Success with Style: Using Writing Style to Predict the Success of Novels

Vikas Ganjigunte Ashok **Song Feng** **Yejin Choi**
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400
vganjiguntea, songfeng, ychoi@cs.stonybrook.edu

Abstract

Predicting the success of literary works is a curious question among publishers and aspiring writers alike. We examine the quantitative connection, if any, between *writing style* and successful literature. Based on novels over several different genres, we probe the predictive power of statistical stylometry in discriminating successful literary works, and identify characteristic stylistic elements that are more prominent in successful writings. Our study reports for the first time that statistical stylometry can be surprisingly effective in discriminating highly successful literature from less successful counterpart, achieving accuracy up to 84%. Closer analyses lead to several new insights into characteristics of the writing style in successful literature, including findings that are contrary to the conventional wisdom with respect to good writing style and readability.

1 Introduction

Predicting the success of novels is a curious question among publishers, professional book reviewers, aspiring and even expert writers alike. There are potentially many influencing factors, some of which concern the intrinsic content and quality of the book, such as interestingness, novelty, style of writing, and engaging storyline, but external factors such as social context and even luck can play a role. As a result, recognizing successful literary work is a hard task even for experts working in the publication industries. Indeed, even some of the best sellers and award winners can go through several rejections before they are picked up by a publisher.[1]

Perhaps due to its obvious complexity of the problem, there has been little previous work that attempts to build statistical models that predict the success of literary works based on their intrinsic content and quality. Some previous studies do touch on the notion of stylistic aspects in successful literature, e.g., extensive studies in Literature discuss literary styles of significant authors (e.g., Ellegård (1962), McGann (1998)), while others consider content characteristics such as plots, characteristics of characters, action, emotion, genre, cast, of the best-selling novels and blockbuster movies (e.g., Harvey (1953), Hall (2012), Yun (2011)).

All these studies however, are qualitative in nature, as they rely on the knowledge and insights of human experts on literature. To our knowledge, no prior work has undertaken a systematic quantitative investigation on the overarching characterization of the writing style in successful literature. In consideration of widely different styles of authorship (e.g., Escalante et al. (2011), Peng et al. (2003), Argamon et al. (2003)), it is not even readily clear whether there might be common stylistic elements that help discriminating highly successful ones from less successful counterpart.

In this work, we present the first study that investigates this unstudied and unexpected connection between stylistic elements and the literary success. The key findings of our research reveal that there exists distinct linguistic patterns shared among suc-

[1] E.g., Paul Harding's "Tinkers" that won 2010 Pulitzer Prize for Fiction and J. K. Rowling's "Harry Potter and the Philosopher's Stone" that sold over 450 million copies.

cessful literature, at least within the same genre, making it possible to build a model with surprisingly high accuracy (up to 84%) in predicting the success of a novel. This result is surprising for two reasons. First, we tackle the hard task of predicting the success of novels written by *previously unseen* authors, avoiding incidental learning of authorship signature, since previous research demonstrated that one can achieve very high accuracy in authorship attribution (as high as 96% in some experimental setup) (e.g., Raghavan et al. (2010), Feng et al. (2012)). Second, we aim to discriminate highly successful novels from less successful, but nonetheless published books written by professional writers, which are undoubtedly of higher quality than average writings. It is important to note that the task we tackle here is much harder than discriminating highly successful works from those that have not even passed the scrutinizing eyes of publishers.

In order to quantify the success of literary works, and to obtain corresponding gold standard labels, one needs to first define *"success"*. For practical convenience, we largely rely on the download counts available at Project Gutenberg as a surrogate to quantify the success of novels. For a small number of novels however, we also consider award recipients (e.g., Pulitzer, Nobel), and Amazon's sales records to define a novel's success. We also extend our empirical study to movie scripts, where we quantify the success of movies based on the average review scores at `imdb.com`. We leave analysis based on other measures of literary success as future research.

In this study, we do not attempt to separate out success based on literary quality (award winners) from success based on popularity (commercial hit, often in spite of bad literary quality), mainly because it is not practically easy to determine whether the high download counts are due to only one reason or the other. We expect that in many cases, the two different aspects of success are likely to coincide, however. In the case of the corpus obtained from Project Gutenberg, where most of our experiments are conducted, we expect that the download counts are more indicative of success based on the literary quality (which then may have resulted in popularity) rather than popularity without quality.

We examine several genres in fiction and movie scripts, e.g., adventure stories, mystery, fiction, historical fiction, sci-fi, short stories, as well as poetry, and present systematic analyses based on lexical and syntactic features which have been known to be effective in a variety of NLP tasks ranging from authorship attribution (e.g., Raghavan et al. (2010)), genre detection (e.g., Rayson et al. (2001), Douglas and Broussard (2000)), gender identification (e.g., Sarawgi et al. (2011)) and native language detection (e.g., Wong and Dras (2011)).

GENRE	#BOOKS	τ^-	τ^+
Adventure	409	17	100
Detective / Mystery	374	25	90
Fiction	1148	7	125
Historical Fiction	374	25	115
Love Stories	342	16	85
Poetry	580	9	70
Science Fiction	902	30	100
Short Stories	1117	9	224

Table 1: # of books available per genre at Gutenberg with download thresholds used to define more successful ($\geq \tau^+$) and less successful ($\leq \tau^-$) classes.

