Level of representation and semantic distance: Rating author personality from texts

Alastair J. Gill (A.Gill@ed.ac.uk)
LEAD-CNRS UMR 5022, University of Burgundy
Dijon 21000, France

Robert M. French (robert.french@u-bourgogne.fr)
LEAD-CNRS UMR 5022, University of Burgundy
Dijon 21000, France

Abstract

Increasingly our perception of others is based on short samples of written text, for example, in e-mail or chat rooms. In this paper we will examine the extent to which text co-occurrence techniques, such as LSA, HAL, and PMI-IR, can be successfully applied to human personality perception based on short written texts. In particular, we compare two approaches: The first compares a "surface similarity" judgment of the text being rated to a description by the author of the text of his/her personality (Simulation 1). The second relies on extracting a very simple representation of author personality from extreme texts and judging the experimental texts on the basis of this representation (Simulation 2). Both of these approaches fail to distinguish personality type. We conclude that co-occurrence techniques, at least used in a relatively canonical way to assess personality from small text samples, are not only inadequate but, most probably, are not doing this in a way that is similar to how we humans rate personality from short text samples.

Introduction

In daily life we may open up our e-mail inbox to discover a message from an unknown individual. We may read through the message and notice that the text's author mentions *parties*, *people*, and *socializing* very frequently. How do we then make a judgment about the author's personality on the basis of these few 'key terms' extracted from the text?

Personality traits are relatively stable over time and relate to an individual's "core qualities". Therefore, judging an individual's personality involves trying to predict future behaviour on the basis of their current or observed behaviour. In this paper we focus on the two traits central to personality theories, Extraversion and Neuroticism (Kline, 1991). Key adjectives that characterize these two traits were taken from Goldberg's five-factor model (FFM; Goldberg, 1992) and used to conceptualise personality in Simulation 1 (see Table 1).

Studies of personality perception show remarkable levels of consensus for these two traits (especially for Extraversion), even in text-only computer-mediated communication (CMC) environments, such as e-mail, chatrooms, and personal websites (Gill, Oberlander & Austin, 2006; Markey & Wells, 2002; Vazire & Gosling, 2004). Furthermore, both Extraversion and Neuroticism influence language at the level of both content and grammar (Oberlander & Gill, 2006; Pennebaker & King, 1999), a fact that has been successfully applied to the task of author personality classification from text (Argamon, Sushant, Koppel, & Pennebaker, 2005; Oberlander & Nowson, 2006).

Although there are models of human processes of personality judgment and perception (cf. Realistic Accuracy Model, Funder, 1995; Weighted-Average Model, Kenny, 1991), these models do not address how representations of personality types – such as those described in the Five-Factor Model (FFM; Goldberg, 1992) – are actually used to determine real world behavior.

In what follows we present two possible explanations of how this might be done. We then test these explanations using three well-known text co-occurrence programs (LSA, HAL, PMI-IR). The first possible explanation, explored in Simulation 1, is that people are simply doing a (largely unconscious) comparison of the overall semantic distance of a number of key terms in the written text directly to the words representing the personality concept: We refer to this as a "surface similarity" judgment. In this case, for example, we would make a rapid mental calculation of the overall semantic similarity between *parties, people, socializing* (words taken from the text under consideration) and *active, enthusiastic, talkative*, words that we know (cf. Goldberg, 1992) to be indicative of extraversion. In Simulation 2, we explore an arguably more realistic, stronger method. How do individual raters use abstract personality concept information (e.g., *active, enthusiastic*, *talkative*) to develop a higher-level representation of an extravert, from which they can then form a shared meaning system (Kenny, 1991; French, 1995). In text rating situations, such a meaning system may give rise to concepts like *parties, fun,* and *exciting* which would be expected to be in extravert writing. This "representative extravert text" would then be compared – in terms of its semantic similarity – to the key terms derived from the text written by the unknown author in order to determine the extent to which he/she seems to be an extravert. Note that the former strategy does not require the building of a higher-level structural representation of the personality of the text's author. Therefore, it would be computationally less intensive and, in a world of constant competition for cognitive resources, it would be the preferred assessment strategy, assuming it was sufficient for accurate personality ratings.

