# A SYSTEMIC FUNCTIONAL APPROACH TO AUTOMATED AUTHORSHIP ANALYSIS

*Shlomo Argamon* and Moshe Koppel***

## INTRODUCTION

Attribution of anonymous texts, if not based on factors external to the text (such as paper and ink type or document provenance, as used in forensic document examination), is largely, if not entirely, based on considerations of language *style*. We will consider here the question of how to best deconstruct a text into quantitative features for purposes of stylistic discrimination. Two key considerations inform our analysis. First, such features should support accurate classification by automated methods. Second, and no less importantly, such features should enable a clear explanation of the stylistic difference between stylistic categories (read: authors) and why a disputed text appears more likely to fall into one or another category. The latter consideration is particularly important when a nonexpert, such as a judge or jury, must evaluate the results and reliability of the analysis.

We start from the intuitive notion that style is indicated in a text by those features of the text that indicate the author's choice of one mode of expression from among a set of equivalent modes for a given content. There are many ways in which such choices manifest themselves in a text. Specific words and phrases may be chosen more frequently by certain authors than others, such as the phrase "cool-headed logician" favored by the Unabomber. Some authors may habitually use certain syntactic

* Linguistic Cognition Laboratory, Department of Computer Science, Illinois Institute of Technology, argamon@iit.edu.

** Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel, moishk@gmail.com.

constructions more frequently, as in Hemingway's preference for short, simple clauses. Differences between authors will also arise at the level of the organization of the text as a whole, as some people may prefer to make reasoned arguments from evidence to conclusions, and others may prefer emotional appeals organized differently.

However, all of these "surface" linguistic phenomena have multiple potential underlying causes, not only authorship. They include the genre, register, and purpose of the text as well as the educational background, social status, and personality of the author and audience.[1] What all these dimensions of variation have in common, though, is independence, to a greater or lesser extent, of the "topic" of the text. Hence the traditional focus in computational authorship attribution on features such as function word usage; vocabulary richness and complexity measures; and frequencies of different syntactic structures; which are essentially nonreferential.

Early statistical attribution techniques relied on relatively small numbers of such features, while developments in machine learning and computational linguistics over the last fifteen to twenty years have enabled larger numbers of features to be generated for stylistic analysis. However, in almost no case is there strong theoretical motivation behind the input feature sets, such that the features have clear interpretations in stylistic terms.

We argue, however, that without a firm basis in a linguistic theory of meaning (not just of syntax), we are unlikely to gain any true insight into the nature of any stylistic distinction being studied. Such understanding is key to both establishing and explaining evidence for a proposed attribution. Otherwise, an attribution method is merely a black box that may appear to work for extrinsic or accidental reasons but not actually give reliable results in a given case. Furthermore, an attribution method that produces insight into the relevant language variation is more likely to be useful and accepted in a forensic context, all else being equal, as the judge and jury will be better able to understand the results.

---

[1] DOUGLAS BIBER & SUSAN CONRAD, REGISTER, GENRE, AND STYLE (P. Austin et al. eds., 2009).

We therefore sketch here a computationally tractable formulation of linguistically and stylistically well-motivated features we have developed that permits text classification based on specific variation in choice of nonreferential meanings. The system produces meaningful information about the stylistic distinctions being analyzed, which can be used for interpretative and forensic purposes. We will explain our methodology and then use it as a case study for what any such methodology should provide.

