Stylometric Analysis of Scientific Articles

Shane Bergsma, Matt Post, David Yarowsky
Department of Computer Science and Human Language Technology Center of Excellence
Johns Hopkins University
Baltimore, MD 21218, USA
sbergsma@jhu.edu, post@cs.jhu.edu, yarowsky@cs.jhu.edu

Abstract

We present an approach to automatically recover hidden attributes of scientific articles, such as whether the author is a native English speaker, whether the author is a male or a female, and whether the paper was published in a conference or workshop proceedings. We train classifiers to predict these attributes in computational linguistics papers. The classifiers perform well in this challenging domain, identifying non-native writing with 95% accuracy (over a baseline of 67%). We show the benefits of using *syntactic features* in stylometry; syntax leads to significant improvements over bag-of-words models on all three tasks, achieving 10% to 25% relative error reduction. We give a detailed analysis of which words and syntax most predict a particular attribute, and we show a strong correlation between our predictions and a paper's number of citations.

1 Introduction

Stylometry aims to recover useful attributes of documents from the style of the writing. In some domains, statistical techniques have successfully deduced author identity (Mosteller and Wallace, 1984), gender (Koppel et al., 2003), native language (Koppel et al., 2005), and even whether an author has dementia (Le et al., 2011). Stylometric analysis is important to marketers, analysts and social scientists because it provides demographic data directly from raw text. There has been growing interest in applying stylometry to the content generated by users of Internet applications, e.g., detecting author ethnicity in social media (Eisenstein et al., 2011; Rao et al., 2011), or whether someone is writing deceptive online reviews (Ott et al., 2011).

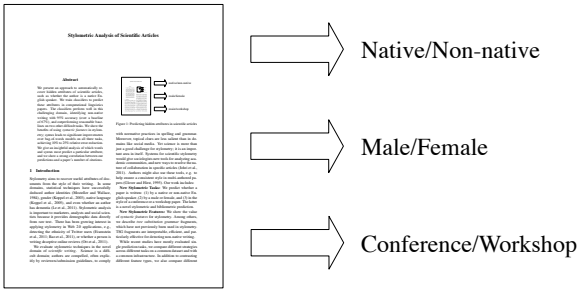


Figure 1: Predicting hidden attributes in scientific articles

We evaluate stylometric techniques in the novel domain of *scientific writing*. Science is a difficult domain; authors are encouraged, often explicitly by reviewers/submission-guidelines, to comply with normative practices in style, spelling and grammar. Moreover, topical clues are less salient than in domains like social media. Success in this challenging domain can bring us closer to correctly analyzing the huge volumes of online text that are currently unmarked for useful author attributes such as gender and native-language.

Yet science is more than just a good stepping-stone for stylometry; it is an important area in itself. Systems for scientific stylometry would give sociologists new tools for analyzing academic communities, and new ways to resolve the nature of collaboration in specific articles (Johri et al., 2011). Authors might also use these tools, e.g., to help ensure a consistent style in multi-authored papers (Glover and Hirst, 1995), or to determine sections of a paper needing revision.

The contributions of our paper include:

New Stylometric Tasks: We predict whether a paper is written: (1) by a native or non-native speaker, (2) by a male or female, and (3) in the style of a conference or workshop paper. The latter is a fully novel stylometric and bibliometric prediction.

New Stylometric Features: We show the value of *syntactic features* for stylometry. Among others, we describe *tree substitution grammar* fragments, which have not previously been used in stylometry. TSG fragments are interpretable, efficient, and particularly effective for detecting non-native writing.

While recent studies have mostly evaluated single prediction tasks, we compare different strategies across different tasks on a common dataset and with a common infrastructure. In addition to contrasting different feature types, we compare different *training strategies*, exploring ways to make use of training instances with label uncertainty.

We also provide a detailed *analysis* that is interesting from a sociolinguistic standpoint. Precisely what words distinguish non-native writing? How does the syntax of female authors differ from males? What are the hallmarks of top-tier papers? Finally, we identify some strong correlations between our predictions and a paper's citation count, even when controlling for paper venue and origin.

2 Related Work

Bibliometrics is the empirical analysis of scholarly literature; *citation analysis* is a well-known bibliometric approach for ranking authors and papers (Borgman and Furner, 2001). Bibliometry and stylometry can share goals but differ in techniques. For example, in a work questioning the blindness of double-blind reviewing, Hill and Provost (2003) predict author identities. They ignore the article body and instead consider (a) potential self-citations and (b) similarity between the article's citation list and the citation lists of known papers. Radev et al. (2009a) perform a bibliometric analysis of computational linguistics. Teufel and Moens (2002) and Qazvinian and Radev (2008) summarize scientific articles, the latter by automatically finding and filtering sentences in other papers that cite the target article.

