Personae: a corpus for author and personality prediction from text

Kim Luyckx, Walter Daelemans

CNTS Language Technology Group
University of Antwerp
Prinsstraat 13, 2000 Antwerp, Belgium
{kim.luyckx, walter.daelemans}@ua.ac.be

Abstract

We present a new corpus for computational stylometry, more specifically authorship attribution and the prediction of author personality from text. Because of the large number of authors (145), the corpus will allow previously impossible studies of variation in features considered predictive for writing style. The innovative meta-information (personality profiles of the authors) associated with these texts allows the study of personality prediction, a not yet very well researched aspect of style. In this paper, we describe the contents of the corpus and show its use in both authorship attribution and personality prediction. We focus on features that have been proven useful in the field of author recognition. Syntactic features like part-of-speech *n*-grams are generally accepted as not being under the author's conscious control and therefore providing good clues for predicting gender or authorship. We want to test whether these features are helpful for personality prediction and authorship attribution on a large set of authors.

Both tasks are approached as text categorization tasks. First a document representation is constructed based on feature selection from the linguistically analyzed corpus (using the Memory-Based Shallow Parser (MBSP)). These are associated with each of the 145 authors or each of the four components of the Myers-Briggs Type Indicator (Introverted-Extraverted, Sensing-iNtuitive, Thinking-Feeling, Judging-Perceiving). Authorship attribution on 145 authors achieves results around 50% accuracy. Preliminary results indicate that the first two personality dimensions can be predicted fairly accurately.

1. Introduction

The style in which a text is written reflects an array of meta-information concerning the text (e.g. topic, register and genre) and its author (e.g. gender, region, age and personality). The field of computational stylometry addresses these aspects of style. We approach stylometry as an automatic text categorization task that labels documents according to a set of predefined categories (Sebastiani, 2002). Like most text categorization systems, it takes a two-stage approach which (i) achieves automatic selection of features that have high predictive value for the categories to be learned, and (ii) uses machine learning algorithms to learn to categorize new documents by using the features selected in the first stage. To allow the selection of linguistic features rather than (*n*-grams of) terms, robust and accurate text analysis tools such as lemmatizers, part of speech taggers, chunkers etc., are necessary. Recently, language technology has progressed to a state of the art in which this has become feasible. This enables the systematic study of the variation of these linguistic properties in texts by different authors (Baayen et al., 1996), time periods, genres or registers (Argamon et al., 2003a), regiolects, and even genders (Argamon et al., 2003a).

In this paper, we focus on personality prediction and authorship attribution. A lot of the research in authorship attribution is performed on a small set of authors, which is an artificial situation. Trying to classify an unseen text as being written by one of two or a few authors is a relatively simple task, which in most cases can be solved with high reliability and accuracies over 95%. Hardly any corpora - except for some based on blogs (Koppel et al., 2006) or Usenet newsgroups (Argamon et al., 2003b) - have more than ten candidate authors. Forensic experts typically use stylometry to indicate which of a small number of suspects is the most likely to have written a short text, without being able to rule out the fact that there might be other people in play. Many studies in stylometry overestimate the importance of linguistic features in experiments discriminating between only two or a small number of authors. We developed the *Personae* corpus to investigate this phenomenon. Documents written by 145 authors allow us to investigate the performance of authorship attribution on a large set of authors.

The corpus is also used for the prediction of the author's personality, a not yet very well researched aspect of style. Our aim is to test whether personality traits such as extraversion are reflected in writing style. Studies in language psychology show that there is a direct correlation between personality and language: personality is projected linguistically and can be perceived through language (Gill, 2003; Gill and Oberlander, 2002; Campbell and Pennebaker, 2003). These studies are however not in a prediction context, but in a descriptive statistics context. The main focus is on extraversion and neuroticism, two of "the most salient and visible personality traits" (Gill, 2003). We want to take the study of personality in text further in three ways:

i. By collecting texts on a non-personality related topic, in this case artificial life. A number of studies in personality prediction and language psychology rely on stream-of-consciousness essays or deep self-analysis (Argamon et al., 2005; Mairesse et al., 2007), and even texts about traumatic experiences (Campbell and Pennebaker, 2003). We also did no extensive cleaning-up of the corpus, in contrast to Nowson and Oberlander (2007).