To explore the two means of evaluating a short written text in order to determine the personality of its author, we adopt statistical text co-occurrence measures of semantic space. These programs are able to compare texts in terms of their general meaning level which make them more suitable for the exploration of human behavior compared to traditional machine learning techniques which search for particular words or types (e.g., Argamon, Sushant, Koppel, & Pennebaker, 2005; Oberlander & Nowson, 2006). The driving idea behind co-occurrence programs, such as, HAL (Lund, Burgess & Atchley, 1995), LSA (Landauer and Dumais, 1997), and PMI-IR (Turney, 2002), is that they can determine the semantics (or, at least, some of the semantics) of a word by analyzing "the company it keeps" in a large corpus of text (Firth, 1957). In short, the average degree of physical proximity over a large number of texts of two words is a measure of their semantic proximity. The size of the *semantic neighbourhood* varies across the different approaches. For example in HAL, it is limited to a few words, whereas for LSA it is the entire document in which the word is found. Although the statistical methods employed to determine co-occurrence vary across the programs, they have demonstrated human-like ability and performance in tasks such as English language learner synonym tasks (e.g., Landauer & Dumais, 1997), classifying the semantic orientation (good vs bad, etc.) of individual words and movie reviews (Turney, 2002; Turney & Littman, 2003), analogical retrieval, (Ramscar & Yarlett, 2003), and even in visual fixations (Huettig et al, 2006; cf. Bullinaria & Levy, in press).

However, critics of co-occurrence techniques as models of human semantic processing argue that to have a truly human understanding of meaning requires human world knowledge and human experience (Glenberg & Robertson, 2000; French & Labiouse, 2002): To correctly judge semantic distances between words, for example, to know how good *John* is as the name of a child's mother, one needs world knowledge, in this case, that mothers are always female, and that John is a male name (French & Labiouse, 2002). Indeed, Bullinaria & Levy (in press) observe that "obviously, co-occurrence statistics *on their own* [original emphasis] will not be sufficient to build complete and reliable lexical representations".

In this paper, we examine the abilities HAL, LSA, and PMI-IR in measuring the semantic similarity between the language of texts actually written by Extravert authors, and words representing Extraverts (such as those used to describe Extraverts, e.g., *enthusiastic*, *talkative* in Simulation 1; or those derived from highly Extravert authors, e.g., *parties, fun, exciting*, in Simulation 2).

Simulation 1

Method

Procedure Here we infer high/low personality orientation for Extraversion and Neuroticism on the basis of direct semantic associations between words in the target texts and "personality trait words" considered to characterize these two traits, taken from Goldberg's Five-Factor Model. The personality orientation of a given word is calculated from the strength of associations with the set of high personality trait words (i.e., words that "define" the trait) minus the strength of its association with a set of low personality trait words (cf. Turney, 2002 and Turney and Litttman, 2003). The precise formula used for this calculation can be found in Turney (2002).

Calculation of Semantic Space

The following programs and parameters were used for the calculation of semantic association:

- **HAL** was implemented using the British National Corpus (BNC), using a rectangular window of 7 words and distance between vectors calculated using cosine, as reported in Huettig et al. (2006).[1]
- **LSA** (Landauer, & Dumais, 1997) uses the University of Colorado at Boulder website[2] using the default semantic space derived from the 'General Reading up to 1st year of college' TASA corpus, and the maximum number of factors available (300). The comparison type used was 'term to term'.
- **PMI-IR** uses the Waterloo MultiText System (WMTS) corpus of around 5×10^{10} English words (due to changes in the functioning of AltaVista; cf. Turney, 2002).[3]

Extraversion		Neuroticism	
High	**Low**	**High**	**Low**
talkative	silent	emotional	calm
bold	timid	nervous	relaxed
assertive	compliant	subjective	objective
spontan-eous	inhibited	worrying	placid
active	passive	volatile	peaceful
energetic	lethargic	insecure	independ-ent
enthusi-astic	apathetic	fearful	inhibited

Table 1: Matched pairs of words associated with high/low Extraversion or Neuroticism (from Goldberg, 1992). These were the words used in Simulation 1 to determine how well the personality traits they characterized were related to the key words taken from the experimental texts.

