Willinsky, J., and M. Quint-Rapoport. *When Complementary and Alternative Medicine Practitioners Use PubMed.* Unpublished paper. University of British Columbia. 2007.

Windschuttle, Keith. "Edward Gibbon and the Enlightenment." *The New Criterion* 15.10. June 1997. <http://www.newcriterion.com/archive/15/jun97/gibbon.htm>.

Wineburg, S. "Historical Problem Solving: A Study of the Cognitive Processes Used in the Evaluation of Documentary and Pictorial Evidence." *Journal of Educational Psychology* 83.1 (1991): 73-87.

Wineburg, S. "Reading Abraham Lincoln: An Expert/Expert Study in the Interpretation of Historical Texts." *Cognitive Science* 22.3 (1998): 319-346.

Wyatt, D., M. Pressley, P.B. El-Dinary, S. Stein, and R. Brown. "Comprehension Strategies, Worth and Credibility Monitoring, and Evaluations: Cold and Hot Cognition When Experts Read Professional Articles That Are Important to Them." *Learning and Individual Differences* 5 (1993): 49–72.

Using syntactic features to predict author personality from text

Kim Luyckx
kim.luyckx@ua.ac.be
University of Antwerp, Belgium

Walter Daelemans
walter.daelemans@ua.ac.be
University of Antwerp, Belgium

Introduction

The style in which a text is written reflects an array of meta-information concerning the text (e.g., topic, register, genre) and its author (e.g., gender, region, age, personality). The field of stylometry addresses these aspects of style. A successful methodology, borrowed from text categorisation research, takes a two-stage approach which (i) achieves automatic selection of features with high predictive value for the categories to be learned, and (ii) uses machine learning algorithms to learn to categorize new documents by using the selected features (Sebastiani, 2002). To allow the selection of linguistic features rather than (*n*-grams of) terms, robust and accurate text analysis tools are necessary. Recently, language technology has progressed to a state of the art in which the systematic study of the variation of these linguistic properties in texts by different authors, time periods, regiolects, genres, registers, or even genders has become feasible.

This paper addresses a not yet very well researched aspect of style, the author's personality. Our aim is to test whether personality traits are reflected in writing style. Descriptive statistics studies in language psychology show a direct correlation: personality is projected linguistically and can be perceived through language (e.g., Gill, 2003; Gill & Oberlander, 2002; Campbell & Pennebaker, 2003). The focus is on extraversion and neuroticism, two of "the most salient and visible personality traits" (Gill, 2003, p. 13). Research in personality prediction (e.g., Argamon et al., 2005; Nowson & Oberlander, 2007; Mairesse et al., 2007) focuses on openness, conscientiousness, extraversion, agreeableness, and neuroticism.

We want to test whether we can automatically predict personality in text by studying the four components of the Myers-Briggs Type Indicator: Introverted-Extraverted, Intuitive-Sensing, Thinking-Feeling, and Judging-Perceiving. We introduce a new corpus, the *Personae* corpus, which consists of Dutch written language, while other studies focus on English. Nevertheless, we believe our techniques to be transferable to other languages.

Related Research in Personality Prediction

Most of the research in personality prediction involves the Five-Factor Model of Personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The 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. Nowson & 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. 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.

Corpus Construction

Our 200,000-word *Personae* corpus consists of 145 BA student essays of about 1,400 words about a documentary on Artificial Life in order to keep genre, register, topic and age relatively constant. These 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. They took an online MBTI test and submitted their profile, the text and some user information. All students released the copyright of their text to the University of Antwerp and explicitly allowed the use of their text and personality profile for research, which makes it possible to distribute the corpus.

The Myers-Briggs Type Indicator (Myers & Myers, 1980) is a forced-choice test based on Jung's personality typology which categorizes a person on four preferences:

- **I**ntroversion and **E**xtraversion (attitudes): I's tend to reflect before they act, while E's act before they reflect.
- i**N**tuition and **S**ensing (information-gathering): N's rely on abstract or theoretical information, while S's trust information that is concrete.
- **F**eeling and **T**hinking (decision-making): While F's decide based on emotions, T's 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 personality traits of extraversion and openness, to a lesser extent with agreeableness and consciousness, but not with neuroticism (McCrae & Costa, 1989).

The participants' characteristics are too homogeneous for experiments concerning gender, mother tongue or region, but we find interesting distributions in at least two of the four MBTI preferences: .45 I vs. .55 E, .54 N vs. .46 S, .72 F vs. .28 F, and .81 J and .19 P.

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

Feature Extraction

While most stylometric studies are based on token-level features (e.g., word length), word forms and their frequencies of occurrence, syntactic features have been proposed as more reliable style markers since they are not under the conscious control of the author (Stamatatos et al., 2001).

We use Memory-Based Shallow Parsing (MBSP) (Daelemans et al., 1999), which gives an incomplete parse of the input text, to extract reliable syntactic features. MBSP tokenizes, performs a part-of-speech analysis, looks for chunks (e.g., noun phrase) and detects subject and object of the sentence and some other grammatical relations.

Features occurring more often than expected (based on the chi-square metric) in either of the two classes are extracted automatically for every document. 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. These features have been proven useful in stylometry (cf. Stamatatos et al., 2001) and are now tested for personality prediction.

