Taking Care of the Linguistic Features of Extraversion

Alastair J. Gill (agill@cogsci.ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Jon Oberlander (J.Oberlander@ed.ac.uk)
Division of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW UK

Abstract

We study how Extraversion or Introversion influences people's language production. A corpus of e-mail texts was gathered from individuals categorised via Eysenck's EPQ-R personality test. One experiment analysed the corpus using existing content analysis tools, and found relatively weak effects of Extraversion. A second experiment used more sensitive bigram-based techniques from statistical natural language processing to replicate earlier findings, and uncover novel patterns of behaviour.

Introduction

Casual acquaintance with Extraverts[1] and Introverts suggests that the former talk a lot more than the latter. But apart from this intuitive difference, how does this personality dimension influence an individual's language production? Before addressing this question, we need to clarify what we mean by Extraversion, and its relevance to cognitive science.

A typical Extravert tends to be sociable, needs people to talk to, craves excitement, takes chances, is easy-going, and optimistic. By contrast, a typical Introvert is quiet, retiring, reserved, plans ahead, and dislikes excitement (Eysenck and Eysenck, 1991).

The personality trait of Extraversion—and the complementary Introversion—is one of the few which researchers generally agree provides 'consistent and valid information' (Jonassen and Grabowski, 1993). Beyond it, there is greater controversy.

For instance, Eysenck's EPQ-R personality test reflects a personality model which incorporates just two further dimensions: Neuroticism, which is mainly characterised by susceptibility to anxiety; and Psychoticism, which is more complicated, but generally related to aggression and individuality. By contrast, the NEO-PI-R model incorporates five factors (Costa and McCrae, 1992). As well as Extraversion and Neuroticism, there are Conscientiousness, Agreeableness and Openness. It is generally agreed that these relate to Psychoticism, but exactly how is

---

[1]The spelling of *Extravert* follows Eysenck, because this paper employs his EPQ-R as the measure of personality, but this does not represent a commitment to a specifically Eysenckian theory of personality.

still the subject of debate (cf. Matthews and Deary, 1998).

Extraversion, and its linguistic consequences—if there are any—is relevant to cognitive research for at least two reasons. First, there is considerable evidence that this personality dimension is related to preferred learning styles and educational achievement, via speed of exam completion, memory retrieval and recall tasks, creativity, mathematical ability, self monitoring and communication ability (Jonassen and Grabowski, 1993). Secondly, there is evidence that computer users attribute personality to interfaces, and respond to it in robust ways (eg. Nass, Moon, Fogg, and Reeves, 1995; Isbister and Nass, 2000). Even in a text-only environment, Extraverts apparently prefer interfaces presenting information using language associated with Extravert traits; Introverts prefer Introverted interfaces. An interface with a matching personality is judged more positively, and rated as more attractive, credible and informative (Nass *et al.*, 1995).

So the personality dimension has some validity, and appears relevant to the diagnosis and projection of personality in human-computer communication, and in computer-based learning. But how does Extraversion influence an individual's language production? In addressing this question, we first outline some hypotheses from the literature, before describing our collection of a controlled corpus of language, and our analysis of it. We then report the results—some unsurprising, others unexpected—and discuss some of their implications.

Previous hypotheses

Work on textual personality within the "Computers Are Social Actors" paradigm has taken the expressive hallmarks of Extraversion or *dominance* (one facet of the dimension) to be confidence, as shown by an avoidance of hedge-expressions such as *perhaps* and *maybe* (Nass *et al.*, 1995), and is related to the empirical work of Bradac and Mulac (1984) on perceptions of powerful and powerless speech.

From an intuitive perspective, Extraverts are described as individuals who think out loud, do most of the talking, are less self-focussed, and tend to skip from topic to topic. Conversely, Introverts mo-

nopolise the conversation on topics important to them, are more self-focussed and prefer to concentrate on discussing one topic in depth (cf. Carment, Miles, and Cervin, 1965). With reference primarily to speech, Furnham (1990) has proposed that Extravert language is less formal, has a more restricted code, uses more verbs, adverbs and pronouns (rather than nouns, adjectives, and prepositions), and uses vocabulary loosely (see also Dewaele and Furnham, 1999, for a review of speech and writing studies).