Derivation of Personality Trait Words

Goldberg's (1992) five-factor model of personality (FFM) provided adjectives to describe the high/low extremes of the Extraversion and Neuroticism personality traits used in Simulation 1 (cf. prose descriptions of EPQ-R; Eysenck &

[1]A local version of this software was made available by Scott McDonald; an online version is available at: http://www.cogsci.ed.ac.uk/~scottm/semantic_space_model.html.
[2]Available from: http://lsa.colorado.edu.
[3]The Perl scripts used for the calculation of PMI-IR were modified from original versions kindly supplied by Peter Turney who also arranged for access to WMTS. An alternative version using Google can be found at: http://www.d.umn.edu/~tpederse.

Eysenck, 1991). Duplicates and multi-word phrases were removed, as were any words that did not appear in the 100 million-word British National Corpus (BNC). Seven matched high-low pairs for Extraversion (e.g., talkative-silent) and Neuroticism (e.g., emotional-calm) were selected in order of their strength in rating the trait, as in Goldberg's original study (cf. Goldberg, 1992, p. 33, Table 2). These matched pairs can be found in Table 1.

Selection of Personality Texts

All experimental texts (a corpus of around 65,000 words) were collected as part of previous experimentation (Gill et al. 2006; Oberlander & Gill, 2006): This consisted of e-mail texts collected from 105 current or recently graduated university students each of whom completed the Eysenck Personality Questionnaire (Revised form, EPQ-R; Eysenck & Eysenck, 1991), thereby providing self-report information for Extraversion and Neuroticism. Thus, for each of the texts we have a self-report by its author of his/her degree of Extraversion and Neuroticism. We did not do co-occurrence analyses of all words in each text, but rather extracted the following key words from the texts:

- The 10 most frequent open-class words, since these represent contentful language at its most general level. These were selected following the removal of the 363 most commonly occurring closed-class words (e.g. prepositions, determiners, conjunctions, and pronouns);
- The 10 most frequent adjectives and;
- The 10 most frequent adverbs.

The adjectives and adverbs were extracted from the texts after automatic tagging for parts of speech (Oberlander & Gill, 2006). We chose these classes of words since they have been used previously for semantic orientation (cf. Turney, 2002).

Relating Semantic Space and Author Personality

HAL, LSA, and PMI-IR were used to derive distances of semantic association for each experimental text with the high/low personality description adjectives for Extraversion and Neuroticism. The semantic orientation value for each of the 105 experimental texts (in the form of top 10 Open-class words, top 10 adjective, and top 10 adverb groups) was then correlated with author self-ratings derived using the EPQ-R (Eysenck & Eysenck, 1991).

Results and Discussion

There was only one significant – but *inverse* – correlation between the ten most frequent open-class words, the ten most frequent adjectives and the ten most frequent adverbs taken from the sample texts and the personality-defining words (see Table 1) from Goldberg's Five-Factor Model (1992). This was the correlation identified by PMI-IR between 10 Adjectives extracted from texts and the high-low Neuroticism trait-defining words from Table 1 ($r=-.25$; $p<.05$). No other significant correlation was found by any of the programs.

The surprising result of this simulation is that the most frequently occurring words (open-class, adjectives and adverbs) taken from short texts written by authors who provided self-ratings of their personality are simply not in the proximal co-occurrence neighborhoods of the trait-defining words established by Goldberg (1992). The difficulty lies perhaps in the fact that Extraverts may not actually write texts which includes language fitting an abstract description of themselves and their behavior. Indeed, this is a particularly important consideration for traits such as Neuroticism, which are often characterized more by internal behavior, rather than outward, expressive behaviors towards others, including, in this case, any description of one's own Neuroticism.

In any event, these results show that personality-appraisal techniques relying on semantic similarities between the most frequently used words in a text and words providing an abstract characterization of a particular trait do not work. It therefore appears likely that human personality raters do not rely on cues from the most frequently used words, but rather know ahead of time the sorts of words to look for. In order to do this, he/she must already have at least a simple *model* (i.e., a more complex internal representation) of an extravert or a neurotic person. This intuition is the basis of the second simulation.