Our empirical results demonstrate that (1) statistical stylometry can be surprisingly effective in discriminating successful literature, achieving accuracy up to 84%, (2) some elements of successful styles are genre-dependent while others are more universal. In addition, this research results in (3) findings that are somewhat contrary to the conventional wisdom with respect to the connection between successful writing styles and readability, (4) interesting correlations between sentiment / connotation and the literary success, and finally, (5) comparative insights between fiction and nonfiction with respect to the successful writing style.

2 Dataset Construction

For our experiments, we procure novels from project Gutenberg[2]. Project Gutenberg houses over $40,000$ books available for free download in electronic format and provides a catalog containing brief descriptions (title, author, genre, language, download count, etc.) of these books. We experiment with genres in Table 1, which have sufficient number of books allowing us to construct reasonably sized datasets.

We use the download counts in Gutenberg-catalog

[2] `http://www.gutenberg.org/`

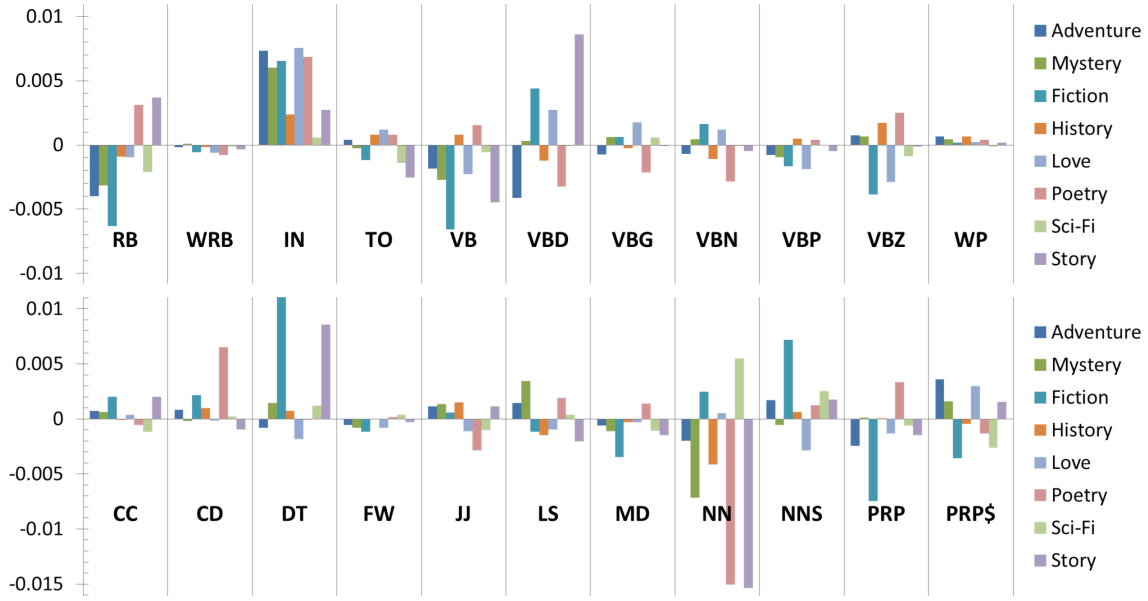

Figure 1: Differences in POS tag distribution between more successful and less successful books across different genres. Negative (positive) value indicates higher percentage in less (more) successful class.

as a surrogate to measure the degree of success of novels. For each genre, we determine a lower bound (τ_+) and an upper bound (τ_-) of download counts as shown in Table 1 to categorize the available books as more successful and less successful respectively. These thresholds are set to obtain at least 50 books for each class, and for each genre. To balance the data, for each genre, we construct a dataset of 100 novels (50 per class).

We make sure that no single author has more than 2 books in the resulting dataset, and in the majority of the cases, only one book has been taken from each author.[3] Furthermore, we make sure that the books from the same author do not show up in both training and test data. These constraints make sure that we learn general linguistic patterns of successful novels, rather than a particular writing style of a few successful authors.

3 Methodology

In what follows, we describe five different aspects of linguistic styles we measure quantitatively. The first three correspond to the features that have been frequently utilized in previous studies in related tasks, e.g., genre detection (e.g., Kessler et al. (1997)) and authorship attribution (e.g., Stamatatos (2009)), while the last two are newly explored in this work.

I. Lexical Choices: unigrams and bigrams.

II. Distribution of Word Categories: Many previous studies have shown that the distribution of part-of-speech (POS) tags alone can reveal surprising insights on genre and authorship (e.g., Koppel and Schler (2003)), hence we examine their distributions with respect to the success of literary works.

III. Distribution of Grammar Rules: Recent studies reported that features based on CFG rules are helpful in authorship attribution (e.g., Raghavan et al. (2010), Feng et al. (2012)). We experiment with four different encodings of production rules:

- Γ: lexicalized production rules (all production rules, including those with terminals)
- Γ^G: lexicalized production rules prepended with the grandparent node.
- γ: unlexicalized production rules (all production rules except those with terminals).
- γ^G: unlexicalized production rules prepended with the grandparent node.