Our system does not consider citations; it is most similar to work that uses raw article text. Hall et al. (2008) build per-year topic models over scientific literature to track the evolution of scientific ideas. Gerrish and Blei (2010) assess the influence of individual articles by modeling their impact on the content of future papers. Yogatama et al. (2011) predict whether a paper will be cited based on both its content and its meta-data such as author names and publication venues. Johri et al. (2011) use per-author topic models to assess the nature of collaboration in a particular article (e.g., *apprenticeship* or *synergy*). One of the tasks in Sarawgi et al. (2011) concerned predicting gender in scientific writing, but they use a corpus of only ten "highly established" authors and make the prediction using twenty papers for each. Finally, Dale and Kilgarriff (2010) initiated a shared task on automatic editing of scientific papers written by non-native speakers, with the objective of developing "tools which can help non-native speakers of English (NNSs) (and maybe some native ones) write academic English prose of the kind that helps a paper get accepted."

Lexical and pragmatic choices in academic writing have also been analyzed within the applied linguistics community (Myers, 1989; Vassileva, 1998).

3 ACL Dataset and Preprocessing

We use papers from the ACL Anthology Network (Radev et al., 2009b, Release 2011) and exploit its manually-curated meta-data such as normalized author names, affiliations (including country, available up to 2009), and citation counts. We convert each PDF to text[1] but remove text before the *Abstract* (to anonymize) and after the *Acknowledgments/References* headings. We split the text into sentences[2] and filter any documents with fewer than 100 (this removes some short/demo papers, malconverted PDFs, etc. – about 23% of the 13K papers with affiliation information). In case the text was garbled, we then filtered the first 3 lines from every file and any line with an '@' symbol (which might be part of an affiliation). We remove footers like *Proceedings of ...*, table/figure captions, and any lines with non-ASCII characters (e.g. math equations). Papers are then parsed via the Berke-

[1] Via the open-source utility `pdftotext`

[2] Splitter from `cogcomp.cs.illinois.edu/page/tools`

Task	Training Set: *Strict*	*Lenient*	Dev Set	Test Set
NativeL	2127	3963	450	477
Venue	2484	3991	400	421
Gender	2125	3497	400	409

Table 1: Number of documents for each task

ley parser (Petrov et al., 2006), and part-of-speech (PoS) tagged using CRFTagger (Phan, 2006).

Training sets always comprise papers from 2001-2007, while test sets are created by randomly shuffling the 2008-2009 portion and then dividing it into development/test sets. We also use papers from 1990-2000 for experiments in §7.3 and §7.4.

4 Stylometric Tasks

Each task has both a *Strict* training set, using only the data for which we are most confident in the labels (as described below), and a *Lenient* set, which forcibly assigns every paper in the training period to some class (Table 1). All test papers are annotated using a *Strict* rule. While our approaches for automatically-assigning labels can be coarse, they allow us to scale our analysis to a realistic cross-section of academic papers, letting us discover some interesting trends.

4.1 *NativeL*: Native vs. Non-Native English

We introduce the task of predicting whether a scientific paper is written by a native English speaker (NES) or non-native speaker (NNS). Prior work has mostly made this prediction in learner corpora (Koppel et al., 2005; Tsur and Rappoport, 2007; Wong and Dras, 2011), although there have been attempts in elicited speech transcripts (Tomokiyo and Jones, 2001) and e-mail (Estival et al., 2007). There has also been a large body of work on *correcting* errors in non-native writing, with a specific focus on difficulties in preposition and article usage (Han et al., 2006; Chodorow et al., 2007; Felice and Pulman, 2007; Tetreault and Chodorow, 2008; Gamon, 2010).

We annotate papers using two pieces of associated meta-data: (1) author first names and (2) countries of affiliation. We manually marked each country for whether English is predominantly spoken there. We then built a list of common first names of English speakers via the top 150 male and female names from the U.S. census.[3] If the *first* author of a paper has an English first name *and* English-speaking-country affiliation, we mark as NES.[4] If *none* of the authors have an English first name *nor* an English-speaking-country affiliation, we mark as NNS. We use this rule to label our development and test data, as well as our *Strict* training set. For *Lenient* training, we decide based solely on whether the first author is from an English-speaking country.

4.2 *Venue*: Top-Tier vs. Workshop

This novel task aims to distinguish top-tier papers from those at workshops, based on style. We use the annual meeting of the ACL as our canonical top-tier venue. For evaluation and *Strict* training, we label all main-session ACL papers as *top-tier*, and all workshop papers as *workshop*. For *Lenient* training, we assign **all** conferences (LREC, Coling, EMNLP, etc.) to be *top-tier* except for their non-main-session papers, which we label as *workshop*.

4.3 *Gender*: Male vs. Female

Because we are classifying an international set of authors, U.S. census names (the usual source of gender ground-truth) provide incomplete information. We therefore use the data of Bergsma and Lin (2006).[5] This data has been widely used in coreference resolution but never in stylometry. Each line in the data lists how often a noun co-occurs with male, female, neutral and plural pronouns; this is commonly taken as an approximation of the true gender distribution. E.g., '*bill clinton*' is 98% male (in 8344 instances) while '*elsie wayne*' is 100% female (in 23). The data also has *aggregate counts* over all nouns with the same first token, e.g., '*elsie* ...' is 94% female (in 255 instances). For *Strict* training/evaluation, we label papers with the following rule based on the first author's first name:

[3] `www.census.gov/genealogy/names/names_files.html` We also manually added common nicknames for these, e.g. *Rob* for *Robert*, *Chris* for *Christopher*, *Dan* for *Daniel*, etc.