Simulation 2

Method

Procedure In Simulation 1, we have shown that a direct co-occurrence comparison of personality-defining concepts (see Table 1) with a pre-selected set of text words does not seem to be enough to enable accurate personality judgments from short textual data. We have proposed that a human judge may use personality information at a conceptual level to create a simple representation of an imagined author of such a short text message and derive personality conclusions based on that. For example, by inferring that an Extraverted individual may write texts that talk about *parties*, *people*, and *socializing*, the judge would then be able to assess how closely the text in front of him or her matched such a schema. In Simulation 2, we consider one simple means of developing a "high-level" representation of personality, and, once again, examine the results using standard co-occurrence programs as in Simulation 1.

Calculation of Semantic Space

The same programs (HAL, LSA, and PMI-IR) with parameters as in Simulation 1 were used for the calculation of semantic space.

Derivation of Personality Trait Words The authors of the short texts used in this study rated themselves in terms of Extraversion and Neuroticism facets of their overall personality. We were therefore able to identify authors in our e-mail corpus who scored greater than 1 standard deviation from the mean for the personality traits under investigation (cf. Oberlander & Gill, 2006). This gave us four groups of individuals: High Extraversion, Low Extraversion, High Neuroticism and Low Neuroticism. These four groups contained the e-mails texts of 11, 4, 6, 9

	Top 7 Open-class words extracted from e-mail texts		Top 7 adjectives extracted from e-mail texts		Top 7 adverbs extracted from e-mail texts	
Trait	**Extraversion**	**Neuroticism**	**Extraversion**	**Neuroticism**	**Extraversion**	**Neuroticism**
High	back	people	other	many	hopefully	even
	nice	going	long	local	as	though
	also	film	cool	big	still	better
	Christmas	write	more	awful	anyway	away
	too	try	great	total	out	actually
	come	home	big	short	however	out
	long	want	busy	positive	better	only
Low	play	know	second	sure	much	still
	much	day	same	same	fairly	rather
	first	come	hard	least	especially	here
	actually	plan	funny	flat	down	ever
	know	year	fresh	usual	recently	down
	give	too	least	long	sure	forward
	down	new	full	exciting	often	long

Table 2: Simulation 2 exemplars derived from High/Low e-mail texts

authors respectively. These texts were then concatenated so as to form one large text for each of the four groups. After removing the most frequent closed-class words, we then selected the seven most frequent Open-class words, Adjectives and Adverbs for each of the four personality groups as in Simulation 1. These empirically derived personality-trait words can be found in Table 2.

Selection of Personality Texts

The same experimental texts were used as in Simulation 1. However, after excluding the texts of the 15 most extreme High Neurotic, Low Neurotic, High Extravert, Low Extravert authors, whose texts were used to derive the "personality representation" words, we used the remaining 90 texts for co-occurrence testing (rather than all of the original 105 texts).

Relating Semantic Space and Author Personality

As in Simulation 1 we attempt to infer high/low personality orientation for Extraversion and Neuroticism on the basis of semantic associations, again using HAL, LSA, and PMI-IR. The only difference is that that personality-trait words were not derived, as they were in Simulation 1, from an abstract model of personality discrimination -- in this case, the Goldberg's Five-Factor Model (Goldberg, 1992) -- but rather were derived directly from the texts written by authors whose personality self-ratings placed them at the extremes of the High-Low continuum for Neuroticism and Extraversion. We considered that these latter sets of words, derived directly from participants' texts, constituted a simple "representation" of the written texts by the strongest representatives of each of the four classes, namely, High Extraversion, Low Extraversion, High Neuroticism, and Low Neuroticism. In short, we felt that this technique would provide an even better chance for LSA, HAL, and PMI-IR to succeed in correctly classifying the remaining 90 texts correctly as to the personalities of their authors.

In all other respects this simulation was identical in methodology to Simulation 1.

Results and Discussion

Quite surprisingly, the results of the co-occurrence analyses using exemplars derived from high/low Extravert and Neurotic authors (Table 3) once again showed that all correlations were less than 0.20 and none of them were significant at the $p<.05$ level. The results from the co-occurrence analyses are not even close to those to the personality perception abilites of human judges for the same material, for example, Gill et al. (2006) found that target-judges agreement of $r=0.89$ ($p<0.05$) for ratings of Extraversion using the e-mail texts of the present study. Markey & Wells (2002) found agreement in ratings of Extraversion following one-on-one CMC chat of $r=.32$ ($p<0.05$), with other forms of CMC, such as personal websites (Vazire & Gosling, 2004) giving self-observer agreement of $r=.26$ and $r=.21$ for Extraversion and Neuroticism (both $p<0.05$), although we note that these were not necessarily text-only. We discuss possible reasons for this disparity in the General Discussion, below.

