who scores low on one dimension but high on the other might not exhibit particularly implicit language. In fact (as we reiterate later), the implicitness predictions were not borne out. On the other hand, as noted previously, it was found that *get to* is a collocation preferred by high extraverts and low neurotics. So it is natural to ask what high extravert–high neurotics "get (up) to." Clearly, this is a question that should be considered in future analyses.

A fourth question is whether the relations between personality and text are specific to particular communication genres or media. A first response is that there must certainly be genre effects. Some of the collocations we have found are characteristic of e-mail but are neither confined to it nor required of it. For instance, multiple exclamation marks may occur in some other genres, such as personal letters, but they are less likely to appear in most others, including e-mail directed to business acquaintances. Equally, as Pennebaker and King (1999) noted, topic also has a noticeable effect: People are more likely to refer to days when recalling and predicting activities in the past and future week. That effect is likely independent

of whether communication is computer-mediated or not. On the other hand, it is interesting that for neuroticism, we found no special bias concerning emotion collocations (or first-person reference). Also, just as there was little evidence for negativity on this dimension, there was only limited evidence of positivity on the extraversion dimension. Low extraverts were notable for their tentativity, but other features did not add up to major differences in positivity. Taken together, these findings are consistent with the idea that e-mail can be less emotional than speech (Whittaker, 2003). In addition, we found no firm results concerning implicitness for either extraversion or neuroticism. It may be that authors do not find producing e-mail under laboratory conditions a very stressful activity, and hence differences in implicitness are not revealed in this context (cf. Dewaele, 2002b). The lesson from this is that, as well as gathering larger bodies of text for discourse-level analysis, follow-up studies should aim to gain from each author multiple texts across multiple contexts. This would allow investigation of the effects of assumed audiences on producers' language.

A fifth question is whether linguistic style (in our e-mail context) is a direct reflection of personality and whether the latter could perhaps be diagnosed from the former. Regarding this connection we note that this study focuses on personality projection, as opposed to personality perception. In studying projection, the only personality associated with the text is that derived from authors' own self-reports on personality questionnaires. By contrast, when studying perception, a second personality is associated with the text—that attributed to it by third-party readers of the texts. In fact, finding the linguistic features associated with perceived (as opposed to projected) personality is much closer to the general task of human-like text classification. We already have evidence that even when personality is projected linguistically, it is not always perceived accurately by human judges (Gill, Oberlander, & Austin, 2006). Hence, there are opportunities for machine learning methods to be applied to the personality classification task and for the machine classifications to be compared with those deriving from the self-reports of authors and those from third-party readers.

A final question is whether the linguistic effects of personality differences are all that important, compared for instance with those due to gender, education, or age. Our response to this is that other personality dimensions or other demographic dimensions are likely to prove at least as interesting as extraversion and neuroticism. We have not pursued these in this study, but recent work on a much larger corpus of naturally occurring personal weblogs suggests that for some linguistic features, the personality dimension of agreeableness may be more important than, say, extraversion—and that gender is also more important then extraversion (Nowson, 2006). The implication of this work is that an author's text will likely reflect several different aspects of his or her personal situation. An individual's character is only one aspect of the situation, and one's level of extraversion and neuroticism are only two aspects of his or her character.

CONCLUSION

The results we have uncovered confirm that there are linguistic differences in collocation patterns that can be systematically related to the differing characters of language users. We tried in general to link extraversion with positivity, sociability, complexity, and implicitness and neuroticism with negativity, self-concern, emphasis, and implicitness. The links were apparent in some, but not all, cases.

For extraversion we found high extravert collocations involving inclusive expressions and connectives more broadly, generating conjoined noun phrases, and collocations involving proper names. Low extraverts were notable for their tentativity, their greater tendency to refer to days of the week, and their less frequent use of adjectives. For neuroticism, we found no special bias concerning emotion collocations or first-person reference. As noted previously, this may be because e-mail is generally less emotional than speech. However, we did find a high neurotic preference for multiple punctuation, article collocations, inclusives, and conjunction. Low neurotics had distinctive collocations involving third-person pronouns and proper names and a broad-ranging use of adverbs. Whereas high neurotics characteristically used clause-initial *well*, low neurotics used *anyway.* We found no firm results concerning implicitness for either extraversion or neuroticism. As noted earlier, this may be because the e-mail task was not sufficiently stressful to elicit differences in implicitness.