[3]The complete list of novels used for each genre in our dataset is available at `http://www.cs.stonybrook.edu/~ychoi/successwithstyle/`

FEATURE	GENRE								Avg	Avg w/o History
	Adven	Myster	Fiction	Histor	Love	Poetr	Sci-fi	Short		
POS	74.0	63.9	72.0	47.0	65.9	63.0	63.0	67.0	64.5	66.9
Unigram	**84.0**	73.0	**75.0**	60.0	**82.0**	71.0	61.0	57.0	70.3	71.8
Bigram	81.0	73.0	**75.0**	51.0	72.0	70.0	59.0	57.0	67.2	69.5
Γ	73.0	71.0	**75.0**	54.0	78.0	**74.0**	71.0	**77.0**	71.6	74.1
Γ^G	75.0	74.0	**75.0**	58.0	81.0	72.0	**76.0**	77.0	**73.5**	**75.7**
γ	72.0	70.0	65.0	53.0	70.0	66.0	64.0	71	66.3	68.2
γ^G	72.0	69.0	74.0	55.0	75.0	69.0	67.0	73.0	69.2	71.2
Γ+Unigram	79.0	73.0	73.0	59.0	80.0	73.0	71.0	73.0	72.6	74.5
Γ^G+Unigram	80.0	74.0	74.0	56.0	**82.0**	72.0	73.0	72.0	72.8	75.2
γ+Unigram	82.0	72.0	73.0	56.0	81.0	69.0	62.0	59.0	69.2	71.1
γ^G+Unigram	80.0	73.0	74.0	58.0	**82.0**	70.0	65.0	58.0	70	71.7
PHR	74.0	65.0	65.0	56.0	64.0	62.0	69.0	71.0	65.7	67.1
PHR+CLS	75.0	69.0	64.0	**61.0**	59.0	62.0	69.0	67.0	65.7	66.4
PHR+Unigram	80.0	74.0	71.0	56.0	79.0	73.0	67.0	66.0	70.7	72.8
PHR+CLS+Unigram	80.0	**75.0**	71.0	56.0	79.0	73.0	66.0	66.0	70.7	72.8

Table 2: Classification results in accuracy (%).

IV. Distribution of Constituents: PCFG grammar rules are overly specific to draw a big picture on the distribution of large, recursive syntactic units. We hypothesize that the distribution of constituents can serve this purpose, and that it will reveal interesting and more interpretable insights into writing styles in highly successful literature. Despite its relative simplicity, we are not aware of previous work that looks at the distribution of constituents directly. In particular, we are interested in examining the distribution of phrasal and/or clausal tags as follows: (i) Phrasal tag percent (PHR) - percentage distribution of phrasal tags.[4] (ii) Clausal tag percent (CLS) - percentage distribution of clausal tags.

V. Distribution of Sentiment and Connotation: Finally, we examine whether the distribution of sentiment and connotation words, and their polarity, has any correlation with respect to the success of literary works. We are not aware of any previous work that looks into this connection.

4 Prediction Performance

We use LibLinear SVM (Fan et al., 2008) with L2 tuned over training data, and all performance is based on 5-fold cross validation. We take 1000 sentences from the beginning of each book. POS features are encoded as unit normalized frequency and all other features are encoded as tf-idf.[5]

Prediction Results Table 2 shows the classification results. The best performance reaches as high as 84% in accuracy. In fact, in all genres except for history, the best performance is at least 74%, if not higher. Another notable observation is that even in the poetry genre, which is not prose, the accuracy gets as high as 74%. This level of performance is not entirely anticipated, given that (1) the test data consists of books written only by previously unseen authors, and (2) each author has widely different writing style, and (3) we do not have training data at scale, and (4) we aim to tackle the hard task of discriminating highly successful ones from less successful, but nonetheless successful ones, as all of them were, after all, good enough to be published.[6]

Prediction with Varying Thresholds of Download Counts Before we proceed to comprehensive analysis of writing style that are prominent in more successful literature (§5), in Table 3, we present how the prediction accuracy varies as we adjust the definition of more-successful and less-successful literature, by gradually increasing (decreasing) the threshold τ^- (τ^+). As we reduce the gap between τ^- and τ^+, the performance decreases, which shows that indeed there are notable statistical differences in linguistic patterns between novels with high and low download counts, and the stylistic difference monotonically increases (thereby higher prediction accuracy) as we increase the gap between two classes.

[4]The percentage of any phrasal tag is the count of occurrence of that tag over the sum of counts of all phrasal tags.

[5]POS tags are obtained using the Stanford POS tagger (Toutanova and Manning, 2000), and parse trees are based on the Stanford parser (Klein and Manning, 2003).

[6]In our pilot study, we also experimented with the binary classification task of discriminating highly successful ones from those that are not even published (unpublished online novels), and it was a much easier task as expected.

τ^-	τ^+	ACCURACY
17	100	84.0
25	90	78.4
35	80	77.6
45	70	76.4
55	60	73.5

Table 3: Accuracy (%) with varying thresholds of download counts for ADVENTURE with unigram features.

This is particularly interesting as the size of training data set is actually monotonically decreasing (making it harder to achieve high accuracy) while we increase the separation between τ^- and τ^+.

5 Analysis of Successful Writing Styles

5.1 Insights Based on Lexical Choices

It is apparent from Table 2 that unigram features yield curiously high performance in many genres. We therefore examine discriminative unigrams for ADVENTURE, shown in Table 4. Interestingly, less successful books rely on verbs that are explicitly descriptive of actions and emotions (e.g., "wanted", "took", "promised", "cried", "cheered", etc.), while more successful books favor verbs that describe thought-processing (e.g., "recognized", "remembered"), and verbs that serve the purpose of quotes and reports (e.g,. "say"). Also, more successful books use discourse connectives and prepositions more frequently, while less successful books rely more on topical words that could be almost cliché, e.g., "love", typical locations, and involve more extreme (e.g., "breathless") and negative words (e.g., "risk").

