Characterizing Stylistic Elements in Syntactic Structure

Song Feng **Ritwik Banerjee** **Yejin Choi**
Department of Computer Science
Stony Brook University
NY 11794, USA
songfeng, rbanerjee, ychoi@cs.stonybrook.edu

Abstract

Much of the writing styles recognized in rhetorical and composition theories involve deep syntactic elements. However, most previous research for *computational* stylometric analysis has relied on shallow lexico-syntactic patterns. Some very recent work has shown that PCFG models can detect distributional difference in syntactic styles, but without offering much insights into exactly what constitute salient stylistic elements in sentence structure characterizing each authorship. In this paper, we present a comprehensive exploration of syntactic elements in writing styles, with particular emphasis on *interpretable characterization* of stylistic elements. We present analytic insights with respect to the authorship attribution task in two different domains.

1 Introduction

Much of the writing styles recognized in rhetorical and composition theories involve deep syntactic elements in style (e.g., Bain (1887), Kemper (1987) Strunk and White (2008)). However, previous research for automatic authorship attribution and computational stylometric analysis have relied mostly on shallow lexico-syntactic patterns (e.g., Mendenhall (1887), Mosteller and Wallace (1984), Stamatatos et al. (2001), Baayen et al. (2002), Koppel and Schler (2003), Zhao and Zobel (2007), Luyckx and Daelemans (2008)).

Some very recent works have shown that PCFG models can detect distributional difference in sentence structure in gender attribution (Sarawgi et al., 2011), authorship attribution (Raghavan et al., 2010), and native language identification (Wong and Dras, 2011). However, still very little has been understood exactly what constitutes salient stylistic elements in sentence structures that characterize each author. Although the work of Wong and Dras (2011) has extracted production rules with highest information gain, their analysis stops short of providing insight any deeper than what simple n-gram-level analysis could also provide.[1] One might even wonder whether PCFG models are hinging mostly on leaf production rules, and whether there are indeed deep syntactic differences at all. This paper attempts to answer these questions.

As an example of syntactic stylistic elements that have been much discussed in rhetorical theories, but have not been analyzed computationally, let us consider two contrasting sentence styles: *loose (cumulative)* and *periodic*:[2] a *loose* sentence places the main clause at the beginning, and then appends subordinate phrases and clauses to develop the main message. In contrast, a *periodic* sentence starts with subordinate phrases and clauses, suspending the most

[1] For instance, missing determiners in English text written by Chinese speakers, or simple n-gram anomaly such as frequent use of "according to" by Chinese speakers (Wong and Dras, 2011).

[2] *Periodic* sentences were favored in classical times, while *loose* sentences became more popular in the modern age.

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1522–1533, Jeju Island, Korea, 12–14 July 2012.

HOBBS	JOSHI	LIN	MCDON
S^ROOT → S , CC S	PP^PRN → IN NP	NP^S → NN CD	NP^NP → DT NN POS
NP^PP → DT	NP^PP → NP PRN SBAR	NP^NP → DT NN NNS	WHNP^SBAR → IN
VP^VP → TO VP	S^ROOT → PP NP VP .	S^ROOT → SBAR , NP VP .	NP^PP → NP SBAR
PP^PP → IN S	PRN^NP → -LRB- PP -RRB-	NP^PP → NP : NP	SBAR^PP → WHADVP S
NP^PP → NP , PP	NP^NP → NNP	S^ROOT → PP , NP VP .	SBAR^S → WHNP S
VP^S → VBZ ADJP S	S^SBAR → PP NP VP	NP^NP → PDT DT NNS	PP^NP → IN SBAR
VP^SINV → VBZ	S^ROOT → LST NP VP .	NP^VP → DT NN SBAR	SBAR^NP → WHNP S
VP^S → VBD S	CONJP^NP → RB RB IN	SBAR^S → WHADVP S	SBAR^PP → SBAR CC SBAR
VP^S → VBG PP	NP^PP → NP PRN PP	PRN^NP → -LRB- NP -RRB-	PP^VP → IN
ADVP^VP → RB PP	NP^NP → NP , NP	NP^PP → NN NN	S^SBAR → VP

Table 1: Top 10 most discriminative production rules for each author in the scientific domain.

LOOSE	*Christopher Columbus finally reached the shores of San Salvador* after months of uncertainty at sea, the threat of mutiny, and a shortage of food and water.
PERIODIC	After months of uncertainty at sea, the threat of mutiny, and a shortage of food and water, *Christopher Columbus finally reached the shores of San Salvador.*

Table 2: Loose/Periodic sentence with identical set of words and POS tags

important part to the end. The example in Table 2 highlights the difference:
Notice that these two sentences comprise of an identical set of words and part-of-speech. Hence, shallow lexico-syntactic analysis will not be able to catch the pronounced stylistic difference that is clear to a human reader.

One might wonder whether we could gain interesting insights simply by looking at the most discriminative production rules in PCFG trees. To address this question, Table 1 shows the top ten most discriminative production rules for authorship attribution for scientific articles,[3] ranked by LIBLINEAR (Fan et al., 2008).[4] Note that terminal production rules are excluded so as to focus directly on syntax.