General Discussion

Humans are able to form, to a significant level of agreement, impressions of each other via short written texts, such as e-mail or chat rooms. However little is known about this process. In this paper we proposed two possible methods of such judgment processes, and implemented them using three widely-used co-occurrence techniques.

First, we explore the possibility, and note the benefits of, a 'fast and frugal' method of personality text classification, which simply assesses, at a surface level, 'how Extraverted' the words in the texts appear (cf. Friedrich, 1993; French, 1995). We simulate this model by calculating co-occurrence associations between our experimental texts and personality trait adjectives taken from a standard model of personality-trait judgment (Goldberg, 1992) and which describe high/low Extraversion and high/low Neuroticism (Simulation 1).

Second, in contrast to this first, simple comparative technique where personality trait words are directly compared to the experimental texts, we propose a more structural approach in which a set of words is derived from the texts written by authors with the (self-evaluated) strongest personalities along the two trait dimensions. We reasoned that these texts would be the best representatives of texts written by authors from the four categories of personality traits. We extracted the key words from these texts and then used these words to judge the remaining texts. Even under these conditions, we still observe no useful correlations obtained by any of the three co-occurrence programs that would allow them to be able to reliably extract a personality judgment from any given text sample.

These results are somewhat surprising in light of the clear success of co-occurrence programs in areas such as synonym matching and assessing opinions from text (Landauer & Dumais, 1997; Lund et al. 1995; Turney, 2002; Turney & Littman, 2003). In our view, this argues for the intrinsic difficulty of the task of personality perception. In other words, it is reasonable to assume that, had there been *some* significant co-occurrence correlations between either standard personality-trait words and words used in short texts (Simulation 1) or frequently used words derived from texts that are arguably representative of texts written by authors at the extremes of the personality traits under consideration (Simulation 2), LSA, HAL, or PMI-IR would have noticed these correlations. However, this was not the case and one must assume, therefore, that these correlations do not exist, at least, for the sets of words that we chose to characterize the texts and the personality traits. This was, presumably, why the performance of all three co-occurrences programs was not even close to human-like performance on this task.

Humans, it turns out, are able to accurately rate short written texts in terms of the personality of their authors (e.g., Gill, et al. 2006; Markey & Wells, 2002). So, why, when given the input, at first blush, reasonable, described in our simulations, do co-occurrence programs fail so completely on this task? The answer almost certainly lies with the selection of the data. We have characterized texts by looking at the most frequently used closed-class words, adjectives and adverbs. This is necessary to provide concise textual representations for the computationally intensive semantic space calculations by the co-occurrence programs. Human raters, in contrast, have access to the full texts (cf. Gill, et al. 2006) and can, therefore, build a far richer representation, aided by years of experience and world knowledge about how various kinds of people write, of the text's author. We have characterized the various personalities by a set of words taken from Golberg's FFM personality judgment model or from texts written by authors at the ends of each personality-trait continuum. Clearly, both the representations of our sample texts on the basis of their top-ten closed-class words, adjectives, and adverbs and these personality-trait representations are insufficient. It is not that LSA, HAL, or PMI-IR are "not working correctly". Presumably, they are working fine. They simply are not doing what humans are doing in performing this task because they lack the extra-text information that people have at their disposal and that has been gathered from years of experience with correlating people behavior with their writing styles. We would argue that humans, on the basis of several words in the text, can build a rich and complex representation – unlike the skeletal representations that we were able to derive from simple analysis of "paradigmatic" texts – of the potential author of the short text message they are considering. This representation is grounded in a lifetime of experience judging people based on what they say and write and, therefore, people are able to intuit far more accurately the personality of the author of the text they are reading than a co-occurrence program that is given only scant textual information on which to make its judgment.

These results would appear to suggest that, in order to do reliable extraction of personality-trait information from short texts, it is necessary to have a representation-building capacity that is, at present, beyond the reach of standard co-occurrence programs.