5.2 Distribution of Sentiment & Connotation

We also determine the distribution of sentiment and connotation words separately for each class (Table 5) to check if there exists a connection with respect to successful writing styles.[7] We first compare distribution of sentiment and connotation for the entire words. As can be seen in Table 5 – **Top**, there are not notable differences. However, when we compare distribution only with respect to discriminative unigrams only (i.e., features with non-zero weights), as shown in Table 5 – **Bottom**, we find substantial differences in all genres. In particular, discriminative unigrams that characterize less successful novels involve significantly more sentiment-laden words.

[7]We use MPQA subjectivity lexicon (Wilson et al., 2005) and connotation lexicon (Feng et al., 2013) for determining sentiment and connotation of words respectively.

Less Successful	
CATEGORY	UNIGRAMS
Negative	never, risk, worse, slaves, hard, murdered, bruised, heavy, prison,
Body Parts	face, arm, body, skins
Location	room, beach, bay, hills, avenue, boat, door
Emotional / Action Verbs	want, went, took, promise, cry, shout, jump, glare, urge
Extreme Words	never, very, breathless, sacred slightest, absolutely, perfectly
Love Related	desires, affairs
More Successful	
CATEGORY	UNIGRAMS
Negation	not
Report / Quote	said, words, says
Self Reference	I, me , my
Connectives	and, which, though, that, as, after, but, where, what, whom, since, whenever
Prepositions	up, into, out, after, in, within
Thinking Verbs	recognized, remembered

Table 4: Discriminative unigrams for ADVENTURE.

5.3 Distribution of Word Categories

Summarized analysis of POS distribution across all genres is reported in Table 6. It can be seen that prepositions, nouns, pronouns, determiners and adjectives are predictive of highly successful books whereas less successful books are characterized by higher percentage of verbs, adverbs, and foreign words. Per genre distributions of POS tags are visualized in Figure 1. Interestingly, some POS tags show almost universal patterns (e.g., prepositions (IN), NNP, WP, VB), while others are more genre-specific.

In Relation to Journalism Style The work of Douglas and Broussard (2000) reveals that informative writing (journalism) involves increased use of nouns, prepositions, determiners and coordinating conjunctions whereas imaginative writing (novels) involves more use of verbs and adverbs, as has been also confirmed by Rayson et al. (2001). Comparing their findings with Table 6, we find that highly

	Adven		Myster		Fiction		Histor		Love		Poetr		Sci-fi		Short	
	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	+
+ve S	4.7	4.9	4.8	4.6	5.6	4.9	5.0	5.1	5.5	5.1	6.3	5.7	4.1	3.7	4.7	4.8
-ve S	4.0	4.0	4.0	4.0	4.3	4.2	4.2	4.2	4.1	4.2	4.3	4.3	2.9	2.9	3.8	4.0
Tot S	8.7	8.9	8.9	8.7	9.9	9.0	9.2	9.3	9.6	9.3	10.6	9.9	7.0	6.7	8.5	8.9
+ve C	22.3	22.5	22.3	22.5	23.7	23.0	23.0	23.2	23.34	23.3	23.8	22.9	21.2	20.6	22.6	22.7
-ve C	19.4	19.6	19.8	19.8	20.3	19.5	19.2	19.4	20.2	20.4	17.7	17.4	16.6	16.7	18.3	18.9
Total C	41.7	42.1	42.1	42.3	44.0	42.5	42.3	42.6	43.5	43.7	41.5	40.3	37.9	37.3	41.0	41.6
+ve S	3.5	1.8	4.1	2.0	3.7	1.4	3.0	1.0	3.4	1.3	3.9	2.0	7.3	5.9	5.1	2.7
-ve S	5.5	3.4	6.3	3.6	5.5	2.9	4.7	1.9	5.1	2.6	5.8	3.3	9.0	8.0	7.3	4.8
Total S	9.1	5.2	10.4	5.6	9.2	4.3	7.7	3.0	8.5	3.9	9.7	5.2	16.3	13.9	12.4	7.5
+ve C	12.9	8.9	14.3	9.8	12.9	8.5	11.5	6.2	12.0	7.7	14.0	9.6	19.6	19.2	16.5	11.9
-ve C	14.1	9.8	15.2	10.9	13.7	9.9	12.4	7.0	12.9	8.5	14.3	10.3	20.0	19.7	17.0	13.3
Total C	27.0	18.7	29.5	20.7	26.6	18.4	23.9	13.2	24.87	16.1	28.3	19.8	39.7	38.9	33.5	25.2

Table 5: **Top:** Distribution of sentiment (connotation) among entire unigrams. **Bottom:** distribution of sentiment (connotation) among discriminative unigrams. 'S' and 'C' stand for sentiment and connotation respectively.

More Successful		
CATEGORY	SUB-CATEGORY	DIFF
Prepositions	General	0.00592
Determiners	General	0.00226
Nouns	Plural	0.00189
	Proper (Singular)	0.00016
Coord. conj.	General	0.00118
Numbers	General	0.00102
Pronouns	Posesseive	0.00081
	General WH	0.00042
	Possessive WH	5.4E-05
Adjectives	General	0.00102
	Superlative	0.00011
Less Successful		
CATEGORY	SUB-CATEGORY	DIFF
Adverbs	General	-0.00272
	General WH	-0.00028
Verbs	Base	-0.00239
	Non-3rd sing. present	-0.00084
	Past tense	-0.00041
	Past participle	-0.00039
	3rd person sing. present	-0.00036
	Modal	-0.00091
Foreign	General	-0.00067
Symbols	General	-0.00018
Interjections	General	-0.00016

Table 6: Top discriminative POS tags.

successful books tend to bear closer resemblance to informative articles.