It does provide some insights, but not to a satisfactory degree. For instance, Hobbs seems to favor inverted declarative sentences (SINV) and adverbs with prepositions (RB PP). While the latter can be easily obtained by simple part-of-speech analysis, the former requires using parse trees. We can also observe that none of the top 10 most discriminative production rules for Hobbs includes SBAR tag, which represents subordinate clauses. But examining discriminative rules alone is limited in providing more comprehensive characterization of idiolects.

Can we unveil something more in deep syntactic structure that can characterize the collective syntactic difference between any two authors? For instance, what can we say about distributional difference between loose and periodic sentences discussed earlier for each author? As can be seen in Table 1, simply enumerating most discriminative rules does not readily answer questions such as above.

In general, production rules in CFGs do not directly map to a wide variety of stylistic elements in rhetorical and composition theories. This is only as expected however, partly because CFGs are not designed for stylometric analysis in the first place, and also because some syntactic elements can go beyond the scope of context free grammars.

As an attempt to reduce this gap between modern statistical parsers and cognitively recognizable stylistic elements, we explore two complementary approaches:

1. Translating some of the well known stylistic elements of rhetorical theories into PCFG analysis (Section 3).

2. Investigating different strategies of analyzing PCFG trees to extract author characteristics that are interesting as well as *interpretable* (Sections 4 & 5).

[3]See Section 2 for the description of the dataset.

[4]We use Berkeley PCFG parser (Petrov and Klein, 2007) for all experiments.

Algorithm 1 Sentence Type-1 Identification

Input: Parse tree $t(N_r)$ of sentence s
Output: Type of s.

```
if S ∈ L^top then
  if SBAR ∉ Ω(N_r) then
    return COMPOUND
  else
    return COMPLEX-COMPOUND
else
  if VP ∈ L^top then
    if SBAR ∉ Ω(N_r) then
      return SIMPLE
    else
      return COMPLEX
return OTHER
```

We present analytic insights with respect to the authorship attribution task in two distinct domains.

2 Data

For the empirical analysis of authorship attribution, we use two different datasets described below. Sections 3, 4 & 5 provide the details of our stylometric analysis.

Scientific Paper We use the ACL Anthology Reference Corpus (Bird et al., 2008). Since it is nearly impossible to determine the gold-standard authorship of a paper written by multiple authors, we select 10 authors who have published at least 8 single-authored papers. We include 8 documents per author, and remove citations, tables, formulas from the text using simple heuristics.[5]

Novels We collect 5 novels from 5 English authors: Charles Dickens, Edward Bulwer-Lytton, Jane Austen, Thomas Hardy and Walter Scott. We select the first 3000 sentences from each novel and group every 50 consecutive sentences into 60 documents per novel per author.

[5]Some might question whether the size of the dataset used here is relatively small in comparison to typical dataset comprised of thousands of documents in conventional text categorization. We point out that authorship attribution is fundamentally different from text categorization in that it is often practically impossible to collect more than several documents for each author. Therefore, it is desirable that the attribution algorithms to detect the authors based on very small samples.

Algorithm 2 Sentence Type-II Identification

Input: Parse tree $t(N_r)$ of sentence s
Output: Type of s.

```
k ← 1
while k ≤ λ do
  if L_k^top ≠ VP then
    if S ∈ Ω(L_k^top) or SBAR ∈ Ω(L_k^top) then
      return PERIODIC
  else
    if S ∈ Ω(L_k^top) or SBAR ∈ Ω(L_k^top) then
      return LOOSE
return OTHER
```

3 Sentence Types

In this section, we examine well-known sentence types that are recognized in the literature, but have not been analyzed computationally.

Type-I Identification – Simple/Complex/ Compound/Complex-Compound: PCFG trees do not provide this information directly, hence we must construct an algorithm to derive it. The key to identifying these sentences is the existence of *dependent* and *independent* clauses. For the former, we rely on the SBAR tag, while for the latter, we first define the sequence of nodes right below the root (e.g., [NP VP .] shown in the horizontal box in Figure 1). We call this the *top structural level*. We then check whether S (in addition to the root S) appears in this sequence.

Formally, let $L^{top} = \{N_i\}$ be the set of nodes in the top structural level, and $\lambda = |L^{top}|$. Let $t(N_r)$ be the tree rooted at N_r, and $\Omega(N_r)$ denote the set of nodes in $t(N_r)$. Algorithm 1 shows the procedure to determine the type-I class of a sentence based on its PCFG tree.[6]

Type-II Identification – Loose/Periodic: A sentence can also be classified as loose or periodic, and we present Algorithm 2 for this identification. We perform a mini-evaluation on 20 previously unseen sentences for each type[7]. Our algorithm was able to perform type-I identification on all sentences correctly. In type-II

[6]Note that Algorithm 1 & 2 rely on the use of Berkeley parser (Petrov and Klein, 2007).