[4] Of course, assuming the first author writes each paper is imperfect. In fact, for some native/non-native collaborations, our system ultimately predicts the 2nd (non-native) author to be the main writer; in one case we confirmed the accuracy of this prediction by personal communication with the authors.

[5] `www.clsp.jhu.edu/~sbergsma/Gender/`

if the name has an aggregate count >30 and female probability >0.85, label as female; otherwise if the aggregate count is >30 and male probability >0.85, label male. This rule captures many of ACL's unambiguously-gendered names, both male (*Nathanael, Jens, Hiroyuki*) and female (*Widad, Yael, Sunita*). For *Lenient* training, we assign all papers based only on whether the male or female probability for the first author is higher. While potentially noisy, there is precedent for assigning a single gender to papers "co-authored by researchers of mixed gender" (Sarawgi et al., 2011).

5 Models and Training Strategies

Model: We take a discriminative approach to stylometry, representing articles as feature vectors (§6) and classifying them using a linear, L2-regularized SVM, trained via LIBLINEAR (Fan et al., 2008). SVMs are state-of-the-art and have been used previously in stylometry (Koppel et al., 2005).

Strategy: We test whether it's better to train with a smaller, more accurate *Strict* set, or a larger but noisier *Lenient* set. We also explore a third strategy, motivated by work in learning from noisy web images (Bergamo and Torresani, 2010), in which we fix the *Strict* labels, but also include the remaining examples as *unlabeled* instances. We then optimize a *Transductive* SVM, solving an optimization problem where we not only choose the feature weights, but also labels for unlabeled training points. Like a regular SVM, the goal is to maximize the margin between the positive and negative vectors, but now the vectors have both fixed and imputed labels. We optimize using Joachims (1999)'s software. While the classifier is trained using a transductive strategy, it is still tested *inductively*, i.e., on unseen data.

6 Stylometric Features

Koppel et al. (2003) describes a range of features that have been used in stylometry, ranging from early manual selection of potentially discriminative words, to approaches based on automated text categorization (Sebastiani, 2002). We use the following three feature classes; the particular features were chosen based on development experiments.

6.1 *Bow* Features

A variety of "discouraging results" in the text categorization literature have shown that simple bag-of-words (*Bow*) representations usually perform better than "more sophisticated" ones (e.g. using syntax) (Sebastiani, 2002). This was also observed in sentiment classification (Pang et al., 2002). One key aim of our research is to see whether this is true of scientific stylometry. Our *Bow* representation uses a feature for each unique lower-case word-type in an article. We also preprocess papers by making all digits '0'. Normalizing digits and filtering capitalized words helps ensure citations and named-entities are excluded from our features. The feature value is the log-count of how often the corresponding word occurs in the document.

6.2 *Style* Features

While text categorization relies on keywords, stylometry focuses on topic-independent measures like function word frequency (Mosteller and Wallace, 1984), sentence length (Yule, 1939), and PoS (Hirst and Feiguina, 2007). We define a *style-word* to be: (1) punctuation, (2) a stopword, or (3) a Latin abbreviation.[6] We create *Style* features for all unigrams and bigrams, replacing non-*style-words* separately with both PoS-tags and spelling signatures.[7] Each feature is an N-gram, the value is its log-count in the article. We also include stylistic *meta-features* such as mean-words-per-sentence and mean-word-length.

6.3 *Syntax* Features

Unlike recent work using generative PCFGs (Raghavan et al., 2010; Sarawgi et al., 2011), we use syntax directly as features in *discriminative* models, which can easily incorporate arbitrary and overlapping syntactic clues. For example, we will see that one indicator of native text is the use of certain determiners as stand-alone noun phrases (NPs), like *this* in Figure 2. This contrasts with a proposed non-native phrase, "this/DT growing/VBG area/NN," where *this* instead modifies a noun. The *Bow* features are clearly unhelpful: *this* occurs in both cases. The

[6] The stopword list is the standard set of 524 SMART-system stopwords (following Tomokiyo and Jones (2001)). Latin abbreviations are *i.e., e.g., etc., c.f., et* or *al.*

[7] E.g., signature '*LC-ing*' means lower-case, ending in *ing*. These are created via a script included with the Berkeley parser.

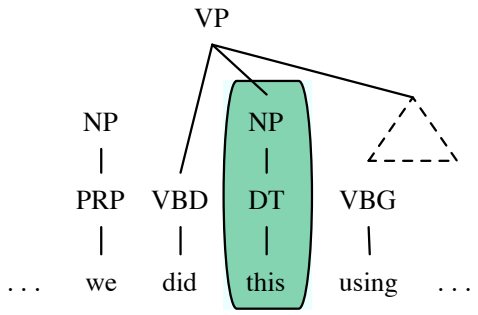


Figure 2: Motivating deeper syntactic features: The shaded TSG fragment indicates native English, but is not directly encoded in *Bow*, *Style*, nor standard CFG-rules.

Style features are likewise unhelpful; *this*-VBG also occurs in both cases. We need the deeper knowledge that a specific determiner is used as a complete NP.