Text analysis approaches have found that transcribed texts rated as belonging to the *warm* facet of Extraversion used fewer negative emotion words and unique words, and more present tense verbs, with *dominant* texts using fewer unique words, positive emotion words and self referents (Berry, Pennebaker, Mueller, and Hiller, 1997). Finally, study of the texts *written* by Extraverts has found that they used fewer negations, tentative words, negative emotion words, causation words, inclusive words, and exclusive words, while using more social and positive emotion words (Pennebaker and King, 1999).

Data Collection

The approach to data collection follows Pennebaker and King (1999). Written texts were collected from 105 University students or recent graduates (37 males, 68 females; mean age = 24.3 years; SD = 4.6; all native English speakers). An introductory e-mail explained the experiment, and pointed subjects to the relevant web-page. After the completion of an online demographic questionnaire and a version of the Eysenck Personality Questionnaire (Revised short form; Eysenck, Eysenck, and Barrett, 1985) (mean score for E = 7.91, SD = 3.25; normative score = 7.42 (male), 7.60 (female)), subjects were asked to compose two e-mails *to a good friend whom they hadn't seen for quite some time*, the style of which is considered to be close to oral communication (Bälter, 1998). One message concerned their activities in the past week; the other discussed their plans for the next week. Subjects were advised to spend around ten minutes per message, composed online and submitted using an HTML form. It was highlighted that responses would be treated in confidence and that subjects could remain anonymous. No payment was made for participation, and integrity of responses was monitored by reading through the transcripts. One additional submission was rejected as not being serious; the resulting corpus contained 210 texts and 65,000 words.

Experiment 1: Dictionary techniques

LIWC and MRC Methods

LIWC Each respondent's texts were individually processed using the LIWC text analysis program (Pennebaker and Francis, 1999). Items were selected

Table 1: Summary of E Score and LIWC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Numbers	−.21		
	Word Count	.20	**.08**	**.0144**

for principle components analysis using the same criteria as Pennebaker and King (1999), namely reliability, topic independence, independence from other variables, and a mean minimum usage of 1%. The validity of the current data was shown using varimax rotation to derive four factors which essentially replicate their prior findings. There was minor variation in some factor loadings, which we attribute to differences in the writing tasks. See Gill and Oberlander (prep) for a fuller discussion.

By correlating their resultant LIWC factors with personality dimensions, Pennebaker and King's results suggest broad style preferences for Extraverts. But this does not identify the relative importance of their categories for identifying text as Extravert.

Thus, to identify which LIWC variables best help identify an author's personality, stepwise linear multiple regression was performed. The variables entered were those which showed at least a small correlation with the personality type—with a significance of $p < .1$—and which satisfied the criteria for inclusion in the previous principle components analysis. However, since requiring variables to have a mean usage of 1% per essay for inclusion in the analysis did not leave any LIWC variables in the regression equation for Extraversion, this criterion was ignored for the results presented below. (Interestingly, by contrast, even with the application of this criterion, Psychoticism and Neuroticism both had several strongly significant LIWC predictor variables.)

MRC In addition to the LIWC-based tests, multiple regression analysis was also performed on psycholinguistic properties of the texts, derived from the MRC Psycholinguistic Database (Coltheart, 1981). Texts were first tagged for Parts of Speech,[2] and each word-POS pair was then looked up in the database. If the word and POS tag matched a pair in the database, psycholinguistic data was returned for that word. When all the words in the text had been processed, mean scores were calculated for categories such as verbal frequency, written frequency, concreteness, age of acquisition, along with additional global information, such as the percentage of a text's words which were captured by the database. As with the LIWC regression, variables showing a correlation with the personality type with a significance of $p < .1$ were entered in to the equation.

---

[2]Using the MXPOST tagger (Ratnaparkhi, 1996).

Table 2: Summary of E Score and MRC multiple regression analysis.

Dependent Variable	Independent Variable	β	R^2	p
E Score	Mean Concreteness	$-.21$	**.05**	**.0278**

Results

The multiple regression analysis of the LIWC variables (Table 1) shows that a greater overall word count for a text ($\beta = .20$), and the occurrence of fewer references to numbers within that text ($\beta = -.21$), indicate Extraversion ($p < .05$). So, Extraverts *do* appear to type more than Introverts, mirroring earlier results on speech (Carment *et al.*, 1965), with the avoidance of numbers embodying a 'looser', less precise use of language (Furnham, 1990). However, the variance accounted for by these variables is relatively low at 8%. In comparable analyses, both Psychoticism and Neuroticism regression equations explain variance greater than 10%.

