Title
Individual differences and implicit language: personality, parts-of-speech and pervasiveness

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26)

ISSN
1069-7977

Authors
Oberlander, Jon
Gill, Alastair J.

Publication Date
2004

Peer reviewed

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 $\geq$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.