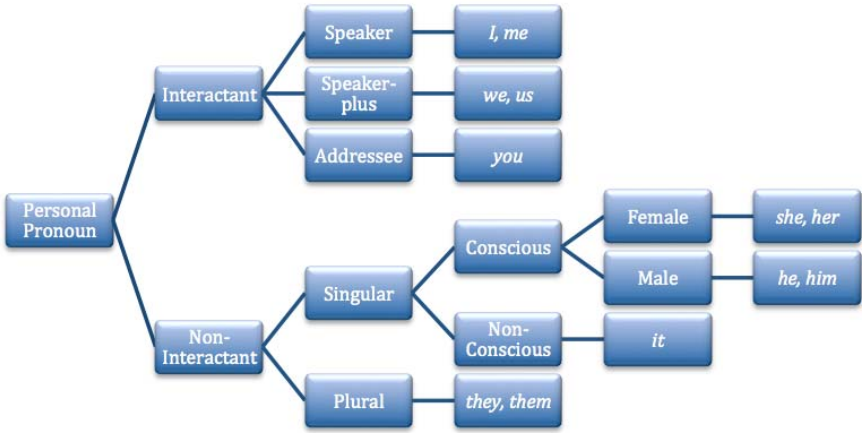
Before we begin, it is worth briefly surveying the variety of problems that fall under the umbrella of "authorship analysis." The simplest form of the problem is where an anonymous document is potentially attributable to one of a relatively small number (two to fifty, or so) of suspects. The question is then simply which of the suspects has a writing style most like that of the anonymous document. More difficult (and much more likely in the real world) is the case where the document might not be authored by any of the suspects at all—in this case we must be able to determine that the document is not enough like any of the suspects to attribute authorship. The hardest version of this scenario is *authorship verification*, where the question is whether a single suspect did or did not author the anonymous document. All such *authorship attribution* scenarios assume a known set of suspects who are being evaluated for authorship of the questioned document. We require some quantity of texts written by each of the suspects to determine authorship. On the other hand, if, as is often the case in police investigations, specific suspects are not known, we must consider the task of *authorship profiling*, determining as much about the author as possible, based upon clues in the document. As we will discuss below, a number of personal characteristics of an author can be reliably estimated from stylistic cues in a document. But first we will consider generally how we can quantitatively characterize the style of a text for computational analysis.

## I. FUNCTIONAL LEXICAL FEATURES

Our methodology is based on Halliday's Systemic Functional Grammar[2] ("SFG"), which we find to be particularly well-suited to the sort of computational analysis we seek. SFG explicitly recognizes and represents various aspects of nonreferential meaning as part of the general grammar, which makes it directly adaptable to stylistic classification.[3] We do not claim, of course, that SFG is the only, or even necessarily the best, approach but rather one that we have found convenient.

We start from the SFG idea that grammar is a set of constraints on how one may express meaning.[4] Grammar is thus a network of possible choices, with more general or abstract choices constraining which more specific choices are allowed. This network of choices is called a *system network*.[5] As a simple example, consider the (partial) system network for pronouns in English, seen below in Figure 1. This network forms a neat hierarchical taxonomy, though not all do. As an approximation we can extract a set of taxonomies (trees) from the full network.

*Figure 1. System diagram for Personal Pronouns, shown as a taxonomic tree.*




---

[2] *See* M.A.K. HALLIDAY & CHRISTIAN M.I.M. MATTHIESSEN, AN INTRODUCTION TO FUNCTIONAL GRAMMAR 37–63 (3d ed. 2004).

[3] *Id.* at 50–53.

[4] *Id.* at 1.

[5] *Id.* at 23.

Given this taxonomy, we may define numeric features describing the statistical "stylistics" of a text via the collection of conditional frequencies of each node in the tree given its parent. Thus, for example, we measure the frequency of "Speaker" pronouns out of all occurrences of "Interactant" pronouns, and so on. This has a straightforward interpretation of measuring the biases of how texts of a given style (e.g., by a given author) prefer certain choices of how to express more general meanings. By using such biases to analyze authorship, we seek to capture relevant *codal variation*, as contrasted with register[6] (variation in these probabilities due to a text's functional context), or *dialect* (variation in how specific meanings are realized (e.g., use of "y'all" for plural "you")).

To give a flavor of these features, here are brief descriptions of several system networks that we have found useful for stylistic classification.[7]

### *A. Conjunctions*

How an author conjoins phrases and clauses is an indication of how the author organizes concepts and relates them to each other. Words and phrases that conjoin clauses (such as "and," "while," and "in other words") are organized in SFG in the CONJUNCTION system network.[8] Types of conjunctions serve to link a clause with its textual context, by denoting how the given clause expands on some aspect of its preceding context. The three top-level options of CONJUNCTION are Elaboration, Extension, and Enhancement, defined as:

- Elaboration: Deepening the content in its context by exemplification or refocusing ("for example," "in other words," "i.e.");

---

[6] Ruqaiya Hasan, *Code, Register, and Social Dialect, in* 2 CLASS, CODES AND CONTROL: APPLIED STUDIES TOWARDS A SOCIOLOGY OF LANGUAGE 224, 253–92 (Basil B. Bernstein ed., 1973).