[7]These were gathered from several online quizzes for English learners. E.g., `http://grammar.about.com`, `http://a4esl.org`

Type	Hobbs	Joshi	Lin	McDon
simple	40.0	41.7	50.2	27.9
cplex	40.8	40.7	37.6	48.4
cpnd	7.9	5.6	3.9	5.5
cpxnd	8.5	9.2	7.7	15.5
other	2.8	2.8	0.6	2.7
loose	27.6	26.4	26.9	30.8
perio	11.1	11.7	15.2	16.4
other	61.3	61.9	57.9	52.8

Table 3: Sentence Types (%) in scientific data.

Type	Dickens	B-Lyt	Austen	Hardy	Scott
simple	26.0	21.2	23.9	25.6	17.5
cplex	24.4	21.8	24.8	25.6	31.8
cpnd	15.3	15.2	12.6	16.3	11.7
cpxnd	20.8	23.3	31.1	18.9	28.7
other	13.5	18.5	7.6	13.6	10.3
loose	11.5	10.8	17.9	14.5	15.3
perio	19.5	13.6	14.0	16.2	18.0
other	69.0	75.6	68.1	69.3	66.7

Table 4: Sentence Types (%) in Novels

identification, it labeled all loose sentences correctly, and achieved 90% accuracy on periodic sentences.

Discussion Tables 3 & 4 show the sentence type distribution in scientific data and novels, respectively.[8] We see that different authors are characterized by different distribution of sentence types. For instance, in Table 3, Lin is a prolific user of simple sentences while McDon prefers employing complex sentences. McDon also uses complex-compound sentences quite often (15.5%), more than twice as frequently as Lin. Notice that all authors use loose sentences much more often than periodic sentences, a known trend in modern English.

In Table 4, we see the opposite trend among 19^{th}-century novels: with the exception of Jane Austen, all authors utilize periodic sentences comparatively more often. We also notice that complex and complex-compound sentences abound, as expected from classic literary proses.

Can we determine authorship solely based on the distribution of sentence types?

We experiment with a SVM classifier using just 6 features (one feature for each sentence type in Table 3), and we achieve accuracy 36.0% with the scientific data. Given that a random baseline would achieve only about 10% accuracy, this demonstrates that the distribution of sentence types does characterize an idiolect to some degree.

[8]Due to space limitation, we present analyses based on 4 authors from the scientific data.

4 Syntactic Elements Based on Production Rules

In this section, we examine three different aspects of syntactic elements based on production rules.

4.1 Syntactic Variations

We conjecture that the *variety* of syntactic structure, which most previous research in computational stylometry has not paid much attention to, provides an interesting insight into authorship. One way to quantify the degree of syntactic variations is to count the unique production rules. In Tables 5, we show the extent of syntactic variations employed by authors using the standard deviation σ and the *coverage* of an author:

$$\mathcal{C}(a) := \frac{|\mathcal{R}(a)|}{|\cup_a \mathcal{R}(a)|} \times 100$$

where $\mathcal{R}(a)$ denotes the set of unique production rules used by author a, and $\cup_a$ iterates over all authors. In order to compare among authors, we also show these parameters normalized with respect to the highest value. Our default setting is to exclude all lexicalized rules in the productions to focus directly on the syntactic variations. In our experiments (Section 6), however, we do augment the rules with (a) ancestor nodes to capture deeper syntactic structure and (b) lexical (*leaf*) nodes.

As hypothesized, these statistics provide us new insights into the authorship. For instance, we find that McDon employs a wider variety of syntactic structure than others, while Lin's writing exhibits relatively the least variation. Moreover, comparing Joshi and Hobbs, it is interesting to see the standard deviation differ a lot

	Hobbs	Joshi	Lin	McDon	Dickens	B-Lyt	Austen	Hardy	Scott
$\mathcal{C}$	36.0	37.6	32.8	42.6	30.9	28.8	36.2	30.0	24.1
$\mathcal{C}_{\text{norm}}$	0.84	0.88	0.77	1.0	0.85	0.79	1.0	0.83	0.67
σ	51.5	39.2	63.3	44.4	88.3	81.6	98.0	125.3	114.7
σ_{norm}	0.81	0.62	1.0	0.7	0.7	0.65	0.78	1.0	0.92

Table 5: Syntactic variations of different authors in the scientific domain.

Hobbs	Joshi	Lin	McDon
# 136	# 142	# 124	# 161
S → S CC S .	S → ADVP PP NP VP .	S → SBAR NP VP .	S → S NP VP .
S → CC NP VP .	S → PP NP ADVP VP .	FRAG → NP : S .	S → S : S .
S → S VP .	S → NP VP	S → NP VP .	S → SBAR VP .
S → NP NP VP .	S → S S CC S .	S → PP VP .	S → SBAR S CC S .
S → PP NP VP .	S → ADVP NP VP .	S → NP ADVP VP .	S → NP PP VP .

Table 6: Most discriminative sentence outlines in the scientific data. $\#N$ shows the number of unique sentence outlines of each author.