We evaluate three feature types that aim to capture such knowledge. In each case, we aggregate the feature counts over all the parse trees constituting a document. The feature value is the log-count of how often each feature occurs. To remove *content* information from the features, we preprocess the parse tree terminals: all non-*style-word* terminals are replaced with their spelling signature (see §6.2).

CFG Rules: We include a feature for every unique, single-level context-free-grammar (CFG) rule application in a paper (following Baayen et al. (1996), Gamon (2004), Hirst and Feiguina (2007), Wong and Dras (2011)). The Figure 2 tree would have features: NP→PRP, NP→DT, DT→*this*, etc. Such features do capture that a determiner was used as an NP, but they do not jointly encode *which* determiner was used. This is an important omission; we'll see that other determiners acting as stand-alone NPs indicate *non-native* writing (e.g., the word *that*, see §7.2).

TSG Fragments: A tree-substitution grammar is a generalization of CFGs that allow rewriting to tree fragments rather than sequences of non-terminals (Joshi and Schabes, 1997). Figure 2 gives the example NP→(DT *this*). This fragment captures both the identity of the determiner and its syntactic function as an NP, as desired. Efficient Bayesian procedures have recently been developed that enable the training of large-scale probabilistic TSG grammars (Post and Gildea, 2009; Cohn et al., 2010).

While TSGs have not been used previously in stylometry, Post (2011) uses them to predict sentence *grammaticality* (i.e. detecting pseudo-sentences following Okanohara and Tsujii (2007) and Cherry and Quirk (2008)). We use Post's TSG training settings and his public code.[8] We parse with the TSG grammar and extract the fragments as features. We also follow Post by having features for aggregate TSG statistics, e.g., how many fragments are of a given size, tree-depth, etc. These syntactic meta-features are somewhat similar to the manually-defined stylometric features of Stamatatos et al. (2001).

C&J Reranking Features: We also extracted the reranking features of Charniak and Johnson (2005). These features were hand-crafted for reranking the output of a parser, but have recently been used for other NLP tasks (Post, 2011; Wong and Dras, 2011). They include lexicalized features for sub-trees and head-to-head dependencies, and aggregate features for conjunct parallelism and the degree of right-branching. We get the features using another script from Post.[9] While TSG fragments tile a parse tree into a few useful fragments, C&J features can produce thousands of features per sentence, and are thus much more computationally-demanding.

7 Experiments and Results

We take the *minority class* as the positive class: NES for *NativeL*, top-tier for *Venue* and female for *Gender*, and calculate the precision/recall of these classes. We tune three hyperparameters for F1-score on development data: (1) the SVM regularization parameter, (2) the threshold for classifying an instance as positive (using the signed hyperplane-distance as the score), and (3) for transductive training (§5), the fraction of unlabeled data to label as positive. Statistical significance on held-out test data is assessed with McNemar's test, $p<0.05$. For F1-score, we use the following reasonable *Baseline*: we label all instances with the label of the minority class (achieving 100% recall but low precision).

7.1 Selection of Syntax and Training Strategy

Development experiments showed that using all features, *Bow+Style+Syntax*, works best on all tasks, but there was no benefit in combining different

[8] http://github.com/mjpost/dptsg

[9] http://github.com/mjpost/extract-spfeatures.

Syntax	Strategy	*NativeL*	*Venue*	*Gender*
Baseline		50.5	45.0	28.7
CFG	*Strict*	93.5	59.9	**42.5**
CFG	*Lenient*	89.9	64.9	39.5
TSG	*Strict*	**93.6**	60.7	40.0
TSG	*Lenient*	90.9	64.4	39.1
C&J	*Strict*	90.5	62.3	37.1
C&J	*Lenient*	86.2	**65.2**	39.0

Table 2: F1 scores for *Bow+Style+Syntax* system on *development data*: The best training strategy and the best syntactic features depend on the task.

Features	*NativeL*	*Venue*	*Gender*
Baseline	49.8	45.5	33.1
Bow	88.8	60.7	42.5
Style	90.6	61.9	39.8
Syntax	88.7	64.6	41.2
Bow+Style	90.4	64.0	45.1
Bow+Syntax	90.3	65.8	42.9
Style+Syntax	89.4	65.5	43.3
Bow+Style+Syntax	**91.6**	**66.7**	**48.2**

Table 3: F1 scores with different features on *held-out test data*: Including style and syntactic features is superior to standard *Bow* features in all cases.

Syntax features. We also found no gain from transductive training, but greater cost, with more hyperparameter tuning and a slower SVM solver. The best *Syntax* features depend on the task (Table 2). Whether *Strict* or *Lenient* training: TSG was best for *NativeL*, C&J was best for *Venue*, and CFG was best for *Gender*. These trends continue on test data, where TSG exceeds CFG (91.6% vs. 91.2%). For the training strategy, *Strict* was best on *NativeL* and *Gender*, while *Lenient* was best on *Venue* (Table 2). This latter result is interesting: recall that for *Venue*, *Lenient* training considers all conferences to be top-tier, but evaluation is just on detecting ACL papers. We suggest some reasons for this below, highlighting some general features of conference papers that extend beyond particular venues.