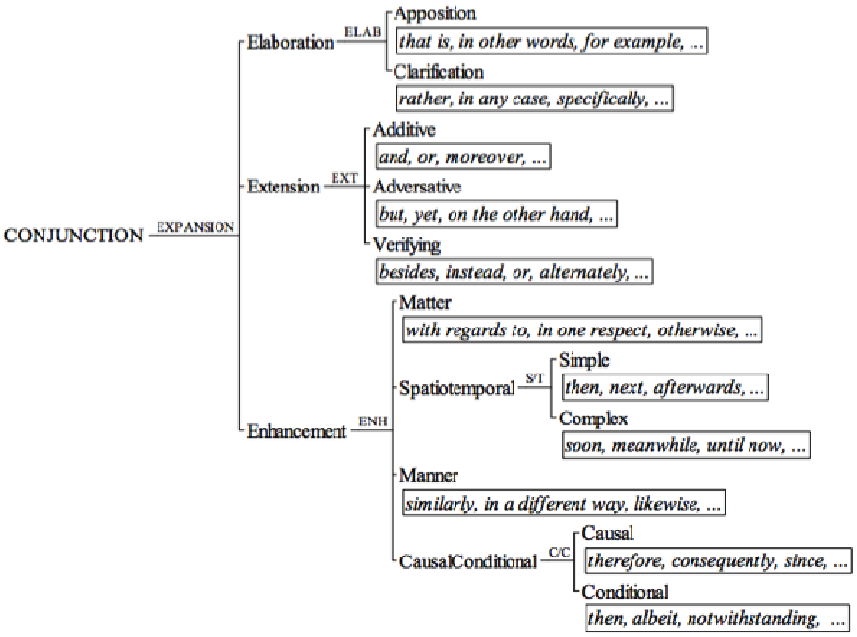
[7] For a more detailed discussion of these features, and the mathematical models involved, see Shlomo Argamon et al., *Stylistic Text Classification Using Functional Lexical Features,* 58 J. AM. SOC'Y INFO. SCI. & TECH. 802, 802–22 (2007).

[8] *See* HALLIDAY & MATTHIESSEN, *supra* note 2, at 538–39.

- Extension: Adding new related information, perhaps contrasting with the current information ("and," "or," "furthermore," "on the other hand");
- Enhancement: Qualifying the context by circumstance or logical connection ("and then," "because," "similarly").[9]

Each option also has several subcategories that further subdivide the ways in which information units in a text can be linked together.

*Figure 2. System diagram for Conjunction.*




### *B. Prepositions*

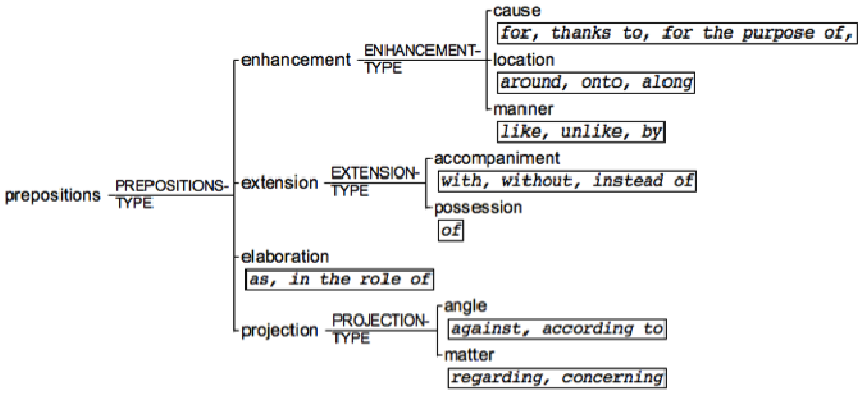
Similarly, prepositions serve to expand the meaning of a phrase or clause by connecting to it a phrase (usually a noun phrase). The high-level structure of the PREPOSITION system is thus similar to that of CONJUNCTION, with four top-level options:

- Elaboration: Exemplification ("as," "in the role of");

[9] *Id.* at 540–48.