(51.5 and 39.2), in spite of their $\mathcal{C}$ scores being similar: 36.0% and 37.6%, respectively. This indicates that Hobbs tends to use a certain subset production rules much more frequently than Joshi. Lin exhibits the highest standard deviation in spite of having least syntactic variation, indicating that he uses a much smaller subset of productions regularly, while ocassionally deviating to other rules.

Similarly, among novels, Jane Austen's writing has the highest amount of variation, while Walter Scott's writing style is the least varied. Even though authors from both datasets display similar $\mathcal{C}$ scores (Table 5), the difference in σ is noteworthy. The significantly higher linguistic variation is to be expected in creative writing of such stature. It is interesting to note that the authors with highest coverage – Austen and Dickens – have much lower deviation in their syntactic structure when compared to Hardy and Scott. This indicates that while Austen and Dickens consistently employ a wider variety of sentence structures in their writing, Hardy and Scott follow a relatively more uniform style with sporadic forays into diverse syntactic constructs.

4.2 Sentence Outlines

Although the approach of Section 4.1 give us a better and more general insight into the characteristics of each author, its ability to provide insight on *deep* syntactic structure is still limited, as it covers production rules at all levels of the tree. We thus shift our focus to the top level of the trees, e.g., the second level (marked in a horizontal box) in Tree (1) of Figure 1, which gives us a better sense of sentence outlines.

Tables 6 and 7 present the most discriminative sentence outlines of each author in the scientific data and novels, respectively. We find that McDon is a prolific user of subordinate clauses, indicating his bias towards using complex sentences. The rule "S → SBAR S CC S" shows his inclination towards complex-compound sentences as well. These inferences are further supported by the observations in Table 3. Another observation of possible interest is the tendency of Joshi and Lin to begin sentences with prepositional phrases.

In comparing Table 6 and Table 7, notice the significantly higher presence of complex and compound-complex structures in the latter[9]. The most discriminating sentence outlines for Jane Austen, for instance, are all indicative of complex-compound sentences. This is further supported by Table 4.

5 Syntactic Elements Based on Tree Topology

In this section, we investigate quantitative techniques to capture stylistic elements in the tree

[9]The presence of "FRAG" is not surprising. Intentional use of verbless sentence fragments, known as *scesis onomaton*, was often employed by authors such as Dickens and Bulwer-Lytton (Quinn, 1995).

Dickens	Bulwer-Lytton	Austen	Hardy	Scott
# 1820	# 1696	# 2137	# 1772	# 1423
SQ → NNP .	SBARQ → WHNP S .	S → S : CC S .	S → S NP VP .	S → NP PRN VP .
FRAG → NP .	FRAG → INTJ NP .	S → S CC S : CC S .	S → ADVP NP VP .	S → PP NP VP .
SINV → NP VP NP .	S → S : S CC S .	S → S : CC S : CC S .	S → FRAG : S .	S → S S : S .
INTJ → UH .	FRAG → CC NP .	S → S : S : CC S .	S → INTJ NP VP .	S → NP PP VP .
SBARQ → WHNP SQ .	FRAG → NP ADJP .	S → SBAR S : CC S .	S → NP VP .	S → ADVP PRN NP VP .

Table 7: Most discriminative sentence outlines in the novel data. $\#N$ shows the number of unique sentence outlines of each author.

Metrics	Scientific Data				Novels				
	Hobbs	Joshi	Lin	McDon	Dickens	B-Lyt	Austen	Hardy	Scott
sen-len $\overline{avg}$	23.7	26.0	21.0	32.2	24.1	26.7	31.4	21.5	34.1
$h^{\mathcal{T}}{}_{\overline{avg}}$	5.8	5.3	5.9	4.8	4.7	5.0	5.4	4.9	5.9
$h^{\mathcal{F}}{}_{\overline{avg}}$	2.4	2.1	2.5	1.9	1.9	1.9	2.1	1.9	2.1
$w^{L}{}_{\overline{avg}}$	5.0	4.8	5.5	4.2	4.1	4.4	4.7	3.8	4.9
$\sigma^{H}{}_{\overline{avg}}$	1.2	1.1	1.1	1.0	1.1	1.1	1.3	1.2	1.4
$\sigma^{S}{}_{\overline{avg}}$	1.9	1.8	1.8	1.7	1.0	1.1	1.2	1.0	1.4

Table 8: Tree topology metrics for scientific data and novels.

topology. Figure 1 shows three different parse trees to accompany our discussion.[10] Notice that sentence (1) is a loose sentence, and sentence (2) is periodic. In general, loose sentences grow deep and unbalanced, while periodic sentences are relatively more balanced and wider.