Similarly, with the MRC Psycholinguistic analysis (Table 2), only the novel finding of a general lowering of a text's concreteness of vocabulary ($\beta = -.21$, $p < .05$) was seen to explain 5% of variance in Extraversion. Again, equations for Psychoticism and Neuroticism explained more than 10% of variance.

Discussion

In both of the dictionary-based analyses of the texts, rather few features appeared to distinguish Extravert/Introvert texts, especially when compared to the numerous LIWC and MRC features which associated with Psychoticism and Neuroticism traits.

How could this be? At least two explanations are possible. First, the LIWC dictionary is a subjectively constructed analysis tool. It is based on judgements by health psychologists of texts written by distressed individuals for therapeutic purposes (Pennebaker and Francis, 1999). For its original purposes, this is a strength; but it also imposes a *top down* limitation on LIWC's functioning. Given this therapeutic origin, it is tempting to suggest that the linguistic features associated with the personality traits of Psychoticism and Neuroticism were more important or relevant to the distressed individuals producing the texts—and that is why these features are better represented in LIWC's dictionary.

The MRC database is also fitted to its specific purposes—for example, matching psycholinguistic stimuli—but this again imposes constraints which might prove artificial when it is applied to a different area of investigation.

Secondly, both dictionaries necessarily operate using strings corresponding to individual words, subsequently classifying them in a predefined way. Neither takes into account the context of a word. Thus it may be that for Psychoticism and Neuroticism the choice of word, or some property of the word is informative—but for Extraverts, it may be that word order or collocations are more relevant.

Experiment 2: NLP techniques

Therefore, we recruit *bottom up* statistical text analysis techniques from corpus linguistics. Specifically, bigram analysis calculates the probability of pairs of adjacent terms, or bigrams, occurring together in that order in a given text. To determine the significance of a bigram's occurrence, a statistic—log likelihood—is calculated, taking into account all the other instances of each element in the bigram pair, and the other words with which they appear.

Since bigrams can be used to calculate the probabilistic space in which language occurs, they have been put to a variety of uses (Collins, 1996; Pedersen, 2001). However, this study uses them simply as an advancement on the classified unigram (that is, single-word) analysis in Experiment 1. Because bigrams contain information about the interconnection and dependencies of words, this second analysis retains some of the contextual information of language use. Equally importantly, since bigrams are not classified subjectively, they provide a form of analysis that is bottom-up, rather than top-down.

Method

The original corpus of texts was divided by degree of Extraversion by selecting respondents whose E score was greater or less than 1 s.d. of the mean (cf. Dewaele and Pavlenko, 2002), with the 21 High Extravert authors scoring more than 11, and the 17 Low Extravert authors scoring less than 5.

Bigrams were calculated for the resulting Extravert and Introvert subcorpora; the former contained over 12,000 words; the latter around 8,000. Bigram profiles were generated for each corpus and their co-occurence significance in the current texts ranked by log-likelihood statistic ($-2\log\lambda$),[3] since for smaller corpora this approximates better to χ^2 than the X^2 statistic (Dunning, 1993). Rankings for each group are based on the top 50 bigrams with frequency of $N \geq 2$, and a significance of $p < .001$. Relative frequency ratios (Damerau, 1993) were then calculated for bigrams that were common to both the subcorpora, and a Spearman Rank correlation was also performed on these bigrams.

Results

Spearman Rank Correlation

The correlation coefficient score of .53 indicates that Extravert and Introvert use of the shared bigrams is significantly correlated at the $p < .005$ (one-tailed, N=28) level, and they are therefore not distinct.

---

[3]Ted Pederson's bigram software is available from: http://www.d.umn.edu/~tpederse/code.html.

Table 3: Shared Extravert and Introvert bigrams.