ii. By collecting a corpus of Dutch written language for the prediction of personality, while other studies focus on English. Nevertheless, we believe our techniques to be transferable to other languages.

iii. By testing whether we can automatically predict personality based on writing. Only a few studies we know of combine the fields of stylometry and language psychology in a prediction context (Argamon et al., 2005; Nowson and Oberlander, 2007; Mairesse et al., 2007).

iv. By extending the task to eight discrimination tasks and four binary classification tasks. Each of the four components of the Myers-Briggs Type Indicator is studied: Introverted-Extraverted, Intuitive-Sensing, Thinking-Feeling, and Judging-Perceiving.

2. Background

2.1. Authorship attribution

The central question in authorship attribution is *Which of the candidate authors wrote a particular document?* Researchers in this field assume that all authors have specific style characteristics that are outside their conscious control. On the basis of those linguistic patterns and markers (e.g., part-of-speech tags), the author of a document can be identified.

Most traditional studies use small sets of authors. Frequencies of rewrite rules (Baayen et al., 1996), *n*-grams of syntactic labels from partial parsing (Hirst and Feiguina, 2007), *n*-grams of parts-of-speech (Diederich et al., 2000), function words (Miranda García and Calle Martín, 2007), and functional lexical features (Argamon et al., 2007) have been shown to be reliable markers of style. New metrics have been proposed for calculating the distance between authors, like the Delta measure (Burrows, 2002) and a measure for intertextual distance suggested by Labbé and Labbé (2006).

The field of authorship attribution is however dominated by studies overestimating the importance of these predictive features in experiments discriminating between only two or a few authors. Taking into account a larger set of authors allows the computation of the degree of variability encountered in text on a single topic of different (types of) features. Recently, research has started to focus on authorship identification on larger sets of authors: 8 (Van Halteren, 2005), 20 (Argamon et al., 2003b), 114 (Madigan et al., 2005), or up to thousands of authors (Koppel et al., 2006) (see Section 5.1).

2.2. Personality prediction

Most of the research in personality prediction involves the Five-Factor Model of Personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These so-called *Big Five* have been criticized for their limited scope, methodology and the absence of an underlying theory. Argamon et al. (2005) predict personality in student essays using functional lexical features. These features represent lexical and structural choices made in the text. The corpus was specifically built for personality prediction. Each student was asked to write a stream-of-consciousness essay and an essay of deep self-analysis. A downside of this approach is, that the students are aware their personality is under investigation, which may influence their writing style.

Nowson and Oberlander (2007) perform feature selection and training on a small and clean weblog corpus, and test on a large, automatically selected corpus. Features include n-grams of words with predictive strength for the binary classification tasks. Openness is excluded from the experiments because of the skewed class distribution. Their study depends on a training corpus with accurate personality scores and clean text, which is hardly compatible with realistic situations. To mimic a realistic testing situation, they do considerably less cleaning up in the larger test corpus.

While the two studies mentioned above took a bottom-up approach, Mairesse et al. (2007) approach personality prediction from a top-down perspective. On a written text corpus, they test the predictive strength of linguistic features that have been proposed in descriptive statistics studies. Similar to Argamon et al. (2005), students - here even *psychology* students - were asked to write stream-of-consciousness essays.

3. Corpus

Our 200,000-word *Personae* corpus consists of 145 student (BA level) essays of about 1400 words about a documentary on Artificial Life. We chose a single topic in order to keep genre, register, topic and age relatively constant. Choosing a non-personality related topic minimizes the effect of awareness of personality being under investigation. The essays contain a factual description of the documentary and the students' opinion about it. The task was voluntary and students producing an essay were rewarded with two cinema tickets. The students also took an online MBTI test and submitted their profile, the text and some user information via a website. All students released the copyright of their text to the University of Antwerp and explicitly allowed the use of their text and associated MBTI personality profile for research, which makes it possible to distribute the corpus.