These findings have been derived using bottom-up stratified corpus comparison, which is sufficiently sensitive to avoid some of the problems associated with dictionary-based methods. Yet interpreting the results is most easily carried out by reusing at least some of the dictionary categories. This is the most general conclusion of this article: The bottom-up, data-driven technique is effective, but interpretation benefits from the use of the existing top-down categories. The more specific results confirm that individual differences persist in the medium of the elicited e-mail discourse.

ACKNOWLEDGMENTS

Alastair J. Gill is now at LEAD-CNRS, Université de Bourgogne.

We are grateful to our anonymous referees and the editor for numerous suggestions that have greatly improved this article and to our colleagues for extensive discussions and advice. In particular, we would like to thank Elizabeth Austin, Carsten Brockmann, Zöe Bruce, James Curran, Jean-Marc Dewaele, Scott Nowson, Judy Robertson, and Keith Stenning. We are also obliged to other reviewers for and audience members at meetings where aspects of this work have been presented. The second author gratefully acknowledges support from two sources:

a studentship the United Kingdom Economic and Social Research Council and a Faber Post-Doctoral Fellowship (05 512 AA 06 S2469).

REFERENCES

Aarts, J., & Granger, S. (1998). Tag sequences in learner corpora: A key to interlanguage grammar and discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 132–141). New York: Addison Wesley Longman.

Ball, C. (1994). Automated text analysis: Cautionary tales. *Literary and Linguistic Computing, 9,* 295–302.

Bälter, O. (1998). *Electronic mail in a working context.* Unpublished doctoral dissertation, Royal Institute of Technology, Stockholm.

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the n-gram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 370–381). Berlin: Springer.

Baron, N. (1998). Letters by phone or speech by other means: The linguistics of e-mail. *Language and Communication, 18,* 133–170.

Biber, D. (1995). *Dimensions of register variation.* Cambridge, England: Cambridge University Press.

Bradac, J. (1990). Language attitudes and impression formation. In H. Giles & W. Robinson (Eds.), *Handbook of language and social psychology* (pp. 387–412). Chichester, England: Wiley.

Buckingham, R., Charles, M., & Beh, H. (2001). Extraversion and neuroticism, partially independent dimensions? *Personality and Individual Differences, 31,* 769–777.

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science, 14,* 60–65.

Carment, D. W., Miles, C. G., & Cervin, V. B. (1965). Persuasiveness and persuasibility as related to intelligence and extraversion. *British Journal of Social and Clinical Psychology, 4,* 1–7.

Cattell, R. B. (1946). *Description and measurement of personality.* London: Harrap.

Colley, A., & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology, 21,* 380–392.

Cope, C. (1969). Linguistic structure and personality development. *Journal of Counseling Psychology, 16,* 1–19.

Costa, P., & McCrae, R. (1984). Personality as a lifelong determinant of well-being. In C. Malatesta & C. Izard (Eds.), *Affective processes in adult development and aging* (pp. 141–157). Beverley Hills, CA: Sage.

Costa, P., & McCrae, R. R. (1992). *NEO–PI–R professional manual.* Odessa, FL: Psychological Assessment Resources.

Damerau, F. (1993). Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing and Management, 29,* 433–448.

Dewaele, J.-M. (2001). Interpreting the maxim of quantity: Interindividual and situational variation in discourse styles of non-native speakers. In E. Nèmeth (Ed.), *Selected papers from the 7th international pragmatics conference* (Vol. 1, pp. 85–99). Antwerp, Belgium: International Pragmatics Association.

Dewaele, J.-M. (2002a). Individual differences in L2 fluency: The effect of neurobiological correlates. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 219–250). Clevedon, England: Multilingual Matters.

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., & Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning, 49,* 509–544.

Dewaele, J.-M., & Furnham, A. (2000). Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences, 28,* 355–365.