Bigram	Extr Cnt	Intr Cnt	Extr Ratio	Intr Ratio	Rel.F Ratio
looking forward	15	4	0.0011	0.0005	2.49
it was	46	22	0.0034	0.0025	1.39
next week	24	12	0.0018	0.0013	1.33
a bit	29	15	0.0022	0.0017	1.28
up with	19	10	0.0014	0.0011	1.26
! !	45	24	0.0033	0.0027	1.24
will be	24	13	0.0018	0.0015	1.22
i was	33	18	0.0025	0.0020	1.22
at the	27	16	0.0020	0.0018	1.12
to see	32	19	0.0024	0.0021	1.12
which is	15	9	0.0011	0.0010	1.11
for a	34	21	0.0025	0.0024	1.07
i have	44	29	0.0033	0.0032	1.01
to get	34	23	0.0025	0.0026	0.98
. i	99	69	0.0074	0.0077	0.95
on friday	11	8	0.0008	0.0009	0.91
, and	48	36	0.0036	0.0040	0.88
and then	23	19	0.0017	0.0021	0.80
in the	41	34	0.0031	0.0038	0.80
apart from	6	5	0.0005	0.0006	0.80
i am	33	28	0.0025	0.0031	0.78
i think	16	14	0.0012	0.0016	0.76
, but	35	31	0.0026	0.0035	0.75
a lot	10	9	0.0007	0.0010	0.74
going to	36	33	0.0027	0.0037	0.72
a few	12	11	0.0009	0.0012	0.72
to do	23	23	0.0017	0.0026	0.66
i've been	9	12	0.0007	0.0013	0.50

Table 4: Bigrams unique to Extravert corpus.

Bigram	Rank	$-2\log\lambda$	Count	Ratio
. .	8	183.48	152	0.0113
of the	33	79.47	40	0.0030
, which	20	100.89	25	0.0019
had a	16	115.60	22	0.0016
which was	24	95.69	19	0.0014
new year	7	192.22	18	0.0013
got a	45	66.65	17	0.0013
a good	46	64.45	16	0.0012
forward to	26	94.76	15	0.0011
need to	28	89.99	15	0.0011
i'll be	22	98.70	14	0.0010
on saturday	27	90.94	13	0.0010
we went	42	67.54	11	0.0008
as well	43	67.18	11	0.0008
couple of	30	84.18	10	0.0007
want to	41	68.01	10	0.0007
the moment	44	67.09	10	0.0007
<END> hi	21	99.44	9	0.0007
able to	50	61.19	9	0.0007
take care	23	96.00	8	0.0006
catch up	39	70.50	7	0.0005
other than	49	62.84	6	0.0005

followed by *hi*, was unique to Extravert texts; and since the <END> marker separates concatenated files in the corpus, here we have a tendency towards message-initial *hi*. By contrast the more formal [<END> *hello*] was found solely in Introvert texts. Use of punctuation also differs between the two groups, with Extraverts preferring multiple exclamation marks [*! !*], and solely using multiple full stops [. .] as in the elliptical (. . .), again a feature of informal style, and 'looser' use of language.

Quantification In terms of quantification, Introverts generally tend to show a preference for a greater use of quantifiers, such as [*a lot*], [*a few*] and uniquely [*all the*], [*one of*], [*lots of*] and [*loads of*], whereas Extraverts show a preference for [*a bit*] and uniquely use [*couple of*]. Not only does this demonstrate an Extravert tendency to be looser and less specific, it also apparently reveals exaggeration on the part of the Introvert.

Social Devices The Extravert use of stylistic expressions such as [*catch up*] and [*take care*] indicate a relaxed and informal style; their omission points to a more socially restrained Introvert. A surprisingly neat equivalence in expression can be found between the Extravert use of [*other than*] rather than [*apart from*], although it is not immediately clear what might give rise to this.

Self/Other Reference References to self in the texts demonstrate differences between Extraverts and Introverts: Introverts make extensive use of the first person singular pronoun ([*i don't*], [*i went*], [*i'm going*], [*i can*], [*i've got*] are all unique to the Introvert text), and also show preference for the following shared bigrams: [*i've done*], [*i think*], [*i am*], [*. i*].

However, further analysis showed Extraverts to be more distinguishable from Ambiverts or Introverts.[4]

Extraverts versus Introverts

The results of the bigram analysis include: bigrams which occurred in both the Extravert and Introvert corpora (Table 3); bigrams which were found uniquely in the Extravert corpus (Table 4); and those found only in the Introvert corpus (Table 5). The shared bigrams are ordered by their relative frequency, with the highest ratios above 1.0 showing the strongest association with Extravert authors, and the smallest ratios less than 1.0 indicating a preference on the part of more Introverted authors (the breakpoint has been indicated by a separating rule). Features which are unique to each subcorpus group can be considered the most distinctive of authorial personality. For current purposes, we divide the features into eight groupings.