The Myers-Briggs Type Indicator (MBTI) (Briggs Myers and Myers, 1980) is a forced-choice test based on Carl Jung's personality typology (Jung, 1921) and designed to categorize a person according to four preferences:

- **I**ntroversion and **E**xtraversion (attitudes): Typically, I's tend to reflect before they act, while E's act before they reflect.

- i**N**tuition and **S**ensing (information-gathering functions): N's trust more abstract or theoretical information, while S's trust information that is concrete.

- **F**eeling and **T**hinking (decision-making functions): While F's decide based on emotions, T's tend to involve logic and reason in their decisions.

- **J**udging and **P**erceiving (lifestyle): J's prefer structure in their lives, while P's like change.

MBTI correlates with the *Big Five* (i.e. the Five-Factor Model of Personality) personality characteristics of extraversion and openness, to a lesser extent with agreeable-

	I	E	N	S	F	T	J	P
% docs	.45	.55	.54	.46	.72	.28	.81	.19
avg. nr. of words	1409	1416	1430	1392	1422	1386	1423	1369
avg. nr. of syll/word	1.48	1.45	1.47	1.46	1.46	1.48	1.46	1.46
avg. sentence length	19.87	18.66	19.66	18.64	18.92	19.93	19.03	19.89

Table 1: Corpus structure

ness and consciousness, but not with neuroticism (McCrae and Costa, 1989).

The participants' characteristics are too homogeneous for experiments concerning gender (77% female), mother tongue (97% native speaker of Flemish-Dutch) or region (77% from the Antwerp region), but we find interesting distributions in at least two of the four MBTI preferences (see Table 1).

Personality measurement in general, and the MBTI is no exception, is a controversial domain. However, especially for scores on IE and NS dimensions, consensus seems to be that they are indeed correlated with personality traits. In the remainder of this paper, we will provide preliminary results on the prediction of personality types from features extracted from the linguistically analyzed essays.

4. Methodology

Text classification starts from a set of training documents (documents of which the author/personality type is known), automatically extracts features that are informative for the class to be predicted and trains a machine learning algorithm that optimally uses these features to do the task for new, previously unseen, documents (Sebastiani, 2002).

4.1. Feature extraction

Syntactic features have been proposed as more reliable style markers than for example token-level features since they are not under the conscious control of the author (Baayen et al., 1996; Argamon et al., 2007). We use the Memory-Based Shallow Parser (MBSP) (Daelemans et al., 1999), which gives an incomplete parse of the input text, to extract reliable syntactic features. MBSP tokenizes the input, performs a part-of-speech analysis, looks for noun phrase, verb phrase and other phrase chunks and detects subject and object of the sentence and a number of other grammatical relations.

Words or parts-of-speech (*n*-grams) occurring more often than expected in either of the categories are extracted automatically for every document. We use the χ^2 metric (see Equation 1), which calculates the expected and observed frequency for every item in every category, to spot features that are able to discriminate between the categories under investigation, i.e. introverted and extraverted authors.

$$\chi^2 = \sum_{i=1}^{k} \frac{(\chi_i - \mu_i)^2}{\sigma_i} \qquad (1)$$

Lexical features (*lex*) are represented binary or numerically, in *n*-grams. *N*-grams of both fine-grained (*pos*) and coarse-grained parts-of-speech (*cgp*) are integrated in the feature vectors as well. The most predictive function words are present in the *fwd* feature set. For all of these features, the χ^2 value is calculated.

An implementation of the Flesch-Kincaid metric indicating the readability of a text along with its components (viz., mean word and sentence length) and the type-token ratio (which indicates vocabulary richness) are also represented (*tok*). These features have been proven useful in the field of stylometry (Stamatatos et al., 2001; Luyckx and Daelemans, 2005; Luyckx et al., 2006) and are now tested for personality prediction.