For a tree t rooted at N_R with a height n, let $\mathcal{T}$ be the set of leaf nodes, and let $\mathcal{F}$ be the set of furcation nodes, and let $\xi(N_i, N_j)$ denote the length of the shortest path from N_i to N_j. Inspired by the work of Shao (1990), we analyze tree topology with the following four measurements:

- Leaf height ($h^{\mathcal{T}} = \{h_i^{\mathcal{T}}, N_i \in \mathcal{T}\}$), where $h_i^{\mathcal{T}} = \xi(N_i, N_R)\ N_i \in \mathcal{T}$. For instance, the leaf height of "free" of Tree (2) in Fig. 1 is 6.
- Furcation height ($h^{\mathcal{F}} = \{h_i^{\mathcal{F}}, N_i \in \mathcal{F}\}$), where $h_i^{\mathcal{F}}$ is the maximum leaf height within the subtree rooted at N_i. In Figure 1, for example, the furcation height of the VP in Tree (2) (marked in triangle) is 3.
- Level width ($w^L = \{w_l, 1 \leq l \leq n\}$), where $w_l = |\{N_i : \xi(N_i, N_R) = l\}|$. E.g., w_4 of Tree (1) in Figure 1 is 6.
- Horizontal $\sigma^H = \{\sigma_i^H, N_i \in \mathcal{F}\}$, and Vertical Imbalance $\sigma^S = \{\sigma_i^S, N_i \in \mathcal{F}\}$. Let $\mathcal{C}$ be the set of child nodes of N_k. If $|\mathcal{C}| \geq 2$, then

$$\sigma_k^H = \sqrt{\frac{1}{n}\sum_{i=1}^{|\mathcal{C}|}(h_i^{\mathcal{F}} - H)^2}$$

where $H = \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} h_i^{\mathcal{F}}$. Similarly,

$$\sigma_k^S = \sqrt{\frac{1}{n}\sum_{i=1}^{|\mathcal{C}|}(s(N_i) - S)^2}$$

where $S = \frac{1}{|\mathcal{C}|}\sum_{i=1}^{|\mathcal{C}|} s(N_i)$ and $s(N_i)$ is the number of leaf nodes of tree rooted at N_i. As shown in Figure 1, the imbalance of the internal node VP in Tree (2) (marked in triangle) is 0.5 horizontally, and 0.5 vertically.

To give an intuition on the relation between these measurements and different tree structures, Table 9 provides the measurements of the three trees shown in Figure 1.

Note that all three sentences are of similar length but show different tree structures. Tree (1) and Tree (2) differ in that Tree (1) is highly unbalanced and grows deep, while Tree

[10] Example sentences are taken from Lin (1997), Joshi (1992), and Lin (1995).

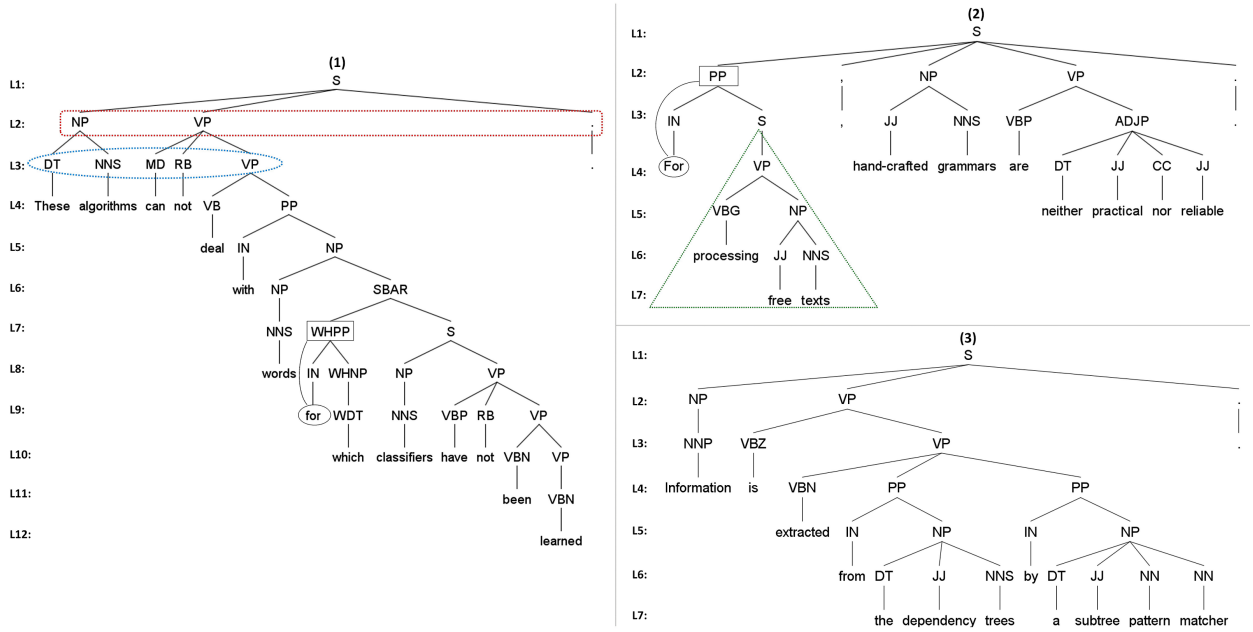

Figure 1: Parsed trees

Metrics	Tree (1)	Tree (2)	Tree (3)
# of tokens	15	13	13
$\max_i \{h_i^{\mathcal{T}}\}$	11	6	6
$\max_i \{w_i^L\}$	6	9	9
$\max_i \{\sigma_i^H\}$	4.97	1.6	1.7
$\max_i \{\sigma_i^S\}$	4	1.5	4.7

Table 9: Tree Topology Statistics for Figure 1.