- Enhancement: Qualifying context temporally, spatially, or causally ("around," "thanks to," "during");
- Extension: Adding related information about an object or event ("of," "without," "besides");
- Projection: Using an object to construe the meaning or significance of another ("against," "regarding," "according to").[10]

*Figure 3. System Diagram for Prepositions.*




### *C. Modality*

The MODALITY system comprises four taxonomies describing choices in how to describe the level of typicality or necessity of facts and events. Syntactically, modality can be realized through modal verbs (e.g., "can," "might," "should," "must"); adverbial adjuncts (e.g., "probably," "preferably"); or projective clauses (e.g., "I think that," "It is necessary that"). The four attributes of any modal expression are:

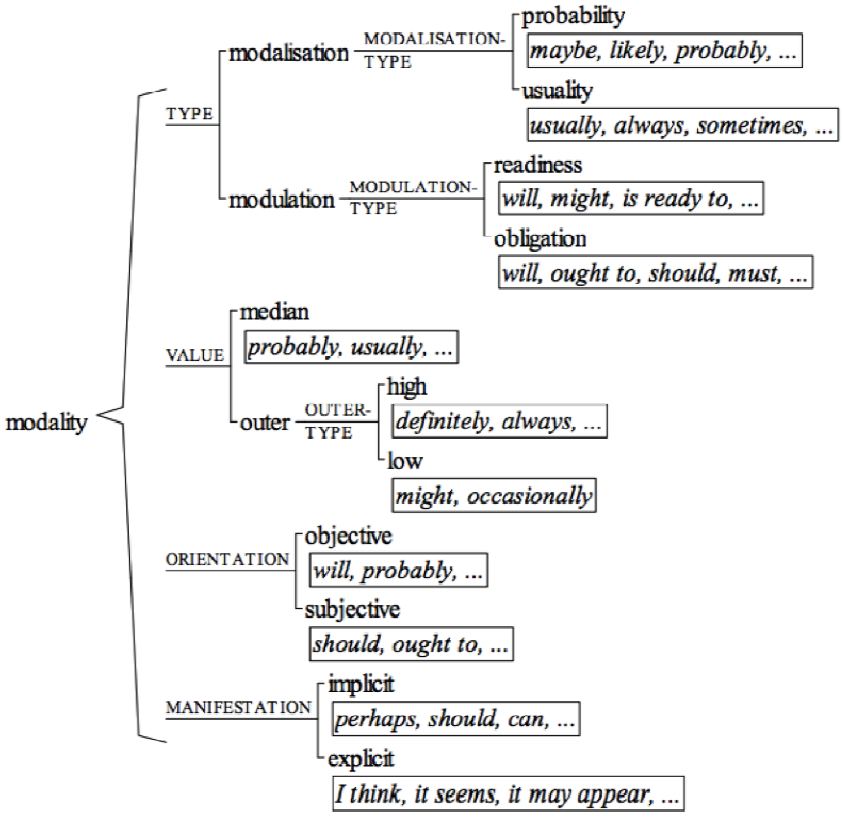
- Type: What kind of modality is being expressed?
  - Modalization: How "typical" is it? ("probably," "seldom")

[10] *See* CHRISTIAN M.I.M. MATTHIESSEN, LEXICO-GRAMMATICAL CARTOGRAPHY: ENGLISH SYSTEMS (1995).

    - Modulation: How "necessary" is it? ("ought to," "allowable")
- Value: What degree of the relevant modality scale is being averred?
    - Median: The "normal" amount ("likely," "usually")
    - Outer: An extreme (either high or low) amount ("never," "maybe," "must")
- Orientation: What is the relation to the speaker/writer of the modality expressed?
    - Objective: Modality expressed irrespective of the speaker/writer ("maybe," "always")
    - Subjective: Modality expressed relative to the speaker/writer ("We think," "I need")
- Manifestation: How is the modal assessment related to the event being assessed?
    - Implicit: Modality realized "in-line" by an adjunct or modal auxiliary ("preferably," "maybe")
    - Explicit: Modality realized by a projective verb, with the nested clause being assessed ("It is better to," "It is possible to")[11]

[11] *See* HALLIDAY & MATTHIESSEN, *supra* note 2, at 612–25.