Surface Realisation Features These gross features are perhaps the most intuitive in their representation of the Extraverts or Introverts. For example, [<END> *hi*], the <END> (end–of–file marker)

---

[4]When comparing the groups High E ($\geq$1s.d.), Mid E ($< \pm$1s.d.) and Low E ($\leq -$1s.d.) (all P and N $< \pm$1s.d.) it was found that Low E and Mid E correlate very significantly ($p < .005$; $\rho = .67$; $N = 19$), whilst High E and Mid E do not significantly correlate at the $p < .05$ level ($\rho = .32$; $N = 24$).

Table 5: Bigrams unique to Introvert corpus.

Bigram	Rank	$-2\log\lambda$	Count	Ratio
. <END>	17	80.13	20	0.0022
i don't	18	78.77	18	0.0020
went to	25	63.53	15	0.0017
to go	34	56.65	14	0.0016
all the	47	43.06	12	0.0013
i went	50	42.70	12	0.0013
one of	32	57.45	11	0.0012
trying to	29	60.75	10	0.0011
i'm going	36	52.84	10	0.0011
i can	46	43.90	10	0.0011
on thursday	20	72.22	9	0.0010
don't know	21	69.76	9	0.0010
i've got	35	55.19	9	0.0010
lots of	26	62.29	8	0.0009
this week	39	48.51	8	0.0009
anyway ,	45	44.79	8	0.0009
should be	40	48.10	7	0.0008
on monday	41	47.91	6	0.0007
two weeks	31	58.65	5	0.0006
loads of	49	42.72	5	0.0006
<END> hello	44	45.05	4	0.0005
exam results	42	47.26	3	0.0003

For Extraverts, the only unique first person bigram is [*i'll be*], and they also show greater use of [*i was*] and [*i will*], although relatively less preferred than Introvert forms. This underscores the increased Introvert tendency to focus on self, whereas the only bigram containing a first person plural is unique to Extraverts ([*we went*]). The Extravert preference for the bigram [*up with*] typically indicates a shared experience (prompting the question *with whom?*) and greater sociability. These results apparently contradict Furnham (1990) on pronouns, but given that the vast majority of pronouns here are first-person singular, thus focusing on self, this is unsurprising.

Valence Bigrams containing negations were used significantly only by Introverts, as in [*i don't*] and [*don't know*] (indeed [*i don't*] is the bigram with most frequent use of *i*) , whilst Extraverts used the bigram [*a good*] which is suggestive of positive affect.[5] Similarly, the Extravert preference for [*looking forward*] and [*forward to*] (presumably as in *looking forward to*) also suggests a more positive disposition.

Ability Personal views on capability are suggested by the different collocations with infinitival *to*.[6] For Extraverts, their ability to do something should they choose is confidently and assertively relayed using *want–, need–,* and *able– (to)*; which they use uniquely. Introverts more timidly and tentatively

---

[5]Further investigation shows that *good* is not directly negated (as in [*not good*]). Compare the Introvert [*i can*], which was generally followed by *not.* Although the effect of negation was not viewed as important by Pennebaker in the functioning of LIWC, it certainly has implications for models of language generation.

[6]This confirms the appropriacy of retaining functors usually filtered out by a stop list (cf. Damerau, 1993).

---

state that they are [*trying to*] or possibly—and at some point in the future—they are [*going to*].

Modality Similarly, collocations with the verb *be* show a distinction in use of modal auxiliaries which has an effect on the projection of certainty. For example, Introverts are unique in their use of the weaker and more tentative *should be*, whereas Extraverts show a greater use of the stronger predictive [*will be*], and are unique in their use of the contracted form [*i'll be*] (*i will be*) (Coates, 1983).

Message Planning/Expression Looking towards surrogates of grammatical construction, Extraverts and Introverts differ in their use of connectives: Introverts show preference for the coordinating conjunctions [*, and*] and [*, but*], whilst Extraverts uniquely show use of the subordinating [*, which*], usually deployed in an evaluative sense.

Discussion

In summary, our results support earlier findings, and suggest some new conclusions.

We found that Extraverts produce texts with more words, which supports the previous findings for speech (Carment *et al.*, 1965), whilst the reduced concreteness of Extravert language is a novel finding. It may be a direct consequence of talking or writing more, if the pressure to produce words at a high rate (in order to hold the floor, for instance) diverts resources away from more detailed lexical planning. Introverts' greater preference for numbers and quantification fits with this, and is compatible with findings concerning the use of articles (Pennebaker and King, 1999), and suggestions of a more imprecise and 'looser' Extravert style (Furnham, 1990).