(2) is much better balanced and grows shorter but wider. Comparing Tree (2) and Tree (3), they have the same max Leaf height, Level width, and Horizontal Imbalance, but the latter has bigger Vertical Imbalance, which quantifies the imbalance in terms of the text span covered by subtrees.

We provide these topological metrics for authors from both datasets in Table 8.

6 Experiments & Evaluation

In our experiments, we utilize a set of features motivated by PCFG trees. These consist of simple production rules and other syntactic features based on tree-traversals. Table 10 describes these features with examples from Tree (2), using the portion marked by the triangle.

These sets of production rules and syntax features are used to build SVM classifiers using LIBLINEAR (Fan et al., 2008), wherein all feature values are encoded as term-frequencies normalized by document size. We run 5-fold cross-validation with training and testing split first as 80%/20%, and then as 20%/80%.

We would like to point out that the latter configuration is of high practical importance in authorship attribution, since we may not always have sufficient training data in realistic situations, e.g., forensics (Luyckx and Daelemans, 2008).

Lexical tokens provide strong clues by creating features that are specific to each author: research topics in the scientific data, and proper nouns such as character names in novels. To lessen such topical bias, we lemmatize and rank words according to their frequency (in the entire dataset), and then consider the top 2,000 words only. Leaf-node productions with words outside this set are disregarded.

Our experimental results (Tables 11 & 12) show that not only do deep syntactic features perform well on their own, but they also significantly improve over lexical features. We also show that adding the style_{11} features further improves performance.

Features	
pr	Rules *excluding* terminal productions. E.g., **VP** → VBG NP
syn_v	Traversal from a non-leaf node to its grandparent (embedded rising). E.g., **VP**ˆS → PP
syn_h	Left-to-right traversal in the set of all non-leaf children of a node. E.g., VBG → NP (for node **VP**)
syn_{v+h}	$syn_v \cup syn_h$
syn_0	No tree traversal. Feature comprises interior nodes only.
$syn_\downarrow$	Union of all edges to child nodes, except when child is a leaf node. E.g., {**VP** → VBG, **VP** → NP}
$syn_\updownarrow$	$syn_\downarrow \cup$ {edge to parent node}
STYLE$_{11}$	The set of 11 extra stylistic features. 6 values from the distribution of sentence types (Section 3), and 5 topological metrics (Section 5) characterizing the height, width and imbalance of a tree.

Variations	
$\hat{pr}$	Each production rule is augmented with the grandparent node.
$*$	Terminal (*leaf*) nodes are included.

Table 10: Features and their lexico-syntactic variations. *Illustration*: $\hat{pr}^*$ denotes the set of production rules pr (including terminal productions) that are augmented with their grandparent nodes.

Features	Scientific		Novels	
		+STYLE$_{11}$		+STYLE$_{11}$
STYLE$_{11}$	20.6	–	43.1	–
Unigram	56.9	–	69.3	–
syn_h	53.7	53.7	68.3	67.9
syn_0	22.9	**31.1**	57.8	**62.5**
$syn_\downarrow$	43.4	**44.0**	63.6	**65.7**
$syn_\updownarrow$	51.1	**51.7**	71.3	**72.8**
syn_{v+h}	54.0	**55.7**	72.0	**73.2**
syn_h^*	63.1	**64.0**	72,1	**73.2**
syn_0^*	56.6	56.0	73.1	**74.1**
$syn_\downarrow^*$	56.3	**57.2**	74.0	**74.9**
$syn_\updownarrow^*$	64.6	**65.4**	**74.9**	**75.3**
syn_{v+h}^*	64.0	**67.7**	74.0	**74.7**
pr	50.3	**53.4**	67.0	66.7
$\hat{pr}$	59.1	**60.6**	69.7	68.7
$pr*$	63.7	**65.1**	71.5	**73.2**
$\hat{pr}*$	66.3	**69.4**	73.6	**74.9**

Table 11: Authorship attribution with **20% training data**. Improvement with addition of STYLE$_{11}$ shown in bold.

To quantify the amount of authorship information carried in the set STYLE$_{11}$, we experiment with a SVM classifier using only 11 features (one for each metric), and achieve accuracy of 42.0% and 52.0% with scientific data and novels, respectively. Given that a random-guess baseline would achieve only 10% and 20% (resp.), and that the classification is based on just 11 features, this experiment demonstrates how effectively the tree topology statistics capture idiolects. In general, lexicalized features yield higher performance even after removing topical words. This is expected since tokens such as function words play an important role in determining authorship (e.g., Mosteller and Wallace (1984), Garcia and Martin (2007), Argamon et al. (2007)).