For the remainder of experiments on each task, we fix the syntactic features and training strategy to those that performed best on development data.

7.2 Test Results and Feature Analysis

Gender remains the most difficult task on *test* data, but our F1 still substantially outperforms the baseline (Table 3). Results on *NativeL* are particularly impressive; in terms of *accuracy*, we classify 94.6% of test articles correctly (the majority-class baseline is 66.9%). Regarding features, just using *Style+Syntax* always works better than using *Bow*. Combining all features always works better still. The gains of *Bow+Style+Syntax* over vanilla *Bow* are statistically significant in each case.

We also highlight important *individual features*:

***NativeL*:** Table 4 gives *Bow* and *Style* features for *NativeL*. Some reflect differences in common native/non-native *topics*; e.g., '*probabilities*' predicts native while '*morphological*' predicts non-native. Several features, like '*obtained*', indicate L1 interference; i.e., many non-natives have a cognate for *obtain* in their native language and thus adopt the English word. As an example, the word *obtained* occurs 3.7 times per paper from Spanish-speaking areas (cognate *obtenir*) versus once per native paper and 0.8 times per German-authored paper.

Natives also prefer certain abbreviations (e.g. '*e.g.*') while non-natives prefer others ('*i.e.*', '*c.f.*', '*etc.*'). Exotic punctuation also suggests native text: the semi-colon, exclamation and question mark all predict NES. Note this also varies by region; semi-colons are most popular in NES countries but papers from Israel and Italy are close behind.

Table 5 gives highly-weighted TSG features for predicting *NativeL*. Note the determiner-as-NP usage described earlier (§ 6.3): *these*, *this* and *each* predict native when used as an NP; *that*-as-an-NP predicts non-native. Furthermore, while not all native speakers use a comma before a conjunction in a list, it's nevertheless a good flag for native writing ('NP→NP, NP**,** (CC *and*) NP'). In terms of non-native syntax, the passive voice is more common ('VP→(VBZ *is*) VP' and 'VP→VBN (PP (IN *as*) NP)'). We also looked for features involving determiners since correct determiner usage is a common difficulty for non-native speakers. We found cases where determiners were missing where natives might have used one ('NP→JJ JJ NN'), but also those where a determiner might be optional and skipped by a native speaker ('NP→(DT *the*) NN NNS'). Note that Table 5

Predicts native		Predicts non-native	
Bow feature	Wt.	*Bow* feature	Wt.
initial	2.25	obtained	-2.15
techniques	2.11	proposed	-2.06
probabilities	1.38	method	-2.06
additional	1.23	morphological	-1.96
fewer	1.02	languages	-1.23
Style feature	Wt.	*Style* feature	Wt.
used to	1.92	, i.e.	-2.60
JJR NN	1.90	have to	-1.65
has VBN	1.90	the *xxxx*-ing	-1.61
example ,	1.75	thus	-1.61
all of	1.73	usually	-1.24
's	1.69	mainly	-1.21
allow	1.47	, because	-1.12
has *xxxx*-ed	1.45	the VBN	-1.12
may be	1.35	JJ for	-1.11
; and	1.21	cf	-0.97
e.g.	1.10	etc.	-0.55
must VB	0.99	associated to	-0.23

Table 4: *NativeL*: Examples of highly-weighted style and content features in the *Bow+Style+Syntax* system.

TSG Fragment	Example
Predicts native English author:	
NP→NNP CD	(*Model*) (*1*)
NP→(DT *these*)	*six of* (*these*)
NP→(DT *that*) NN	*in* (*that*) (*language*)
NP→(DT *this*)	*we did* (*this*) *using ...*
VP→(VBN *used*) S	(*used*) (*to describe it*)
NP→NP, NP, (CC *and*) NP	(*X*), (*Y*), (*and*) (*Z*)
NP→(DT *each*)	(*each*) consists of ...
Predicts non-native English author:	
VP→(VBZ *is*) VP	*it* (*is*) (*shown below*)
VP→VBN (PP (IN *as*) NP)	(*considered*) (*as*) (*a term*)
NP→JJ JJ NN	*in* (*other*) (*large*) (*corpus*)
NP→DT JJ (CD *one*)	(*a*) (*correct*) (*one*)
NP→(DT *the*) NN NNS	*seen in* (*the*) (*test*) (*data*)
NP→(DT *that*)	*larger than* (*that*) *of ...*
QP→(IN *about*) CD	(*about*) (*200,000*) *words*

Table 5: *NativeL*: Highly-weighted syntactic features (descending order of absolute weight) and examples in the *Bow+Style+Syntax* system.

examples are based on actual usage in ACL papers. We also found that *complex* NPs were more associated with native text. Features such as 'NP→DT JJ NN NN NN', and 'NP→DT NN NN NNS' predict native writing.

Non-natives also rely more on boilerplate. For example, the exact phrase "The/This paper is organized as follows" occurs 3 times as often in non-native compared to native text (in 7.5% of all non-native papers). Sentence re-use is only indirectly captured by our features; it would be interesting to encode flags for it directly.

In general, we found very few highly-weighted features that pinpoint 'ungrammatical' non-native writing (the feature '*associated to*' in Table 4 is a rare example). Our classifiers largely detect non-native writing on a stylistic rather than grammatical basis.