4.2. Experimental set-up

For authorship attribution on 145 authors, training is done by means of 5-fold cross-validation. *K*-fold cross-validation allows us to get a reliable indication of how well the learner will do when it is asked to make new predictions on the held-out test set. It also allows us to experiment with different algorithm parameter settings without using the test data. The data set is divided into five subsets containing two fragments of equal size per author. Five times one of the subsets is used as test set and the other subsets as training set.

For personality prediction, training is done by means of 10-fold cross-validation. The data set is divided into nine subsets of 15 authors and one of 10 authors. Ten times one of the subsets is used as test set and the other subsets as training set.

The feature vectors that are fed into the machine learning algorithm contain the top-*n* features with highest χ^2 value. For the experiments in personality prediction, every author is represented by one feature vector, resulting in 145 vectors per fold (divided over training and test). For the authorship attribution experiments, on the other hand, every text fragment is split in ten equal parts, each part being represented by means of a feature vector, resulting in 1450 vectors per fold (divided over training and test).

For classification, we use TiMBL (Memory-based learning) (Daelemans et al., 1999), a supervised inductive algorithm for learning classification tasks based on the *k-nn* algorithm with various extensions for dealing with nominal features and feature relevance weighting. Memory-based learning stores feature representations of training instances in memory without abstraction and classifies new instances by matching their feature representation to all instances in memory. From these ”nearest neighbors”, the class of the test item is extrapolated. We did no extensive model selection (optimization) of the parameters for TiMBL yet, since these are exploratory experiments.

5. Results and discussion

5.1. Authorship attribution

Tables 2 and 3 show classification accuracy using the five-fold cross-validation mechanism as explained in Section 4.2. The micro-average of all correctly classified instances over the five folds was calculated. We present results for all feature sets so that we can discuss the most relevant type of features for authorship attribution on the 145 authors in the *Personae* corpus. The random baseline generates the positive class with a probability estimated on the frequency of that class in training. The majority baseline, on the other hand, always generates the class that has the majority in training - which is one of the 145 authors in this study. Random baseline for authorship attribution on 145 authors (each of them with 2 fragments in test) is 0.00% accuracy, and the majority baseline is 0.69% accuracy.

Features	Accuracy
tok	29.17%
fwd	**32.83%**
lex 1	**34.00%**
lex 2	22.90%
lex 3	12.00%
cgp 1	30.00%
cgp 2	31.17%
cgp 3	28.21%
pos 1	**34.48%**
pos 2	30.55%
pos 3	17.10%

Table 2: TiMBL results in authorship attribution on 145 authors: separate feature sets

The results in Table 2 suggest that words (content and content words) and parts-of-speech are the most reliable markers of style in this large set of authors. Incrementally combining good working singular feature sets achieves results between 40.69% and 49.21% accuracy (see Table 3), which is a vast improvement over the majority baseline. This indicates that providing (combinations of) deeper linguistic features rather than only lexical of token features improves the system significantly.

Features	Accuracy
lex1 + pos1	40.69%
lex1 + pos1 + tok	48.28%
lex1 + pos1 + tok + fwd1	48.28%
lex1 + tok	**49.21%**

Table 3: TiMBL results in authorship attribution on 145 authors: combinations

For 145 authors, an accuracy around 50% agrees with our expectations, since increasing the number of authors to be discerned makes the task considerably more difficult. Most other studies in authorship attribution report accuracies of more than 95%. Trying to classify an unseen text as being written by one of two or a few authors is a relatively simple task. We tried to simulate an experiment with two authors, using the approach and features we present in this paper. In order to minimize the effect of chance in selecting two authors from the *Personae* corpus, we selected a hundred random samples of two authors (see Table 1). On average, we achieve an accuracy of 96.90% with separate feature sets (viz., *cgp2*), and even 98.65% with combinations of feature sets (viz., *lex1+pos1+tok*), which is in line with results reported in other studies on small sets of authors.