A more important observation, however, is that even after removing the leaf production rules, accuracy as high as 93% (scientific) and 92.2% (novels) are obtained using syntactic features, which demonstrates that there are syntactic patterns unique to each author. Also notice that using only production rules, we achieve higher accuracy in novels (90.1%), but the addition of STYLE$_{11}$ features yields better results with scientific data (93.0%).

Using different amounts of training data provides insight about the influence of lexical clues. In the scientific dataset, increasing the amount of training data decreases the average performance difference between lexicalized and unlexicalized features: 13.5% to 11.6%. In novels, however, we see the opposite trend: 6.1% increases to 8.1%.

We further observe that with scientific data, increasing the amount of training data improves the average performance across all unlexicalized feature-sets from 50.0% to 82.9%, an improvement of 32.8%. For novels, the corresponding improvement is small in comparison: 17.0%.

This difference is expected. While authors such as Dickens or Hardy have their unique writing styles that a classifier can learn based on few documents, capturing idiolects in the more rigid domain of scientific writing is far from obvious with little training data.

Features	Scientific		Novels	
		+style$_{11}$		+style$_{11}$
style$_{11}$	42.0	–	52.0	–
Unigram	88.0	–	92.7	–
syn_h	85.0	85.0	87.6	**88.9**
syn_0	40.0	**53.0**	66.4	**72.3**
$syn_\downarrow$	78.0	**82.0**	80.3	**82.3**
$syn_\updownarrow$	85.0	**92.0**	89.3	**92.2**
syn_{v+h}	89.0	**93.0**	90.1	**91.2**
syn^*_h	93.0	93.0	93.7	**93.9**
syn^*_0	92.0	**94.0**	92.1	**93.2**
$syn^*_\downarrow$	93.0	**94.0**	93.4	**94.5**
$syn^*_\updownarrow$	93.0	**95.0**	94.9	**95.2**
syn^*_{v+h}	94.0	**96.0**	94.7	**94.8**
pr	85.0	**86.0**	86.7	86.7
$\hat{pr}$	87.0	**89.0**	88.2	**89.3**
$pr*$	93.0	**94.0**	92.1	**93.2**
$\hat{pr}*$	94.0	**95.0**	94.5	**95.1**

Table 12: Authorship attribution with **80% training data**.

Turning to lexicalized features, we note that with more training data, lexical cues perform better in scientific domain than in novels. With 80% data used for training, the average performance of lexicalized feature-sets with science data is 94.4%, and slightly lower at 94.3% for novels. With less training data, however, these figures are 63.5% and 74.3% respectively.

Finally, we point out that adding the style features derived from sentence types and tree topologies almost always improves the performance. In scientific data, syn^*_{v+h} with style$_{11}$ features shows the best performance (96%), while $syn^*_\updownarrow$ yields the best results for novels (95.2%). For unlexicalized features, adding style$_{11}$ to syn_{v+h} and $syn_\updownarrow$ yields respective improvements of 4.0% and 2.9% in the two datasets.

7 Related Work

There are several hurdles in authorship attribution. First and foremost, writing style is extremely domain-dependent. Much of previous research has focused on several domains of writing, such as informal modern writing in blogs and online messages (Zheng et al., 2006), relatively formal contemporary texts such as news articles (Raghavan et al., 2010), or classical literature like novels and proses (e.g., (Burrows, 2002), (Hoover, 2004)).

The nature of these features have also varied considerably. Character level n-grams have been used by several researchers; most notably by Peng et al. (2003), by Houvardas and Stamatatos (2006) for feature selection, and by Stamatatos (2006) in ensemble learning. Keselj et al. (2003) employed frequency measures on n-grams for authorship attribution.

Others, such as Zhao and Zobel (2005), Argamon and Levitan (2004), Garcia and Martin (2007), have used word-level approaches instead, incorporating the differential use of function words by authors.

More sophisticated linguistic cues have been explored as well: parts-of-speech n-grams (Diederich et al., 2003), word-level statistics together with POS-sequences (Luyckx and Daelemans, 2008), syntactic labels from partial parsing (Hirst and Feiguina, 2007), etc. The use of syntactic features from parse trees in authorship attribution was initiated by Baayen et al. (1996), and more recently, Raghavan et al. (2010) have directly employed PCFG language models in this area.

Syntactic features from PCFG parse trees have also been used for gender attribution (Sarawgi et al., 2011), genre identification (Stamatatos et al., 2000), native language identification (Wong and Dras, 2011) and readability assessment (Pitler and Nenkova, 2008). The primary focus of most previous research, however, was to attain better classification accuracy, rather than providing linguistic interpretations of individual authorship and their stylistic elements.

Our work is the first to attempt authorship attribution of scientific papers, a contemporary domain where language is very formal, and the stylistic variations have limited scope. In addition to exploring this new domain, we also present a comparative study expounding the role of syntactic features for authorship attribution in classical literature. Furthermore, our work is also the first to utilize tree topological

features (Chan et al., 2010) in the context of stylometric analysis.