***Venue*:** Table 6 provides important *Bow* and *Style* features for the *Venue* task (syntactic features omitted due to space). While some features are topical (e.g. '*biomedical*'), the table gives a blueprint for writing a solid main-conference paper. That is, good papers often have an explicit probability model (or algorithm), experimental baselines, error analysis, and statistical significance checking. On the other hand, there might be a bias at main conferences for focused, incremental papers; features of workshop papers highlight the exploration of '*interesting*' new ideas/domains. Here, the objective might only be to show what is '*possible*' or what one is '*able to*' do. Main conference papers prefer work that improves '*performance*' by '*#%*' on established tasks.

***Gender*:** The CFG features for *Gender* are given in Table 7. Several of the most highly-weighted female features include pronouns (e.g. PRP$). A higher frequency of pronouns in female writing has been attested previously (Argamon et al., 2003), but has not been traced to particular syntactic constructions. Likewise, we observe a higher frequency of not just negation (noted previously) but adverbs (RB) in general (e.g. 'VP→MD RB VP'). In terms of *Bow* features (not shown), the words *contrast* and *comparison* highly predict female, as do topical clues like *verb* and *resource*. The top-three male *Bow* features are (in order): *simply*, *perform*, *parsing*.

7.3 Author Rankings

While our objective is to predict attributes of *papers*, we also show how that we can identify *author* attributes using a larger body of work. We make *NativeL* and *Gender* predictions for all papers in the

Predicts ACL		Predicts Workshop	
Bow feature	Wt.	*Bow* feature	Wt.
model	2.64	semantic	-2.16
probability	1.66	analysis	-1.65
performance	1.40	verb	-1.35
baseline	1.36	lexical	-1.33
=	1.26	study	-0.92
algorithm	1.18	biomedical	-0.87
large	1.16	preliminary	-0.69
error	1.15	interesting	-0.69
outperforms	1.02	aim	-0.64
significant	0.96	manually	-0.62
statistically	0.75	appears	-0.54
Style feature	Wt.	*Style* feature	Wt.
by VBG	1.04	able to	-0.99
#%	0.82	*xxxx*-ed out	-0.77
NN over	0.79	further NN	-0.71
than the	0.79	NN should	-0.69
improvement	0.75	will be	-0.61
best	0.71	possible	-0.57
xxxx-s by	0.70	have not	-0.56
much JJR	0.67	currently	-0.56

Table 6: *Venue*: Examples of highly-weighted style content features in the *Bow+Style+Syntax* system.

1990-2000 era using our *Bow+Style+Syntax* system. For each author+affiliation with $\geq$3 first-authored papers, we take the average classifier score on these papers.

Table 8 shows cases where our model strongly predicts native, showing top authors with foreign affiliations and top authors in English-speaking countries.[10] While not perfect, the predictions correctly identify some native authors that would be difficult to detect using only name and location data. For example, *Dekai Wu* (Hong Kong) speaks English natively; *Christer Samuelsson* lists near-native English on his C.V.; etc. Likewise, we have also been able to accurately identify a set of *non-native* speakers with common American names that were working at American universities.

Table 9 provides some of the extreme predictions of our system on *Gender*. The extreme male and female predictions are based on both style and content; females tend to work on summarization, discourse, etc., while many males focus on parsing. We also tried making these lists without *Bow* features, but the extreme examples still reflect topic to some extent. Topics themselves have their own style, which the style features capture; it is difficult to fully separate style from topic.

[10]Note again that this is based on the affiliation of these authors during the 1990s; e.g. Gerald Penn published three papers while at the University of Tübingen.

CFG Rule	Example
Predicts female author:	
NP→PRP$ NN NN	(*our*) (*upper*) (*bound*)
QP→RB CD	(*roughly*) (*6000*)
NP→NP, CC NP	(*a new* NE *tag*), (*or*) (*no* NE *tag*)
NP→PRP$ JJ JJ NN	(*our*) (*first*) (*new*) (*approach*)
VP→MD RB VP	(*may*) (*not*) (*be useful*)
ADVP→RB RBR	(*significantly*) (*more*)
Predicts male author:	
ADVP→RB RB	(*only*) (*superficially*)
NP→NP, SBAR	*we use* (*XYZ*), (*which is ...*)
S→S: S.	(*Trust me*): (*I'm a doctor*)
S→S, NP VP	(*To do so*), (*it*) (*needs help*)
WHNP→WP NN	*depending on* (*what*) (*path*) *is ...*
PP→IN PRN	(*in*) ((*Jelinek, 1976*))

Table 7: *Gender*: Highly-weighted syntactic features (descending order of weight) and examples in the *Bow+Style+Syntax* system.