5.4 Distribution of Constituents

It can be seen in Table 2 that deep syntactic features expressed in terms of different encodings of production rules consistently yield good performance across almost all genres. Production rules are overly specific to gain more generalized, interpretable, high-level insights however (Feng et al., 2012). Therefore, similarly as word categories (POS), we consider the categories of nonterminal nodes of the parse trees, in particular, phrasal and clausal tags, as they represent the gist of constituent structure that goes beyond shallow syntactic information represented by POS.

Table 8 shows how the distribution of phrasal and clausal tags differ for successful books when computed over all genres. Positive (negative) DIFF values indicate that the corresponding tags are favored in more successful (less successful) books when counted across all genres. We also report the number of genres (#Genres) in which the individual difference is positive / negative.

In terms of phrasal tags, we find that more successful books are composed of higher percentage of PP, NP and wh-noun phrases (WHNP), whereas less successful books are composed of higher percentage of VP, adverb phrases (ADVP), interjections (INTJ) and fragments (FRAG). Notice that this observation is inline with our earlier findings with respect to the distribution of POS.

In regard to clausal tags, more successful books involve more clausal tags that are necessary for complex sentence structure and inverted sentence structure (SBAR, SBARQ and SQ) whereas less successful books rely more on simple sentence structure (S). Figure 2 shows the visualization of the distribution of these phrasal and clausal tags.

It is also worth to mention that phrasal and clausal

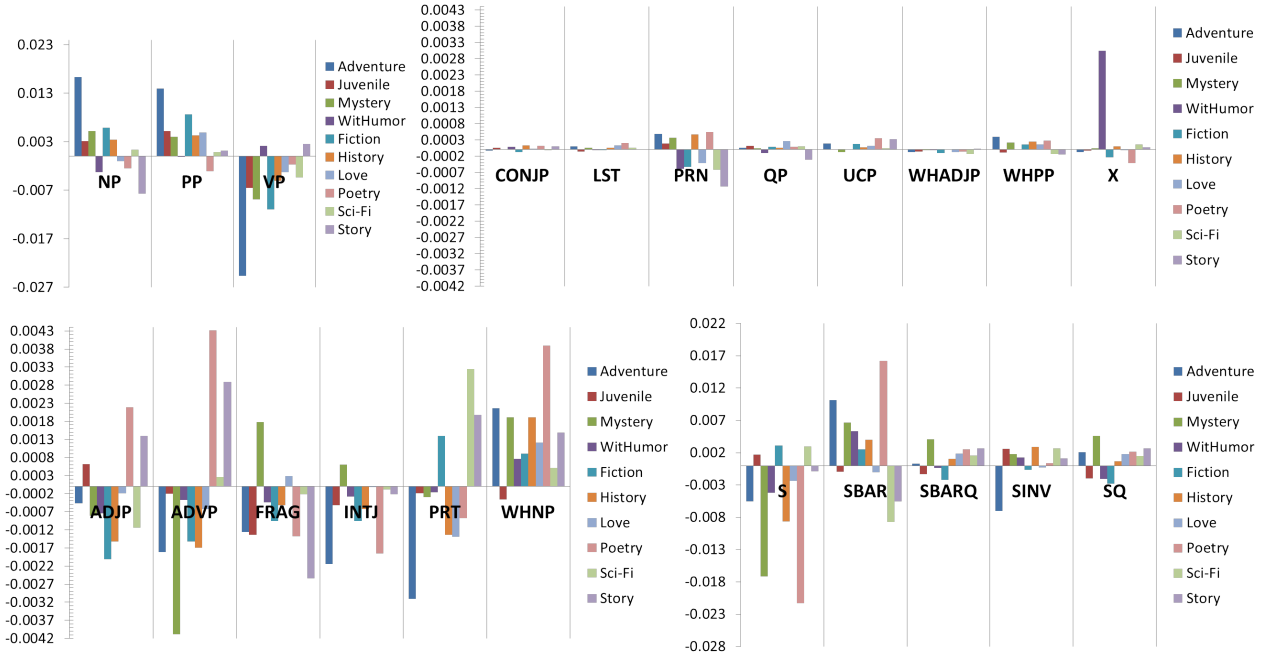

Figure 2: Difference between phrasal and clausal tag percentage distributions of more successful and less successful books across different genres. Specifically, we plot D^{-}–D^{+}, where D^{+} is the phrasal tag distribution (in %) of more successful books and D^{-} is the phrasal tag distribution (in %) of less successful books.

READABILITY INDICES	More Succ.	Less Succ.
FOG index	9.88	**9.80**
Flesch index	87.48	**87.64**

Table 7: Readability: Lower FOG and higher Flesch indicate higher readability (numbers in Boldface).

tags alone can yield classification performance that are generally better than that of POS tags, in spite of the very small feature set (26 tags in total). In fact, constituent tags deliver the best performance in case of historical fiction genre (Table 2).