*Figure 4. System diagram for Modality. Note the four parallel taxonomies.*




## II. EXPERIMENTS IN AUTHORSHIP PROFILING

The uses of these features can be seen both in authorship attribution and in *authorship profiling*, where we seek to determine characteristics of a text's author (such as sex, age, or personality), even in the absence of any specific candidate authors. We describe here some experiments we have done on authorship profiling for author sex, age, native language, and personality.[12]

In these experiments, we compared the use of functional lexical features as above with content-based features, namely,

[12] *See* Shlomo Argamon et al., *Automatically Profiling the Author of an Anonymous Text*, COMM. ACM, Feb. 2009, at 119.

individual words. In order to keep the number of features reasonably small, we consider just the 1,000 words that appear sufficiently frequently in the corpus and that discriminate best between the classes of interest (determined by "information-gain" on a holdout set).

We note that the use of content-based features for authorship studies can be problematic. One must be even more wary of content markers potentially being artifacts of a particular writing situation or experimental setup and thus producing overly optimistic results that will not be borne out in real-life applications. For example, were we to seek to identify Arthur Conan Doyle's writing by the high frequency of the words "Sherlock," "Holmes," and "Watson," we would misattribute any works not part of that detective series. We will therefore be careful to distinguish results that exploit content-based features from those that do not.

Whatever features are used in a particular experiment, we represent a document as a numerical vector $X$. Once labeled training documents have been represented in this way, we can apply machine-learning algorithms to learn classifiers that assign new documents to categories. Generally speaking, the most effective multiclass (i.e., more than two classes) classifiers for authorship studies all share the same structure: we learn a weight vector $W_j$ for each category $c_j$ and then assign a document, $X$, to the class for which the inner product $W_j * X$ is maximal. The weight vector is learned based on a *training set* of data points, each labeled with its correct classification. There are a number of effective algorithms for learning such weight vectors; we use here Bayesian Multinomial Regression ("BMR"),[13] which we have found to be both efficient and accurate. BMR is a probabilistically well-founded multivariate variant of logistic regression, which tends to work well for problems with large numbers of variables (as here).[14] BMR has

---

[13] *See* Alexander Genkin et al., *Large-Scale Bayesian Logistic Regression for Text Categorization*, 49 TECHNOMETRICS 291, 291–304 (2007).

[14] When seeking to construct predictive models from data with a very large number of variables, it is possible that a model can easily be found to fit the known data accidentally, just because there are many parameters in the model that can be adjusted. Such a model will then not classify new data

also been shown specifically to be effective for text classification and related problems.[15] Other learning methods such as support vector machines[16] generally work just as well.

### *A. Test Data*

In the experiments described below, we sought to profile documents by four common author characteristics: sex, age, native language, and personality type. The first three of these have obvious application in the investigative and forensic contexts. Personality type is more useful for investigations but can also provide corroborative evidence for identification when personality information about a suspect is known. We first describe in this section the data sets, comprising labeled collections of texts, that we used to learn and test our classification models. In the following section, we will describe the experimental procedure and results.

**Sex and Age.** Our corpus[17] for both author sex and age consists of the full set of postings of 19,320 blog authors (each text is the full set of posts by a given author) writing in English. The (self-reported) age and gender of each author is known and for each age interval the corpus includes an equal number of male and female authors. The texts range in length from several hundreds to tens of thousands of words, with a mean length of 7,250 words per author. Based on each blogger's reported age, we label each blog in our corpus as belonging to one of three

---

well. This problem is known as *overfitting*. *See* Tom Dietterich, *Overfitting and Undercomputing in Machine Learning*, ACM COMPUTING SURVS., Sept. 1995, at 326–27. BMR, and other modern learning algorithms, seek to minimize this problem by various mathematical methods.

[15] *See* Genkin et al., *supra* note 13; *see also* Moshe Koppel et al., *Automatically Classifying Documents by Ideological and Organizational Affiliation*, PROC. 2009 IEEE INT'L CONF. ON INTELLIGENCE & SECURITY INFORMATICS, at 176.

[16] *See* NELLO CRISTIANINI & JOHN SHAWE-TAYLOR, AN INTRODUCTION TO SUPPORT VECTOR MACHINES AND OTHER KERNEL-BASED LEARNING METHODS 7 (2000).

[17] First described in Jonathan Schler et al., *Effects of Age and Gender on Blogging,* AAAI SPRING SYMPOSIUM: COMPUTATIONAL APPROACHES TO ANALYZING WEBLOGS, 2006, at 199.

age groups: thirteen to seventeen (42.7%), twenty-three to twenty-seven (41.9%) and thirty-three to forty-seven (15.5%). Intermediate age groups were removed to avoid ambiguity since many of the blogs were written over a period of several years. Our objective is to identify to which of these three age intervals an anonymous author belongs.