Highest NES Scores, non-English-country: *Gerald Penn,*[10] *Ezra W. Black, Nigel Collier, Jean-Luc Gauvain, Dan Cristea, Graham J. Russell, Kenneth R. Beesley, Dekai Wu, Christer Samuelsson, Raquel Martinez*
Highest NES Scores, English-country: *Eric V. Siegel, Lance A. Ramshaw, Stephanie Seneff, Victor W. Zue, Joshua Goodman, Patti J. Price, Stuart M. Shieber, Jean Carletta, Lynn Lambert, Gina-Anne Levow*

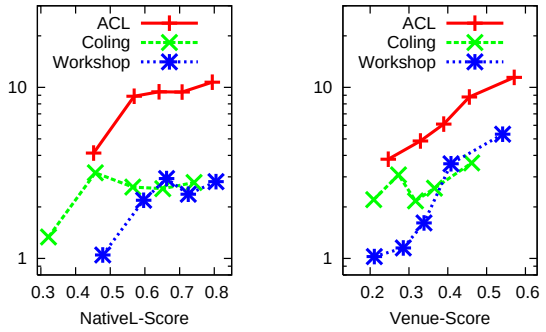
Table 8: Authors scoring highest on *NativeL*, in descending order, based exclusively on article text.

7.4 Correlation with Citations

We also test whether our systems' *stylometric* scores correlate with the most common *bibliometric* measure: citation count. To reduce the impact of *topic*, we only use *Style+Syntax* features. We plot results separately for *ACL*, *Coling* and *Workshop* papers (1990-2000 era). Papers at each venue are sorted by their classifier scores and binned into five score bins. Each point in the plot is the mean-score/mean-number-of-citations for papers in a bin (within-community citation data is via the AAN §3

Highest Model Scores (Male): *John Aberdeen, Chao-Huang Chang, Giorgio Satta, Stanley F. Chen, GuoDong Zhou, Carl Weir, Akira Ushioda, Hideki Tanaka, Koichi Takeda, Douglas B. Paul, Hideo Watanabe, Adam L. Berger, Kevin Knight, Jason M. Eisner*
Highest Model Scores (Female): *Julia B. Hirschberg, Johanna D. Moore, Judy L. Delin, Paola Merlo, Rebecca J. Passonneau, Bonnie Lynn Webber, Beth M. Sundheim, Jennifer Chu-Carroll, Ching-Long Yeh, Mary Ellen Okurowski, Erik-Jan Van Der Linden*

Table 9: Authors scoring highest (absolute values) on *Gender*, in descending order, based exclusively on article text.



(a) (b)

Figure 3: Correlation between predictions (x-axis) and mean number of citations (y-axis, *log-scale*).

and excludes self citations). We use a truncated mean for citation counts, leaving off the top/bottom five papers in each bin.

For *NativeL*, we only plot papers marked as **native** by our *Strict* rule (i.e. English name/country). Papers with the lowest *NativeL*-scores receive many fewer citations, but they soon level off (Figure 3(a)). Many *junior* researchers at English universities are non-native speakers; early-career non-natives might receive fewer citations than well-known peers. The correlation between citations and *Venue*-scores is even stronger (Figure 3(b)); the top-ranked workshop papers receive five times as many citations as the lowest ones, and are cited better than a good portion of ACL papers. These figures suggest that citation-predictors can get useful information beyond typical *Bow* features (Yogatama et al., 2011). Although we focused on a past era, stylistic/syntactic features should also be more robust to the evolution of scientific topics; we plan to next test whether we can better *forecast* future citations. It would also be interesting to see whether these trends transfer to other academic disciplines.

7.5 Further Experiments on *NativeL*

For *NativeL*, we also created a special test corpus of 273 papers written by first-time ACL authors (2008-2009 era). This set closely aligns with the system's potential use as a tool to help new authors compose papers. Two (native-speaking) annotators manually annotated each paper for whether it was primarily written by a native or non-native speaker (considering both content and author names/affiliations). The annotators agreed on 90% of decisions, with an inter-annotator kappa of 66%. We divided the papers into a test set and a development set. We applied our *Bow+Style+Syntax* system exactly as trained above, except we tuned its hyperparameters on the new development data. The system performed quite well on this set, reaching 68% F1 over a baseline of only 27%. Moreover, the system also reached 90% accuracy, matching the level of human agreement.

8 Conclusion

We have proposed, developed and successfully evaluated significant new tasks and methods in the stylometric analysis of scientific articles, including the novel resolution of publication venue based on paper style, and novel syntactic features based on tree substitution grammar fragments. In all cases, our syntactic and stylistic features significantly improve over a bag-of-words baseline, achieving 10% to 25% relative error reduction in all three major tasks. We have included a detailed analysis of discriminative stylometric features, and we showed a strong correlation between our predictions and a paper's number of citations. We observed evidence for L1-interference in non-native writing, for differences in topic between males and females, and for distinctive language usage which can successfully identify papers published in top-tier conferences versus wokrshop proceedings. We believe that this work can stimulate new research at the intersection of computational linguistics and bibliometrics.

References

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text*, 23(3), August.

Harald Baayen, Fiona Tweedie, and Hans van Halteren. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132.

Alessandro Bergamo and Lorenzo Torresani. 2010. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Proc. NIPS*, pages 181–189.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proc. Coling-ACL*, pages 33–40.

Christine L. Borgman and Jonathan Furner. 2001. Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, 36:3–72.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proc. ACL*, pages 173–180.

Colin Cherry and Chris Quirk. 2008. Discriminative, syntactic language modeling through latent SVMs. In *Proc. AMTA*.