Connection to Readability Pitler and Nenkova (2008) provide comprehensive insights into assessment of readability. In their work, among the most discriminating features characterizing text with better readability is increased use of verb phrases (VP). Interestingly, contrary to the conventional wisdom – that readability is of desirable quality of good writings – our findings in Table 2 suggest that the increased use of VP correlates strongly with the writing style of the opposite spectrum of highly successful novels.

As a secondary way of probing the connection between readability and the writing style of successful literature, we also compute two different readability measures that have been used widely in prior literature (e.g., Sierra et al. (1992), Blumenstock (2008), Ali et al. (2010)): (i) Flesch reading ease score (Flesch, 1948), (ii) Gunning FOG index (Gunning, 1968). The overall weighted average readability scores are reported in Table 7. Again, we find that less successful novels have higher readability compared to more successful ones.

The work of Sawyer et al. (2008) provides yet another interesting contrasting point, where the authors found that award winning academic papers in marketing journals correlate strongly with increased readability, characterized by higher percentage of simple sentences. We conjecture that this opposite trend is likely to be due to difference between fiction and nonfiction, leaving further investigation as future research.

In sum, our analysis reveals an intriguing and unexpected observation on the connection between *readability and the literary success — that they correlate into the opposite directions*. Surely our findings only demonstrate correlation, not to be con-

Phrasal	+	−	DIFF	$\#^{+}_{Gen}/\#^{-}_{Gen}$
ADJP	0.030	0.031	-6E-4	5/3
ADJP	0.030	0.031	-6E-4	5/3
ADVP	0.052	0.054	-0.002	2/6
CONJP	3E-4	3E-4	2E-5	5/3
FRAG	0.008	0.008	-1E-4	2/6
LST	2E-4	1E-4	5E-5	6/2
NAC	9E-6	6E-6	3E-6	5/3
NP	0.459	0.453	0.005	6/2
NX	1E-4	1E-4	-4E-7	3/5
PP	0.122	0.117	0.005	7/1
PRN	0.005	0.004	2E-4	4/4
PRT	0.010	0.010	-5E-4	3/5
QP	0.001	0.001	7E-5	6/2
RRC	8E-5	8E-5	6E-6	6/2
UCP	8E-4	7E-4	1E-4	8/0
VP	0.292	0.300	-0.008	1/7
WHADJP	2E-4	2E-4	-5E-5	1/7
WHAVP	0	0	0	-
WHNP	0.013	0.012	0.001	8/0
WHPP	0.001	9E-4	1E-4	6/2
X	0.001	0.001	-4E-5	4/4
Clausal	+	−	DIFF	$\#^{+}_{Gen}/\#^{-}_{Gen}$
SBAR	0.166	0.164	0.002	4/4
SQ	0.020	0.018	0.002	7/1
SBARQ	0.014	0.013	0.001	7/1
SINV	0.018	0.018	-6E-5	5/3
S	0.781	0.785	-0.004	3/5

Table 8: Overall Phrasal / Clausal Tag Distribution and analysis. All values are rounded to [3-5] decimal places.

MORE SUCCESSFUL				
BOOK (Q)	$P_{D_{KL}}$	$UP_{D_{KL}}$	S_u	S_{Γ^*}
"Don Quixote" – Miguel De Cervantes	0.139	0.152	+	+
"Other Voices, Other Rooms" – Truman Capote	**0.014**	0.010	+	+
"The Fixer" – Bernard Malamud	0.013	0.015	+	+
"Robinson Crusoe" – Daniel Defoe	0.042	0.051	+	+
"The old man and the sea" – Ernest Hemingway	**0.065**	0.060	+	+
"A Tale of Two Cities" – Charles Dickens	0.027	0.030	+	+
"Independence Day" – Richard Ford	0.031	0.026	+	+
"Rabbit At Rest" – John Updike	0.047	0.048	+	+
"American Pastoral" – Philip Roth	0.039	0.043	+	+
"Dr Jackel and Mr. Hyde" – Robert Stevenson	0.036	0.037	+	+
LESS SUCCESSFUL				
"The lost symbol" – Dan Brown	0.046	0.042	-	-
"The magic barrel" – Bernard Malamud	0.0288	0.0284	+	-
"Two Soldiers" – William Faulkner	0.130	0.117	-	+
"My life as a man" – Philip Roth	0.046	**0.052**	-	+

Table 9: Prediction on books beyond Gutenberg. Shaded entries indicate incorrect predictions.

fused as causation, between readability and literary success. We conjecture that the conceptual complexity of highly successful literary work might require syntactic complexity that goes against readability.

6 Literature beyond Project Gutenberg

One might wonder how the prediction algorithms trained on the dataset based on Project Gutenberg might perform on books not included at Gutenberg. This section attempts to address such a question. Due to the limited availability of electronically available books that are free of charge however, we could not procure more than a handful of books.[8]

6.1 Highly Successful Books

First, we apply the classifiers trained on the Project Gutenberg dataset (all genres merged) on a few extremely successful novels (Pulitzer prize, National Award recipients, etc). Table 9 shows the results of two classification options: (1) KL-divergence based, and (2) unigram-feature based.

Although KL-divergence based prediction was not part of the classifiers that we explored in the previous sections, we include it here mainly to provide better insights as to which well-known books share closer structural similarity to either more or less successful writing style. As a probability model, we use the distributions of phrasal tags, as those can give us insights on deep syntactic structure while suppressing potential noises due to topical variances. Table 9 shows symmetrised KL-divergence between each of the previously unseen novels and the collection of books from Gutenberg corresponding to more successful (less successful) labels. For prediction, the label with smaller KL is chosen.