Digman, J. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology, 41,* 417–440.

Eysenck, H. (1970). *The biological basis of personality.* Springfield, IL: Thomas.

Eysenck, H., & Eysenck, S. B. G. (1975). *The Eysenck Personality Questionnaire.* London: Hodder and Stoughton.

Eysenck, H., & Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire–Revised.* London: Hodder and Stoughton.

Eysenck, S., Eysenck, H., & 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 & W. Robinson (Eds.), *Handbook of language and social psychology* (pp. 73–95). Chichester, England: Wiley.

Gifford, R., & Hine, D. W. (1994). The role of verbal behaviour in the encoding and decoding of interpersonal dispositions. *Journal of Research in Personality, 28,* 115–132.

Gill, A. (2004). *Personality and language: The projection and perception of personality in computer-mediated communication.* Unpublished doctoral dissertation, University of Edinburgh, United Kingdom.

Gill, A., Harrison, A., & Oberlander, J. (2004). Interpersonality: Individual differences and interpersonal priming. In *Proceedings of the 26th annual conference of the cognitive science society* (pp. 464–469). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gill, A., Oberlander, J., & Austin, E. (2006). Rating e-mail personality at zero acquaintance. *Personality and Individual Differences, 40,* 497–507.

Goldberg, L. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34.

Granger, S., & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119–131). New York: Addison Wesley Longman.

Heylighen, F., & Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science, 7,* 293–340.

Howeler, M. (1972). Diversity of word usage as a stress indicator in an interview situation. *Journal of Psychological Research, 1,* 243–248.

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1994). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19,* 313–330.

Matthews, G., Deary, I., & Whiteman, M. (2003). *Personality traits* (2nd ed.). Cambridge, England: Cambridge University Press.

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer* (pp. 186–198). New York: Addison Wesley Longman.

Nowson, S. (2006). *The language of weblogs: A study of genre and individual differences.* Unpublished doctoral dissertation, University of Edinburgh, United Kingdom.

Nowson, S., Oberlander, J., & Gill, A. (2005). Weblogs, genres and individual differences. In *Proceedings of the 27th annual meeting of the Cognitive Science Society* (pp. 1666–1671). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Oberlander, J., & Gill, A. (2004). Individual differences and implicit language: Personality, parts-of-speech and pervasiveness. In *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 1035–1040). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pennebaker, J. W., & Francis, M. (1999). *Linguistic inquiry and word count (LIWC).* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC2001).* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pennebaker, J. W., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality & 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* (pp. 133–142). University of Pennsylvania.

Rayson, P. (2003). *Wmatrix: A statistical method and software tool for linguistic analysis through corpus comparison.* Unpublished doctoral dissertation, Lancaster University, England.

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

Scherer, K. (1979). Personality markers in speech. In K. R. Scherer & H. Giles (Eds.), *Social markers in speech* (pp. 147–209). Cambridge, England: Cambridge University Press.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford, England: Oxford University Press.

Smith, C. (1992). Introduction: Inferences from verbal material. In C. Smith (Ed.), *Motivation and personality: Handbook of thematic content analysis* (pp. 1–17). Cambridge, England: Cambridge University Press.

Stallman, R. (1994). *Gnu Emacs manual* (10th ed.). Boston: Free Software Foundation Press.

Thorne, A. (1987). The press of personality: A study of conversations between introverts and extraverts. *Journal of Personality & Social Psychology, 53,* 718–726.

Tribble, C. (2000). Genres, keywords, teaching: Towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75–90). Frankfurt, Germany: Peter Lang.

Whittaker, S. (2003). Theories and methods in mediated communication. In A. Graesser, M. Gernsbacher, & S. Goldman (Eds.), *The handbook of discourse processes* (pp. 243–286). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wiggins, J., & Pincus, A. (1992). Personality: Structure and assessment. *Annual Review of Psychology, 43,* 473–504.

Wilson, M. (1987). *MRC Psycholinguistic Database: Machine usable dictionary* (Tech. Rep.). Oxford, England: Oxford Text Archive.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20,* 1–28.