Conclusion

In this paper we have explored the possibility of using well-known co-occurrence programs (LSA, HAL, and PMI-IR) to perform personality judgments based on short, written texts. We first examined an approach based on a surface similarity judgment of the text being compared to words taken from a standard, abstract personality-trait model (Goldberg, 1992). None of the co-occurrence programs using this approach were able to reliably extract any personality-trait information from the texts. Next we attempted to use words taken from paradigmatic examples of texts written by authors who identified themselves as high/low Extravert or Neurotic via the EPQ-R self-rating personality questionnaire. Here again, all co-occurrence programs failed to find any reliable correlation between key words taken from these paradigmatic texts and a sample of 90 short written texts, for each of which we had a personality self-rating by the text's author.

This leads us to the conclusion that, at the present time, co-occurrence programs are not appropriate tools for this kind of evaluation. We suggest, however, that this is because they cannot presently develop the high-level representations of personality that we humans can. Once they begin to acquire this ability, we believe their ability to judge personality from short texts will gradually come in line with that of humans. Finally, this work suggests the somewhat unsuspected difficulty of automatic personality assessment. It would seem that humans, in order to perform this task, rely on information garnered from years of experience correlating people's behavior with their writing styles and this information is simply not present in an analysis of even a relatively large sample of texts (over a hundred).

Acknowledgments

This work was supported in part by a Région de Bourgogne FABER Post-Doctoral Fellowship (05512AA06S2469) and by European Commission Sixth Framework grant NEST 516542 (second author). Thanks to Jon Oberlander, Matthias Mehl, Jim Friedrich and our anonymous reviewers for their helpful comments on earlier versions of this paper.

We gratefully acknowledge the generosity of Scott McDonald and Peter Turney for making their software available and for their assistance in responding to questions.

References

Agamon, S., Sushant, D., Koppel, M. & Pennebaker, J. (2005). Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and Classification Society of North America.*

Bullinaria, J.A. & Levy, J.P. (in press). Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study. *Behavior Research Methods*, 38.

Eysenck, H. & Eysenck, S. B. G. (1991). *The Eysenck Personality Questionnaire-Revised.* Hodder & Stoughton, Sevenoaks.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society; reprinted in Palmer, F., (ed. 1968), Selected Papers of J. R. Firth, Longman, Harlow.

French, R.M. (1995). *The Subtlety of Sameness.* Cambridge: MIT Press.

French, R. M. & Labiouse, C. (2002). Four Problems with Extracting Human Semantics from Large Text Corpora. *Proceedings of the 24th Annual Conference of the Cognitive Science Society.* NJ:LEA.

Friedrich, J. (1993), Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena, *Psychological Review*, 100 (2), 298-319.

Funder, D.C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102, 652-670.

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

Goldberg, L. (1992). Development of markers for the Big-Five factor structure. *Psychological Assessment*, 4, 26-42.

Huettig, F., Quinlan, P. T., McDonald, S. A. & Altmann, G. T. M. (2006). Models of high-dimensional semantic space predict language-mediated eye movements in the visual world. *Acta Psychologica*, 121, 65-80.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 104(2):211–240.

Lund, K., Burgess, C., & Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society,* pages 660-665.

Markey, P. and Wells, S. (2002). Interpersonal perception in internet chat rooms. *Journal of Research in Personality*, 36:134–146.

Oberlander, J. & Gill, A.J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 42, 239–270.

Oberlander, J. and Nowson, S. (2006). Whose thumb is it anyway? Classifying author personality from weblog text. In *Proceedings 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia.*

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

Ramscar M.J.A. and Yarlett D.G. (2003). Semantic Grounding in Analogical Processing: An Environmental Approach, *Cognitive Science*, 27, 41-71.

Turney, P.D. (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, pp. 417-424.

Turney, P.D., & Littman, M.L. (2003). Measuring praise and criticism: Inference of semantic orientation from association, *ACM Transactions on Information Systems (TOIS)*, 21(4), 315-346.

Vazire, S., & Gosling, S. D. (2004). E-perceptions: Personality impressions based on personal web sites. *Journal of Personality and Social Psychology, 87*, 123–132.