Experiments in Personality Prediction and Discussion

We report on experiments on eight binary classification tasks (e.g., I vs. not-I) (cf. Table 1) and four tasks in which the goal is to distinguish between the two poles in the preferences (e.g., I vs. E) (cf. Table 2). Results are based on ten-fold cross-validation experiments with TiMBL (Daelemans & van den Bosch, 2005), an implementation of memory-based learning (MBL). MBL stores feature representations of training instances in memory without abstraction and classifies new instances by matching their feature representation to all instances in memory. We also report random and majority baseline results. Per training

document, a feature vector is constructed, containing comma-separated binary or numeric features and a class label. During training, TiMBL builds a model based on the training data by means of which the unseen test instances can be classified.

Task	Feature set	Precision	Recall	F-score	Accuracy
Introverted	lex 3-grams *random*	56.70% *44.1%*	84.62% *46.2%*	67.90%	64.14%
Extraverted	cgp 3-grams random	58.09% 54.6.%	98.75% *52.5%*	73.15%	60.00%
iNtuitive	cgp 3-grams *random*	56.92% *48.7.%*	94.87% *48.7%*	71.15%	58.62%
Sensing	pos 3-grams *random*	50.81% *40.3.%*	94.03% *40.3.%*	65.97%	55.17%
Feeling	lex 3-grams *random*	73.76% *72.6%*	99.05% *73.3.%*	84.55%	73.79%
Thinking	lex 1-grams *random*	40.00% *28.2.%*	50.00% *27.5%*	44.44%	65.52%
Judging	lex 3-grams *random*	81.82% *77.6%*	100.00% *76.9%*	90.00%	82.07%
Perceiving	lex 2-grams *random*	26.76% *6.9%*	67.86% *7.1%*	38.38%	57.93%

Table 1:TiMBL results for eight binary classification tasks

Table 1 suggests that tasks for which the class distributions are not skewed (I, E, N and S) achieve F-scores between 64.1% and 73.2%. As expected, results for Feeling and Judging are high, but the features and methodology still allow for a score around 40% for tasks with little training data.

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

Table 2:TiMBL results for four discrimination tasks

Table 2 shows results on the four discrimination tasks, which allows us to compare with results from other studies in personality prediction. Argamon et al. (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 & 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 et al. (2005) (57%), Nowson & 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.

Conclusions and Future Work

Experiments with TiMBL suggest that the first two personality dimensions (Introverted-Extraverted and iNtuitive-Sensing) can be predicted fairly accurately. We also achieve good results in six of the eight binary classification tasks. 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 study of stylistic variation between the 145 authors. A lot of the research in author recognition is performed on a closed-class task, which is an artificial situation. Hardly any corpora – except for some based on blogs (Koppel et al., 2006) – have more than ten candidate authors. The corpus allows the computation of the degree of variability encountered in text on a single topic of different (types) of features when taking into account a relatively large set of authors. This will be a useful complementary resource in a field dominated by studies potentially overestimating the importance of these features in experiments discriminating between only two or a small number of authors.

Acknowledgements

This study has been carried out in the framework of the Stylometry project at the University of Antwerp. The "Computational Techniques for Stylometry for Dutch" project is funded by the National Fund for Scientific Research (FWO) in Belgium.

References

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

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

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

Daelemans, W., Bucholz, S. and Veenstra, J. (1999), Memory-Based Shallow Parsing, *Proceedings of the 3rd Conference on Computational Natural Language Learning CoNLL*, pp. 53-60.

Gill, A. (2003), Personality and language: The projection and perception of personality in computer-mediated communication, PhD thesis, University of Edinburgh.

Gill, A. & Oberlander J. (2002), Taking care of the linguistic features of extraversion, *Proceedings of the 24th Annual Conference of the Cognitive Science Society,* pp. 363-368.

Koppel, M., Schler, J., Argamon, S. and Messeri, E. (2006), Authorship attribution with thousands of candidate authors, *Proceedings of the 29th ACM SIGIR Conference on Research and Development on Information Retrieval*, pp. 659-660.

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

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

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

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

Sebastiani, F. (2002), Machine learning in automated text categorization, *ACM Computing Surveys* 34(1), 1-47.

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

An Interdisciplinary Perspective on Building Learning Communities Within the Digital Humanities

Simon Mahony
simon.mahony@kcl.ac.uk
King's College London

Introduction

Recent research at the Centre for Computing in the Humanities at King's College London has focussed on the role and place of the digital humanities in the academic curriculum of Higher Education (see Jessop:2005, Jessop:forthcoming). This work is based on the experience of both our undergraduate and postgraduate programmes focusing particularly on the way in which students are encouraged to integrate the content of a variety of digital humanities courses and apply it to their own research project. In the case of the undergraduates this is developed in conjunction with their home department. These courses are designed to train not just the new generation of young scholars in our discipline but also the majority who will gain employment in a variety of professions in industry and commerce.

Our students come from a range of disciplines and backgrounds within the humanities and what is highlighted in each case is the necessity to ensure that their projects meet the scholarly criteria of their home disciplines and the interdisciplinary aspects of humanities computing. This emphasises the need for training the students in collaborative method and reflective practice; the need to build a community of learning which will lead to a community of practice. This paper discusses recent research and initiatives within distance learning, focussing on how these can be repurposed for campus-based courses, and is illustrated by the findings of their use in a digital humanities course.

Context

There have been a number of initiatives that are pertinent to this topic. The published report on the accomplishments of the Summit on Digital Tools for the Humanities convened in 2005 at the University of Virginia (http://www.iath.virginia.edu/dtsummit/) identified areas where innovative change was taking place that could lead to what they referred to as "a new stage in humanistic scholarship". The style of collaboration enabled by digital learning community tools is identified as one such area. This has been further reinforced at the National Endowment of the Humanities hosted Summit Meeting of Digital Humanities Centers and Funders held in April 2007 at the University of Maryland.