Based only on the distribution of 26 phrasal tags, the KL divergence classifier is able to make correct

[8] We report our prediction results on *all* books beyond Project Gutenberg of which we managed to get electronic copies, i.e., the results in Table 9 are not cherry-picked.

predictions on 7 out of 10 books, a surprisingly high performance based on mere 26 features. Of course, considering only the distribution of phrasal tags is significantly less informed than considering numerous other features that have shown substantially better performance, e.g., unigrams and CFG rewrite rules. Therefore, we also present the SVM classifier trained on unigram features. It turns out unigram features are powerful enough to make correct predictions for all ten books in Table 9.

Hemingway and Minimalism It is good to think about where and why KL-divergence-based approach fails. In fact, when we included Hemingway's *The Old Man and the Sea* into the test set, we were expecting some level of confusions when relying only on high-level syntactic structure, as Hemingway's signature style is minimalism, with 70% of his sentences corresponding to simple sentences. Not surprisingly, more adequately informed classifiers, e.g., SVM with unigram features, are still able to recognize Hemingway's writings as those of highly successful ones.

6.2 Less Successful Books

In order to obtain less successful books, we consider the *Amazon seller's rank* included in the product details of a book. The less successful books considered in Table 9 had an Amazon seller's rank beyond $200k$ (higher rank indicating less commercial success) except Dan Brown's *The lost symbol*, which we included mainly because of negative critiques it had attracted from media despite its commercial success. As shown in Table 9, all three classifiers make (arguably) correct predictions on Dan Brown's book.[9] This result also supports our earlier assumption on the nature of novels available at Project Gutenberg — that they would be more representative of literary success than general popularity (with or without literary quality).

7 Predicting Success of Movie Scripts

We have seen successful results in the novel domain, but can stylometry-based prediction work on very different domains, such as screenplays? Unlike novels, movie scripts are mostly in dialogues, which are likely to be more informal. Also, what to keep in mind is that much of the success of movies depends on factors beyond the quality of writing of the scripts, such as the quality of acting, the popularity of actors, budgets, artistic taste of directors and producers, editing and so forth.

[9]Most notable pattern based on phrasal tag analysis is a significantly increased use of fragments (FRAG), which associates strongly with less successful books in our dataset.

FEATURE	Adven	Fanta	Roman	Thril
POS	62.0	58.0	61.7	56.0
Unigram	62.0	81.3	70.0	**80.0**
Bigrams	73.3	84.7	80.8	76.0
Γ	66.0	81.3	70.0	76.0
Γ^G	62.0	69.3	**86.7**	60.0
γ	62.0	81.3	78.3	76.0
γ^G	69.3	77.3	77.5	68.0
Γ+Uni	62.0	85.3	70.0	76.0
Γ^G+Uni	54.7	81.3	70.0	76.0
γ+Uni	58.0	**89.3**	70.0	76.0
γ^G+Uni	58.0	84.7	70.0	76.0
PHR	46.0	42.7	65.8	**80.0**
PHR+CLR	**76.7**	31.3	65.8	**80.0**
PHR+Uni	62.0	81.3	70.0	**80.0**
PHR+CLR+Uni	62.0	81.3	70.0	**80.0**

Table 10: Classification results on movie dialogue data (rating ≥ 8 vs rating ≤ 5.5).

We use the Movie Script Dataset introduced in Danescu-Niculescu-Mizil and Lee (2011). It includes the dialogue scripts of 617 movies. The average rating of all movies is 6.87. We consider movies with IMDb rating $\geq$ 8 as "more successful", the ones with IMDb rating $\leq$ 5.5 as "less successful". We combine all the dialogues of each movie and filter out the movies with less than 200 sentences. There are 11 genres ("ADVENTURE", "FANTASY", "ROMANCE", "THRILLER", "ACTION", "COMEDY", "CRIME", "DRAMA", "HORROR", "MYSTERY", "SCI-FI") with 15 movies or more per class, we take 15 movies per class and perform classification tasks with the same experiment setting as Table 2.

Table 10, we show some of the example genres with relatively successful outcome, reaching as high as 89.3% accuracy in FANTASY genre. We would like to note however that in many other genres, the prediction did not work as well as it did for the novel domain. We suspect that there are at least two reasons for this: it must be partly due to very limited data size — only 15 instances per class with the rating threshold we selected for defining the success of

movies. The second reason is due to many other external factors that can also influence the success of movies, as discussed earlier.

8 Related Work

Predicting success of novels and movies: To the best of our knowledge, our work is the first that provides quantitative insights into the unstudied connection between the writing style and the success of literary works. There have been some previous work that aims to gain insights into the secret recipe of successful books, but most were qualitative, based only on a dozen of books, focusing mainly on the high-level content of the books, such as the personalities of protagonists, antagonists, the nature of plots (e.g., Harvey (1953), Yun (2011)). In contrast, our work examines a considerably larger collection of books (800 in total) over eight different sub-genres, providing insights into lexical, syntactic, and discourse patterns that characterize the writing styles commonly shared among the successful literature. Another relevant work has been on a different domain of movies (Yun, 2011), however, the prediction is based only on external, non-textual information such as the reputation of actors and directors, and the power of distribution systems etc, without analyzing the actual content of the movie scripts.