Extraverts' use of other or social referents, and Introverts' preference for self referents confirms Berry *et al.* (1997)'s previous findings for Extraversion and its *dominant/submissive* facets. Another possible manifestation of the increased Extravert social ability and ease in interaction is expressed by their use of surface features and social devices. We also note in passing the tendency of Extraverts to refer to days of the weekend, where Introverts refer to weekdays.

Our results on valence are consistent with previous findings on Introverts' preference for negations and negative emotion words, and the Extravert tendency for positive affect words is consistent with results for *warmth*. However, they do suggest that care should be taken over the relation between Extraversion and *dominant* facet features (cf. Isbister and Nass, 2000).

Expressions of definite modality and ability appear to be associated with Extraversion, although they may not be the same forms as those discussed in the context of powerful/less speech. Adoption of definite modalities can also be related to avoidance of tentativity (Pennebaker and King, 1999).

Turning to connectives, we note that our Introvert

preference for [, *and*] and [, *but*] is consistent with studies using LIWC which found that the dictionary categories of Inclusion and Exclusion were both inversely correlate with Extraversion. However, [*other than*] and [*apart from*] would both fall into the same LIWC category, yet appear to distinguish opposite ends of the personality dimension.

Conclusion

By combining techniques from psycholinguistics and statistical natural language processing, we have been able to replicate previous findings on the expression of Extraversion through language, and uncover some new linguistic behaviours. Where existing content analysis tools could not detect reliable differences, more sensitive linguistic tools proved their worth.

Further, more technically sophisticated analyses can be carried out on this data, and we envisage the use of machine learning techniques to identify distinctive features from the texts, along with bigram analysis exploiting Parts of Speech tags. Additionally, the role of gender could be investigated.

Our findings could be exploited within the field of automatic language generation. As they stand, stochastic techniques would be needed; however, a cognitively-based personality model would allow a deeper approach, and that is our eventual goal.

Acknowledgements

References

Bälter, O. (1998). *Electronic Mail in a Working Context*. Ph.D. thesis, Royal Institute of Technology, Stockholm.

Berry, D., Pennebaker, J., Mueller, J., and Hiller, W. (1997). Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, **23**, 526–537.

Bradac, J. and Mulac, A. (1984). A molecular view of powerful and powerless speech styles. *Communication Monographs*, **51**, 307–319.

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

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. Croom Helm, London.

Collins, M. J. (1996). A new statistical parser based on bigram lexical dependencies. In *Proc of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184–191.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, **33**, 497–505.

Costa, P. and McCrae, R. R. (1992). *NEO PI-R Professional Manual*. Psychological Assessment Resources, Odessa, Florida.

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

Dewaele, J.-M. and Furnham, A. (1999). Extraversion: The unloved variable in applied linguistic research. *Language Learning*, **49**(3), 509–544.

Dewaele, J.-M. and Pavlenko, A. (2002). Emotion vocabulary in interlanguage. *Language Learning*, **52**(2), 265–324.

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Eysenck, H. and Eysenck, S. (1991). *Eysenck Personality Questionnaire-Revised*. Hodder, London.

Eysenck, S., Eysenck, H., and Barrett, P. (1985). A revised version of the psychoticism scale. *Personality and Individual Differences*, **6**(1), 21–29.

Furnham, A. (1990). Language and personality. In H. Giles and W. Robinson, editors, *Handbook of Language and Social Psychology*, pages 73–95. Wiley, Chichester.

Gill, A. and Oberlander, J. (in prep.). Dictionary approaches to personality language. *in prep.*

Isbister, K. and Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *Int. J Human-Computer Studies*, **53**, 251–267.

Jonassen, D. and Grabowski, B. (1993). *Handbook of Individual Differences, Learning and Instruction*. Laurence Erlbaum Associates, Hillsdale, NJ.

Matthews, G. and Deary, I. (1998). *Personality Traits*. Cambridge University Press, Cambridge.

Nass, C., Moon, Y., Fogg, B., and Reeves, B. (1995). Can computer personalities be human personalities? *Int J Human-Computer Studies*, **43**, 223–239.

Pedersen, T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*.

Pennebaker, W. and Francis, M. (1999). *Linguistic Inquiry and Word Count (LIWC)*. Lawrence Erlbaum Associates, Mahwah, NJ.

Pennebaker, W. and King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, **77**(6), 1296–1312.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.