Quantitative Authorship Attribution of Authors by Employing Supervised Statistical Analyses —Alice Bradley Sheldon and Her Contemporaries—

Miki Kimura

1. Introduction

This study presents a case study of a quantitative authorship attribution dealing with the works of Alice Bradley Sheldon (1915–1987), an American writer of feminist science fiction. When she first began writing in 1967, Sheldon used the male pen name James Tiptree, Jr., both to conceal her identity and as a commercial strategy. In this manner, she successfully disguised her gender for approximately a decade.

During the decade in which the name James Tiptree, Jr., concealed Sheldon's identity, many critics discussed the author's gender because James Tiptree indicated that he had written under a pseudonym. The best-known literary critique was Robert Silverberg's introduction to the 1975 collection of Tiptree's stories, *Warm Worlds and Otherwise*, in which Silverberg (1975) wrote:

> It has been suggested that Tiptree is female, a theory that I find absurd, for there is to me something ineluctably masculine about Tiptree's writing. I don't think the novels of Jane Austen could have been written by a man nor the stories of Ernest Hemingway by a woman, and in the same way I believe the author of the James Tiptree stories is male.

Silverberg continued with a favorable comparison of Tiptree to Ernest Hemingway:

> Hemingway was a deeper and trickier writer than he pretended to be; so too with Tiptree, who conceals behind an aw-shucks artlessness an astonishing skill for shaping scenes and misdirecting readers into unexpected abysses of experience. Additionally there is, too, that prevailing masculinity about both of them.

Later, Lefanu (1989) also compared Tiptree's stories to Hemingway's, noting that Tiptree's manner of writing was decidedly masculine:

> The masculine manner of Tiptree's style is cunning contrivance that reveals, first, the limitations of a machismo-oriented culture and the limitations of science fiction when that oriented culture is incorporated unquestioningly into its fictive conventions.

Silverberg (1997: 282) dealt with Sheldon's revelation and posed a question whether there exists "masculine" and "feminine" science fiction or not by referring to Silverberg (1975). In addition to these critics, Le Guin (1976) suggested that there was a major difference between the works attributed to James Tiptree, Jr., and those attributed to Raccoona Sheldon:

> About Raccoona, by the way, there some of your true lovers did kind of suspect something. Vonda and I have wondered if Raccoona wasn't Tiptree, several times in the past. However, I will tell you a strange thing: I really truly don't like any of Raccoona's stories I've read (only 2 I think, or 3) as well as most of Tip's. They are different.

Sheldon claimed that her bibliography was ambisexual, which could help her disguise her gender. In reality, before she started writing short stories, Sheldon was an army pilot and, after retirement, joined the CIA with her husband. After joining the CIA, she entered a graduate school to study to become a psychologist; she began writing science fiction during her career as a psychologist. Therefore, her bibliography was somewhat ambisexual or masculine. After Sheldon revealed her true identity, she confessed the reasons why she used

these two pseudonyms in an interview in 1986.

In Japan, too, the author has many critics. The most prominent writer publishing papers on Alice Sheldon is Mari Kotani. Ms. Kotani investigated the author's manner of writing and style and suggested that inter-authorial variation existed in Alice Sheldon's texts when compared with Ernest Hemingway in her articles written in 1994 and 1999. In addition to the inter-authorial variations between texts by Alice Sheldon and Hemingway, Kotani (1999) suggested intra-authorial variations in Alice Sheldon's writings, noting that after Alice Sheldon revealed her identity, the manner of writing had changed. In this sense, works written by Alice Sheldon have chronological variations based on Sheldon's biographical story.

To investigate the similarities and dissimilarities between the writings of Sheldon and Hemingway, the first step was to compile complete corpora containing all of their published work. However, two confounding factors existed when comparing their works, namely, different genres and periods. To overcome these problems, additional corpora were compiled for Theodore Sturgeon, Arthur C. Clarke, Ursula K. Le Guin, and Octavia E. Butler.

2. Data and methods

In this study, I performed a quantitative stylistic analysis of Sheldon's work, comparing it with the following male and female writers: Hemingway, Sturgeon, Clarke, Le Guin, and Butler. All except Hemingway were science fiction writers whose careers overlapped Sheldon's. By comparing all of the writings of Sheldon with those of Hemingway and her own contemporaries, I hoped to find clues about Sheldon's allegedly masculine writing style.

The Sheldon corpus compiled for this study contained all of Alice Sheldon's published works under both of her pen names (72 works with 865,802 word tokens). The other five corpora contained all of Hemingway's works (69 works with 271,475 word tokens), Sturgeon's works (222 works with 1,777,561 word tokens), Clarke's works (104 works with 467,983 word tokens), Le Guin's works (45 works with

589,481 word tokens), and Butler's works (93 works with 867,396 word tokens). To maintain relatively consistent sample sizes, 70 of Sturgeon's works were randomly selected from his corpus to use in this study. As I mentioned earlier, additional corpora that were compiled for Sturgeon, Clarke, Le Guin, and Butler will solve the problem for confounding factors in comparison.

The emphasis in this research was primarily on variations between the six authors' works (i.e., inter-author variations). Of particular interest are two questions raised by critics such as Silverberg (1975), Lefanu (1989), and others.