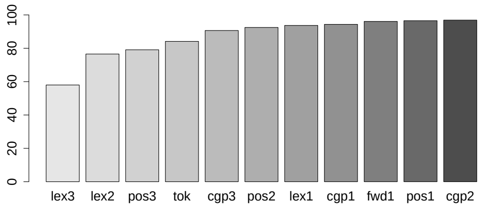

Figure 1: TiMBL results in authorship attribution on 100 random samples of 2 authors

Comparing our results to those of other studies focusing on large sets of authors is also difficult since most of them use much more data (over 10,000 words per author)(e.g. Argamon et al. (2007; Burrows (2007; Gamon (2004; Hirst and Feiguina (2007; Madigan et al. (2005), while the *Personae* corpus consists of about 1400 words per author. Argamon et al. (2003b) report on results in authorship attribution on twenty authors in a corpus of Usenet newsgroups on a variety of topics. Depending on the topic, results vary from 25% (books, computer theory) to 45% accuracy (computer language) for the 20-author task. Linguistic profiling, a technique presented by Van Halteren (2005), takes large numbers of linguistic features to compare separate authors to average profiles. In a set of eight authors, a linguistic profiling system correctly classifies 97% of the test documents. Madigan et al. (2005) use a collection of data released by Reuters consisting of 114 authors, each represented by 200 texts, minimally. Results of Bayesian multinomial logistic regression on this corpus show error rates between 97% and 20%, depending on the type of features applied. This is only partially comparable to the authorship attribution results on 145 authors presented in this paper because of the large amount of data in (Madigan et al., 2005), while our system works on limited data. In a study of weblog corpora, Koppel et al. (2006) show that authorship attribution with thousands of candidate authors is reasonably reliable, since the system gave an answer in 31.3% of the cases, while the answer is correct in almost 90% of the cases.

5.2. Personality prediction

We report on experiments on eight binary classification tasks (e.g., I vs. not-I) (cf. Table 5) and four tasks in which the goal is to distinguish between the two poles in the preferences (e.g., I vs. E) (cf. Table 6). Results are based on ten-fold cross-validation experiments with TiMBL (Memory-based learning (Daelemans and van den Bosch, 2005)). During training, TiMBL builds a model based on the training data by means of which the unseen

Task	Features	Precision	Recall	F-score	Accuracy
I	lex 3	56.70%	84.62%	67.90%	64.14%
	random	*44.1%*	*46.2%*		
E	cgp 3	58.09%	98.75%	73.15%	60.00%
	random	*54.6%*	*52.5%*		
N	cgp 3	56.92%	94.87%	71.15%	58.62%
	random	*48.7%*	*48.7%*		
S	pos 3	50.81%	94.03%	65.97%	55.17%
	random	*40.3%*	*40.3%*		
F	lex 3	73.76%	99.05%	84.55%	73.79%
	random	*72.6%*	*73.3%*		
T	lex 3	40.00%	50.00%	44.44%	65.52%
	random	*28.2%*	*27.5%*		
J	lex 3	81.82%	100.00%	90.00%	82.07%
	random	*77.6%*	*76.9%*		
P	lex 2	26.76%	67.86%	38.38%	57.93%
	random	*6.9%*	*7.1%*		

Table 5: TiMBL results for eight binary classification tasks

Class	Features
I	- ; conclusie misschien meer/inzicht *- ; conclusion maybe more/insight*
E	! uitvoeren we zij zelf de/mens *! execute we they the/human*
N	aangezien simuleren term de/mogelijkheid *since simulate term the/possibility*
S	gebeurt hersenen tastzin een/spontaan *happens brain sense a/spontaneous*
F	! beste denk ik een/beetje *! best think I a/little/bit*
T	: theorie omdat de/socio-politieke *: theory because the/socio-political*
J	mechanisme proces systeem mijn/mening *mechanism process system my/opinion*
P	ontwikkelingen symbiose tijdens van/levende *developments symbiosis during of/living*

Table 4: Predictive features per personality type (*with English translation)*

test instances can be classified. We present random and majority baselines. For Tables 5 and 6, the micro-average of every element in the contingency table over the ten folds was calculated. An example of some lexical predictive features for the personality types is presented in Table 4.