8 Conclusion

In this paper, we have presented a comprehensive exploration of syntactic elements in writing styles, with particular emphasis on interpretable characterization of stylistic elements, thus distinguishing our work from other recent work on syntactic stylometric analysis. Our analytical study provides novel statistically supported insights into stylistic elements that have not been computationally analyzed in previous literature. In the future, we plan to investigate the use of syntactic feature generators for text categorization (e.g., Collins and Duffy (2002), Moschitti (2008), Pighin and Moschitti (2009)) for stylometry analysis.

Acknowledgments Yejin Choi is partially supported by the Stony Brook University Office of the Vice President for Research. We thank reviewers for many insightful and helpful comments.

References

Shlomo Argamon and Shlomo Levitan. 2004. Measuring the usefulness of function words for authorship attribution. *Literary and Linguistic Computing*, pages 1–3.

Shlomo Argamon, Casey Whitelaw, Paul Chase, Sobhan Raj Hota, Navendu Garg, and Shlomo Levitan. 2007. Stylistic text classification using functional lexical features: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 58(6):802–822.

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

H. Baayen, H. van Halteren, A. Neijt, and F. Tweedie. 2002. An experiment in authorship attribution. In *6th JADT*. Citeseer.

A. Bain. 1887. *English Composition and Rhetoric: Intellectual elements of style*. D. Appleton and company.

S. Bird, R. Dale, B.J. Dorr, B. Gibson, M.T. Joseph, M.Y. Kan, D. Lee, B. Powley, D.R. Radev, and Y.F. Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proc. of the 6th International Conference on Language Resources and Evaluation Conference (LREC08)*, pages 1755–1759.

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

Samuel W. K. Chan, Lawrence Y. L. Cheung, and Mickey W. C. Chong. 2010. Tree topological features for unlexicalized parsing. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 117–125, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 263–270, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Diederich, J. Kindermann, E. Leopold, and G. Paass. 2003. Authorship attribution with support vector machines. *Applied Intelligence*, 19(1):109–123.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Antonion Miranda Garcia and Javier Calle Martin. 2007. Function words in authorship attribution studies. *Literary and Linguistic Computing*, 22(1):49–66.

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

D. L. Hoover. 2004. Testing burrow's delta. *Literary and Linguistic Computing*, 19(4):453–475.

J. Houvardas and E. Stamatatos. 2006. N-gram feature selection for author identification. In *Proc. of the 12^{th} International Conference on Artificial Intelligence: Methodology, Systems and Applications*, volume 4183 of *LNCS*, pages 77–86, Varna, Bulgaria. Springer.

Aravind K. Joshi. 1992. Statistical language modeling. In *Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. Association for Computational Linguistics.

S. Kemper. 1987. Life-span changes in syntactic complexity. *Journal of gerontology*, 42(3):323.

Vlado Keselj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proc. of the*

Pacific Association for Computational Linguistics, pages 255–264.

M. Koppel and J. Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI*, volume 3, pages 69–72. Citeseer.

D. Lin. 1995. University of manitoba: description of the pie system used for muc-6. In *Proceedings of the 6th conference on Message understanding*, pages 113–126. Association for Computational Linguistics.

D. Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 64–71. Association for Computational Linguistics.

Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *COLING '08*, pages 513–520.

T.C. Mendenhall. 1887. The characteristic curves of composition. *Science*, ns-9(214S):237–246.

Alessandro Moschitti. 2008. Kernel methods, syntax and semantics for relational text categorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 253–262, New York, NY, USA. ACM.

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

Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 267–274, Stroudsburg, PA, USA. Association for Computational Linguistics.

S. Petrov and D. Klein. 2007. Improved inference for unlexicalized parsing. In *Proceedings of NAACL HLT 2007*, pages 404–411.

Daniele Pighin and Alessandro Moschitti. 2009. Reverse engineering of tree kernel feature spaces. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 111–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: a unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 186–195, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arthus Quinn. 1995. *Figures of Speech: 60 Ways To Turn A Phrase*. Routledge.

Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 38–42, Uppsala, Sweden. Association for Computational Linguistics.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, CoNLL '11, pages 78–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

K.T. Shao. 1990. Tree balance. *Systematic Biology*, 39(3):266.

Efstathios Stamatatos, George Kokkinakis, and Nikos Fakotakis. 2000. Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26(4):471–495.

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

E. Stamatatos. 2006. Ensemble-based author identification using character n-grams. *ReCALL*, page 4146.

W. Strunk and E.B. White. 2008. *The elements of style*. Penguin Group USA.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1600–1610, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Proceedings of the Second Asia conference on Asia Information Retrieval Technology*, AIRS'05, pages 174–189, Berlin, Heidelberg. Springer-Verlag.

Y. Zhao and J. Zobel. 2007. Searching with style: Authorship attribution in classic literature. In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pages 59–68. Australian Computer Society, Inc.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features

and classification techniques. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):378–393.