**Native Language.** We used the International Corpus of Learner English ("ICLE"),[18] which was assembled for the precise purpose of studying the English writing of nonnative English speakers from a variety of countries. All the writers in the corpus are university students (mostly in their third or fourth year) studying English as a second language. All are roughly the same age (in their twenties) and are assigned to the same proficiency level in English. All texts are short student essays on a similar set of topics, so they are in the same genre. We consider five subcorpora from Russia, the Czech Republic, Bulgaria, France, and Spain. To balance the corpus, we took 258 authors from each subcorpus (randomly discarding any surplus). All texts in the resulting corpus are between 579 and 846 words long. Our objective is to determine which of the five languages is the native tongue of an anonymous author writing in English.

**Personality.** We used essays written by psychology undergraduates at the University of Texas at Austin collected by James W. Pennebaker.[19] Students were instructed to write a short "stream of consciousness" essay wherein they tracked their thoughts and feelings over a twenty minute free-writing period. The essays range in length from 251 to 1,951 words. Each writer also filled out a questionnaire testing for the "Big Five" personality dimensions: neuroticism, extraversion, openness, conscientiousness, and agreeableness. We consider here just the dimension of neuroticism (roughly, tendency to worry or be anxious), as methods and results for other personality factors are qualitatively similar. We defined "positive" examples to be the

---

[18] *International Corpus of Learner English,* UNIVERSITE CATHOLIQUE DE LOUVAIN, http://www.uclouvain.be/en-cecl-icle.html (last visited Mar. 2, 2013).

[19] Shlomo Argamon et al., *Lexical Predictors of Personality Type,* PROC. JOINT ANN. MEETING INTERFACE & CLASSIFICATION SOC'Y N. AM., 2005.

participants with neuroticism scores in the upper third of the authors, and 'negative' examples to be those with scores in the lowest third. The rest of the data were ignored, and the final corpus consists of 198 examples.

### *B. Procedure and Results*

Accuracy results for the above profiling tasks are given in Table 1 for different combinations of features. Recall that a training set is required for the system to learn a classification model for any given task. The accuracy of the system must be evaluated on data separate from the training data, since even perfect performance on the training data is easy to achieve and meaningless in terms of the real-world potential accuracy of the system. Hence each dataset needs to be divided into disjoint training and test sets for evaluation. To maximize use of limited data, a standard technique, called *ten-fold cross-validation*, is used to divide the data randomly into ten equal parts, then to perform ten train-test runs, each run training on nine-tenths of the data and testing on the remaining tenth. The average accuracy over these ten runs is a good estimate of the actual performance of the system on new data.

Accuracy is measured simply as the percentage of text examples that the system classified correctly. In any given classification problem, there is a baseline performance, given by the percentage of the data falling into the majority class. This percentage indicates the performance of the trivial classifier that just classifies every example as that majority class. If the accuracy of our classification system is significantly higher than this baseline performance, the system can be said to work; the higher the accuracy, the better it works.

Consider now the results for authorship profiling given in Table 1. We first note that while in most cases (other than neuroticism) content words help, style features often give good results on their own. More informative are the highest weighted features for each output class, given in Table 2. For sex, the style features that prove to be most useful for gender discrimination are determiners and certain prepositions (markers of male writing) and pronouns (markers of female writing),

which is consistent with other studies. For age, we see a preference for more formal writing in the older bloggers (prepositions and determiners), though the content features in this case give more insight, in terms of the usual concerns of people in different age groups. For native language, we see some interesting stylistic patterns, in that native speakers of Slavic languages have clear preferences for personal pronouns, particularly first person, while Romance language speakers have distinctive (and different) patterns of verb auxiliary use. The content features in this case, while more dispositive, are clearly not useful in any context where deception would come into play, as they can be easily planted by a deceptive writer.

Finally, we see that neurotics tend to refer more often to themselves, use pronouns as subjects rather than as objects in a clause, and consider explicitly who benefits from some action (through prepositional phrases involving, e.g., "for" and "in order to"); nonneurotics, on the other hand, tend to use less precise specification of objects or events (determiners and adjectives such as "a" or "little") and show more concern with how things are or should be done (via prepositions such as "by" or "with" and modals such as "ought to" or "should").