In Table 5, the personality prediction system is evaluated in terms of precision, recall, F-score, and accuracy, in order to get good grip on the system's performance. In heavily skewed classes, accuracy is less suitable to evaluate the system, because it does not take class distributions into account. Precision indicates the number of correctly classified instances of the positive class (resp., I, E, N, S, F, T, J, P), while recall represents the number of incorrectly classified negative instances (resp., not-I, not-E, not-N, not-S, not-F, not-T, not-J, not-P). F-score is the harmonic weighted mean of precision and recall, and accuracy indicates the number of correctly classified instances over the positive and negative classes. Results in accuracy are reported just for completeness, but we will not use them in our analyses since they are unfit for dealing with skewed classes.

The results in Table 5 reveal that four of the eight classes achieve an F-score of around 70% with the best scoring feature set. For heavily skewed classes with almost no counterexamples like F (only 28% of the instances is negative) and J (18% of the instances is negative), results are high, as expected (cf. random baselines). They vary between 85% and 90% F-score. For classes with hardly any positive instances, results are low, viz. 44% for F and 38% for P.

Task	Feature set	F-score [INFJ]	F-score [ESTP]	Avg. F-score	Acc.
I/E	lex 3	67.53%	63.24%	65.38%	65.52%
	random				*49.7%*
	majority				*55.2%*
N/S	pos 3	58.65%	64.97%	61.81%	62.07%
	random				*44.8%*
	majority				*53.8%*
F/T	lex 3	84.55%	13.64%	49.09%	73.79%
	random				*60.7%*
	majority				*72.4%*
J/P	lex 3	90.00%	13.33%	51.67%	82.07%
	random				*63.5%*
	majority				*80.7%*

Table 6: TiMBL results on four discrimination tasks

Table 6 shows results on the four discrimination tasks, which allows us to compare with results from other studies in personality prediction. Argamon and Levitan (2005) find appraisal adjectives and modifiers to be reliable markers (58% accuracy) of neuroticism, while extraversion can be predicted by function words with 57% accuracy. Nowson and Oberlander (2007) predict high/low extraversion with a 50.6% accuracy, while the system achieves 55.8% accuracy on neuroticism, 52.9% on agreeableness, and 56.6% on conscientiousness. Openness is excluded because of the skewed class distribution. Taking a top-down approach, Mairesse et al. (2007) report accuracies of 55.0% for extraversion, 55.3% for conscientiousness, 55.8% agreeableness, 57.4% for neuroticism, and 62.1% for openness.

For the I-E task - correlated to extraversion in the Big Five - we achieve an accuracy of 65.5%, which is better than Argamon and Levitan (2005) (57%), Nowson and Oberlander (2007) (51%), and Mairesse et al. (2007) (55%). For the N-S task - correlated to openness - we achieve the same result as Mairesse et al. (2007) (62%).

For the *F-T* and *J-P* tasks, the results hardly achieve higher than majority baseline, but nevertheless something is learned for the minority class, which indicates that the features selected work for personality prediction, even with heavily skewed class distributions.

6. Conclusions and Further Research

Results from experiments in authorship attribution on 145 authors indicate that in almost 50% of the cases, a text from one of the 145 authors is classified correctly. Using combinations of good working lexical and syntactic features leads to significant improvements. Exploratory experiments in personality prediction suggest that the first two personality dimensions (Introverted-Extraverted and iNtuitive-Sensing) can be predicted fairly accurately. Thanks to improvements in shallow text analysis, we can use syntactic features for the prediction of personality type and author.

Further research using the *Personae* corpus will involve a 'one vs. all' study in author verification, similar to the research by Argamon et al. (2003b), but on a significantly larger set of authors. We will also explore the use of Genetic Algorithm (GA) optimization for personality prediction and authorship attribution.

7. Acknowledgements

This study has been carried out in the framework of the "Computational Techniques for Stylometry for Dutch" project, funded by the National Fund for Scientific Research (FWO) in Belgium. We would like to thank the reviewers for their comments which helped improve this paper.

8. References

S. Argamon and S. Levitan. 2005. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 2005 joint conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH)*, pages 1–3.

S. Argamon, M. Koppel, J. Fine, and A. Shimoni. 2003a. Gender, genre, and writing style in formal written texts. *Text*, 23(3):321–346.