Text quality and readability: Louis (2012) explored various features that measure the quality of text, which has some high-level connections to our work. Combining the insights from Louis (2012) with our results, we find that the characteristics of text quality explored in Louis (2012), readability of text in particular, do not correspond to the prominent writing style of highly successful literature. There have been a number of other work that focused on predicting and measuring readability (e.g., Kate et al. (2010), Pitler and Nenkova (2008), Schwarm and Ostendorf (2005), Heilman and Eskenazi (2006) and Collins-Thompson et al. (2004)) employing various linguistic features.

There is an important difference however, in regard to the nature of the selected text for analysis: most studies in readability focus on differentiating good writings from noticeably bad writings, often involving machine generated text or those written by ESL students. In contrast, our work essentially deals with differentiating good writings from even better writings. After all, all the books that we analyzed are written by expert writers who passed the scrutinizing eyes of publishers, hence it is reasonable to expect that the writing quality of even less successful books is respectful.

Predicting success among academic papers: In the domain of academic papers, which belongs to the broad genre of non-fiction, the work of Sawyer et al. (2008) investigated the stylistic characteristics of award winning papers in marketing journals, and found that the readability plays an important role. Combined with our study which focuses on fiction and creative writing, it suggests that the recipe for successful publications can be very different depending on whether it belongs to fiction or nonfiction. The work of Bergsma et al. (2012) is also somewhat relevant to ours in that their work included differentiating the writing styles of workshop papers from major conference papers, where the latter would be generally considered to be more successful.

9 Conclusion

We presented the first quantitative study that learns to predict the success of literary works based on their writing styles. Our empirical results demonstrated that statistical stylometry can be surprisingly effective in discriminating successful literature, achieving accuracy up to 84% in the novel domain and 89% in the movie domain. Furthermore, our study resulted in several insights including: lexical and syntactic elements of successful styles, the connection between successful writing style and readability, the connection between sentiment / connotation and the literary success, and last but not least, comparative insights between successful writing styles of fiction and nonfiction.

Acknowledgments This research was supported in part by the Stony Brook University Office of the Vice President for Research, and in part by gift from Google. We thank anonymous reviewers, Steve Greenspan, and Mike Collins for helpful comments and suggestions, Alex Berg for the title, and Arun Nampally for helping with the preliminary work.

References

Omar Ali, Ilias N Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2010. Automating news content analysis: An application to gender bias and readability. *Journal of Machine Learning Research-Proceedings Track*, 11:36–43.

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *TEXT-THE HAGUE THEN AMSTERDAM THEN BERLIN-*, 23(3):321–346.

Shane Bergsma, Matt Post, and David Yarowsky. 2012. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 327–337. Association for Computational Linguistics.

Joshua E Blumenstock. 2008. Automatically assessing the quality of wikipedia articles.

Kevyn Collins-Thompson, James P. Callan, and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Dan Douglas and Kathleen M Broussard. 2000. Longman grammar of spoken and written english. *TESOL Quarterly*, 34(4):787–788.

Alvar Ellegård. 1962. *A Statistical method for determining authorship: the Junius Letters, 1769-1772*, volume 13. Göteborg: Acta Universitatis Gothoburgensis.

Hugo J Escalante, Thamar Solorio, and M Montes-y Gómez. 2011. Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 288–298.

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

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Characterizing stylistic elements in syntactic structure. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1522–1533. Association for Computational Linguistics.

Song Feng, Jun Sak Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, Angust. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Robert Gunning. 1968. *The technique of clear writing*. McGraw-Hill New York.

James W Hall. 2012. *Hit Lit: Cracking the Code of the Twentieth Century's Biggest Bestsellers*. Random House Digital, Inc.

John Harvey. 1953. The content characteristics of best-selling novels. *Public Opinion Quarterly*, 17(1):91–114.

Michael Heilman and Maxine Eskenazi. 2006. Language learning: Challenges for intelligent tutoring systems. In *Proceedings of the workshop of intelligent tutoring systems for ill-defined tutoring systems. Eight international conference on intelligent tutoring systems*, pages 20–28.

Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554. Association for Computational Linguistics.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. 1997. Automatic detection of text genre. 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 32–38. Association for Computational Linguistics.

Dan Klein and Christopher D Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics.

Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, volume 69, page 72. Citeseer.

Annie Louis. 2012. Automatic metrics for genre-specific text quality. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, page 54.

Jerome McGann. 1998. *The Poetics of Sensibility: A Revolution in Literary Style*. Oxford University Press.

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*, pages 267–274. 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.

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. Association for Computational Linguistics.

Paul Rayson, Andrew Wilson, and Geoffrey Leech. 2001. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306.

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*, pages 78–86. Association for Computational Linguistics.

Alan G Sawyer, Juliano Laran, and Jun Xu. 2008. The readability of marketing journals: Are award-winning articles better written? *Journal of Marketing*, 72(1):108–117.

Sarah E. Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arlene E Sierra, Mark A Bisesi, Terry L Rosenbaum, and E James Potchen. 1992. Readability of the radiologic report. *Investigative radiology*, 27(3):236–239.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, 60(3):538–556.

Kristina Toutanova and Christopher D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *In EMNLP/VLC 2000*, pages 63–70.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.

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*, pages 1600–1610. Association for Computational Linguistics.

Chang-Joo Yun. 2011. Performance evaluation of intelligent prediction models on the popularity of motion pictures. In *Interaction Sciences (ICIS), 2011 4th International Conference on*, pages 118–123. IEEE.