In other experiments we have done using features of lexicogrammar indicative of writers' attitudes, we found (unsurprisingly) texts by neurotic individuals to be characterized more by focus on, e.g., negative orientation and affect, whereas texts by nonneurotics focused more on positive orientation and appreciation.[20] That is, neurotics evaluated objects and propositions more negatively and more in terms of feelings, while nonneurotics did so more positively and more in terms of objective characteristics.

---

[20] *See id*.

***Table 1.*** *Classification accuracy (10-fold cross-validation) for authorship profiling using different feature sets.*

| | **Baseline** | **Style** | **Content** | **Style+Content** |
|---|---|---|---|---|
| **Gender (2 classes)** | *50.0* | 72.0 | 75.1 | **76.1** |
| **Age (3 classes)** | *42.7* | 66.9 | 75.5 | **77.7** |
| **Language (5 classes)** | *20.0* | 65.1 | **82.3** | 79.3 |
| **Neuroticism (2 classes)** | *50.0* | **65.7** | 53.0 | 63.1 |

***Table 2.*** *Most important Style and Content features (by information gain) for each class of texts in each profiling problem.*

| **Class** | **Style Features** | Content Features |
|---|---|---|
| Female | **personal pronoun, *I*, *me*, *him*, *my*** | ***cute*, *love*, *boyfriend*, *mom*, *feel*** |
| Male | **determiner, *the*, *of*, preposition-matter, *as*** | *system, software, game, based, site* |
| Teens | ***im*, *so*, *thats*, *dont*, *cant*** | *haha, school, lol, wanna, bored* |
| Twenties | **preposition, determiner, *of*, *the*, *in*** | *apartment, office, work, job, bar* |
| Thirties+ | **preposition, *the*, determiner, *of*, *in*** | *years, wife, husband, daughter, children* |
| Bulgarian | **conjunction-extension, pronoun-interactant,** *however*, **pronoun-conscious,** *and* | *bulgaria, university, imagination, bulgarian, theoretical* |
| Czech | **personal pronoun,** *usually*, *did*, *not*, *very* | *czech, republic, able, care, started* |
| French | *indeed*, **conjunction-elaboration,** *will*, **auxverb-future, auxverb-probability** | *identity, europe, european, nation, gap* |
| Russian | *can't*, *i*, *can*, *over*, *every* | *russia, russian, crimes, moscow, crime* |
| Spanish | **determiner-specific**, *this, going_to, because, although* | *spain, restoration, comedy, related, hardcastle* |
| Neurotic | *myself*, **subject pronoun, reflexive pronoun, preposition-behalf, pronoun-speaker** | *put, feel, worry, says, hurt* |
| Nonneurotic | ***little*, auxverbs-obligation, nonspecific determiner, *up*, preposition-agent** | *reading, next, cool, tired, bed* |

## III. DISCUSSION

We have sketched here a framework for addressing authorship attribution as a question of evaluating codal variation by estimating the probabilities of different grammatical choices by different authors or kinds of authors. These features perform as well or better in our empirical tests as other sorts of features and (often) have the advantage of giving meaningful insight into the underlying stylistic differences between authors.

As we have argued above and elsewhere,[21] such insight should be considered a key criterion for authorship attribution methods, along with accuracy and reliability. Without such understanding, it is extremely difficult, or impossible, to have real confidence that results in any specific instance are reliable, due to the large number and variety of possible confounding factors (dialect and register variation and the like). Results that can be meaningfully interpreted, however, also make the task of conveying their import to nonexperts, including judges and juries, much easier.

It also seems likely that an operationalization of idiolect as a systematic skewing of probabilities in system taxonomies, as developed above, helps to put the problem of author analysis into a larger theoretical context. This context recognizes language variation due to code, as in authorial differences, as well as variation due to register and genre. By identifying author analysis as one aspect of a continuum of similar kinds of variation, we may hope to disentangle the omnipresent effects of register and genre variation when analyzing authorship.

---

[21] *See, e.g.*, Shlomo Argamon & Moshe Koppel, *The Rest of the Story: Finding Meaning in Stylistic Variation, in* THE STRUCTURE OF STYLE: ALGORITHMIC APPROACHES TO UNDERSTANDING MANNER AND MEANING 79 (Shlomo Argamon et al. eds., 2010); *see also* Argamon et al., *supra* note 7.