S. Argamon, M. Saric, and S. Stein. 2003b. Learning algorithms and features for multiple authorship discrimination. In *Proceedings of the 2003 International Joint Conferences on Artificial Intelligence (IJCAI): Workshop on Computational Approaches to Style Analysis and Synthesis*, pages 475–480.

S. Argamon, S. Dhawle, M. Koppel, and J. Pennebaker. 2005. Lexical predictors of personality type. In *Joint Annual Meeting of the Interface and the Classification Society of North America*.

S. Argamon, C. Whitelaw, P. Chase, S. Dhawle, S. Hota, N. Carg, and S. Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society of Information Science and Technology*, 58(6):802–822.

H.R. Baayen, H. Van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–131.

I. Briggs Myers and P.B. Myers. 1980. *Gifts differing: Understanding personality type*. Mountain View, CA: Davies-Black Publishing.

J. Burrows. 2002. 'Delta': a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

J. Burrows. 2007. All the way through: Testing for authorship in different frequency data. *Literary and Linguistic Computing*, 22(1):27–47.

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

W. Daelemans and A. van den Bosch. 2005. *Memory-Based Language Processing*. Studies in Natural Language Processing. Cambridge, UK: Cambridge University Press.

W. Daelemans, S. Bucholz, and J. Veenstra. 1999. Memory-Based Shallow Parsing. In *Proceedings of the Third Conference on Computational Natural Language Learning (CoNLL)*, pages 53–60.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2000. Authorship attribution with Support Vector Machines. *Applied Intelligence*, 19(1-2):109–123.

M. Gamon. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. In *Proceedings of the 2004 International Conference on Computational Linguistics (COLING)*, pages 611–617.

A.J. Gill and J. Oberlander. 2002. Taking care of the linguistic features of extraversion. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society (CogSci)*, pages 363–368.

A.J. Gill. 2003. *Personality and language: The projection and perception of personality in computer-mediated communication*. Ph.D. thesis, University of Edinburgh.

G. Hirst and O. Feiguina. 2007. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417.

C.G. Jung. 1921. *Psychologische Typen*. Walter-Verlag.

M. Koppel, J. Schler, S. Argamon, and E. Messeri. 2006. Authorship attribution with thousands of candidate authors. In *Proceedings of the 29th International Conference of the Special Interest Group on Information Retrieval (SIGIR)*, pages 659–660.

C. Labbé and D. Labbé. 2006. A tool for literary studies: intertextual distance and tree classification. *Literary and Linguistic Computing*, 21(3):311–326.

K. Luyckx and W. Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. In *Proceedings of the fifteenth meeting of Computational Linguistics in the Netherlands (CLIN 2004)*, pages 149–160.

K. Luyckx, W. Daelemans, and E. Vanhoutte. 2006. Stylogenetics: clustering-based stylistic analysis of literary corpora. In *Proceedings of the 5th Language Resources and Evaluation Conference: Workshop "Towards Computational Models of Literary Analysis" (LREC)*.

D. Madigan, A. Genkin, D. Lewis, S. Argamon, D. Fradkin, and L. Ye. 2005. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*.

F. Mairesse, M. Walker, M. Mehl, and R. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

R. McCrae and P. Costa. 1989. Reinterpreting the Myers-Briggs Type Indicator from the perspective of the Five-Factor Model of Personality. *Journal of Personality*, 57(1):17–40.

A. Miranda García and J. Calle Martín. 2007. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66.

S. Nowson and J. Oberlander. 2007. Identifying more bloggers. Towards large scale personality classification of personal weblogs. In *Proceedings of the 2007 International Conference on Weblogs and Social Media (ICWSM)*.

F. Sebastiani. 2002. Machine learning in automated text categorization. *Association for Computing Machinery (ACM) Computing Surveys*, 34(1):1–47.

E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214.

H. Van Halteren. 2005. Linguistic profiling for author recognition and verification. In *Proceedings of the 2005 Meeting of the Association for Computational Linguistics (ACL)*.