Martin Chodorow, Joel R. Tetreault, and Na-Rae Han. 2007. Detection of grammatical errors involving prepositions. In *Proc. ACL-SIGSEM Workshop on Prepositions*, pages 25–30.

Trevor Cohn, Phil Blunsom, and Sharon Goldwater. 2010. Inducing tree-substitution grammars. *J. Mach. Learn. Res.*, 11:3053–3096.

Robert Dale and Adam Kilgarriff. 2010. Helping our own: Text massaging for computational linguistics as a new shared task. In *Proc. 6th International Natural Language Generation Conference*, pages 261–265.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proc. ACL*, pages 1365–1374.

Dominique Estival, Tanja Gaustad, Son-Bao Pham, Will Radford, and Ben Hutchinson. 2007. Author profiling for English emails. In *Proc. PACLING*, pages 263–272.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874.

Rachele De Felice and Stephen G. Pulman. 2007. Automatically acquiring models of preposition use. In *Proc. ACL-SIGSEM Workshop on Prepositions*, pages 45–50.

Michael Gamon. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In *Proc. Coling*, pages 611–617.

Michael Gamon. 2010. Using mostly native data to correct errors in learners' writing: a meta-classifier approach. In *Proc. HLT-NAACL*, pages 163–171.

Sean Gerrish and David M. Blei. 2010. A language-based approach to measuring scholarly impact. In *Proc. ICML*, pages 375–382.

Angela Glover and Graeme Hirst. 1995. Detecting stylistic inconsistencies in collaborative writing. In *Writers at work: Professional writing in the computerized environment*, pages 147–168.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proc. EMNLP*, pages 363–371.

Na-Rae Han, Martin Chodorow, and Claudia Leacock. 2006. Detecting errors in English article usage by non-native speakers. *Nat. Lang. Eng.*, 12(2):115–129.

Shawndra Hill and Foster Provost. 2003. The myth of the double-blind review?: Author identification using only citations. *SIGKDD Explor. Newsl.*, 5:179–184.

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

Thorsten Joachims. 1999. Transductive inference for text classification using support vector machines. In *Proc. ICML*, pages 200–209.

Nikhil Johri, Daniel Ramage, Daniel McFarland, and Daniel Jurafsky. 2011. A study of academic collaborations in computational linguistics using a latent mixture of authors model. In *Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 124–132.

Aravind K. Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages: Beyond Words*, volume 3, pages 71–122.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2003. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proc. KDD*, pages 624–628.

Xuan Le, Ian Lancashire, Graeme Hirst, and Regina Jokel. 2011. Longitudinal detection of dementia through lexical and syntactic changes in writing: A case study of three British novelists. *Literary and Linguistic Computing*, 26(4):435–461.

Frederick Mosteller and David L. Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag.

Greg Myers. 1989. The pragmatics of politeness in scientific articles. *Applied Linguistics*, 10(1):1–35.

Daisuke Okanohara and Jun'ichi Tsujii. 2007. A discriminative language model with pseudo-negative samples. In *Proc. ACL*, pages 73–80.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proc. ACL*, pages 309–319.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proc. EMNLP*, pages 79–86.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proc. Coling-ACL*, pages 433–440.

Xuan-Hieu Phan. 2006. CRFTagger: CRF English POS Tagger. `crftagger.sourceforge.net`.

Matt Post and Daniel Gildea. 2009. Bayesian learning of a tree substitution grammar. In *Proc. ACL-IJCNLP*, pages 45–48.

Matt Post. 2011. Judging grammaticality with tree substitution grammar derivations. In *Proc. ACL*, pages 217–222.

Vahed Qazvinian and Dragomir R. Radev. 2008. Scientific paper summarization using citation summary networks. In *Proc. Coling*, pages 689–696.

Dragomir R. Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2009a. A bibliometric and network analysis of the field of computational linguistics. *Journal of the American Society for Information Science and Technology*.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009b. The ACL anthology network corpus. In *Proc. ACL Workshop on Natural Language Processing and Information Retrieval for Digital Libraries*, pages 54–61.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proc. ACL*, pages 38–42.

Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Proc. ICWSM*, pages 598–601.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proc. CoNLL*, pages 78–86.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34:1–47.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. 2001. Automatic text categorization in terms of genre and author. *Computational Linguistics*, 26(4):471–495.

Joel R. Tetreault and Martin Chodorow. 2008. The ups and downs of preposition error detection in ESL writing. In *Proc. Coling*, pages 865–872.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–445.

Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from 'round here, are you? Naive Bayes detection of non-native utterances. In *Proc. NAACL*.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proc. Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16.

Irena Vassileva. 1998. Who am I/who are we in academic writing? *International Journal of Applied Linguistics*, 8(2):163–185.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proc. EMNLP*, pages 1600–1610.

Dani Yogatama, Michael Heilman, Brendan O'Connor, Chris Dyer, Bryan R. Routledge, and Noah A. Smith. 2011. Predicting a scientific community's response to an article. In *Proc. EMNLP*, pages 594–604.

G. Udny Yule. 1939. On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship. *Biometrika*, 30(3/4):363–390.