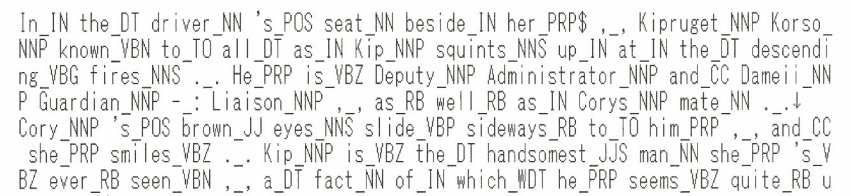
Following the approaches of Hirst and Feiguina (2007), and Hou and Jiang (2014), our quantitative analyses used syntactic variables as effective discriminants — specifically, the distribution of parts of speech (POS). The POS tags that we used in this analysis were attached by using a software named "GoTagger." Table 1 represents the tag set of Gotagger. Figure 1 represents the actual outputs.

Unigrams, bigrams, and trigrams of POS were chosen as the variables. With these variables, we conducted two analyses: support vector machines (SVM) and random forests. Because of the difference of the sample sizes (i.e., five works were written by Raccoona Sheldon and 67 by James Tiptree), when trying to inspect the intra-author variation in Alice Sheldon's works by employing supervised methods, the problem of overfitting occurs. Therefore, the intra-author variations were not inspected by the two supervised learning methods in this study. Moreover the results of these statistical analyses were compared with each other. For those analyses in which the discriminate variables had high sensitivity, the results captured inter-author variations between the works of Alice Sheldon and those of the three male authors. This means that Alice Sheldon may not have the style that many literary critics opine. However, some opposing evidence exists as well, showing that Sheldon's work was sometimes similar to that of Hemingway. The next section provides the full results.

Table 1. Tag Set of Gotagger

Abb.	Parts of Speech	Abb.	Parts of Speech
CC	Coordinating conjunction	PRP$	Possessive pronoun
CD	Cardinal number	RB	Adverb
DT	Determiner	RBR	Adverb, comparative
EX	Existential *there*	RBS	Adverb, superlative
FW	Foreign word	RP	Particle
IN	Preposition/subord. conjunction	SYM	Symbol
JJ	Adjective	TO	*to*
JJR	Adjective, comparative	UH	Interjection
JJS	Adjective, superlative	VB	Verb, base form
LS	List item marker	VBD	Verb, past tense
MD	Modal	VBG	Verb, gerund/present participle
NN	Noun, singular or mass	VBN	Verb, past participle
NNS	Noun, plural	VBP	Verb, non-3rd ps. sing.
NNP	Proper noun, singular	VBZ	Verb, 3rd ps. sing. Present
NNPS	Proper noun, plural	WDT	*wh*-determiner
PDT	Predeterminer	WP	*wh*-pronoun
POS	Possessive ending	WP$	Possessive *wh*-pronoun
PRP	Personal pronoun	WRB	*wh*-adverb

Figure 1. Output of Gotagger.



3. Results

Table 2 lists the results derived from the SVM method using POS unigrams as variables. The classification accuracy of the SVM analysis for discriminating between the three authors was 98.66%, which was

Table 2. Results from Support Vector Machines

	Butler	Clarke	Hemingway	Le Guin	Sheldon	Sturgeon
Butler	**93**	0	0	0	0	0
Clarke	0	**104**	0	0	0	0
Hemingway	0	1	**68**	0	0	0
Le Guin	0	0	0	**41**	3	1
Sheldon	0	0	1	0	**71**	0
Sturgeon	0	0	0	0	0	**70**

significantly greater than that specified by the criterion for classification accuracy (22.96%) when considering the difference of the sample sizes (cf. Kobayashi and Abe (2014)). When we used either POS bigrams or trigrams as variables, the classification accuracy increased to 99.12%.

The second of the two supervised learning methods used random forests. Its results, shown in Table 3, had a classification accuracy of 89.62%, which again was significantly greater than the criterion for classification accuracy (22.96%). When we used POS bigrams and POS trigrams as variables, the classification accuracy was 87.42%, and 81.90%, respectively. From these results, the supervised learning methods can successfully discriminate between the styles of Alice Sheldon and others. However, six works written by Alice Sheldon were misclassified as works by Hemingway. Two works written by Alice Sheldon were misclassified as works by Clarke and two were misclassified as works by Le Guin. Alice Sheldon's writing style was found to be relatively similar to that of Ernest Hemingway, as many literary critics have opined.

Table 4 presents the results from SVM and random forests when employing three kinds of syntactic variables. The classification accuracy of the SVM analysis for discriminating between the six authors was 98.66% when using unigrams of POS. When we used either POS bigrams or trigrams as variables, the classification accuracy increased to 99.12%. The second of the two supervised learning methods used

Table 3. Results from Random Forests

	Butler	Clarke	Hemingway	Le Guin	Sheldon	Sturgeon
Butler	**91**	0	0	1	1	0
Clarke	0	**97**	1	1	2	3
Hemingway	2	3	**57**	3	2	2
Le Guin	1	2	6	**29**	6	1
Sheldon	0	2	6	2	**62**	0
Sturgeon	0	0	0	0	0	**70**

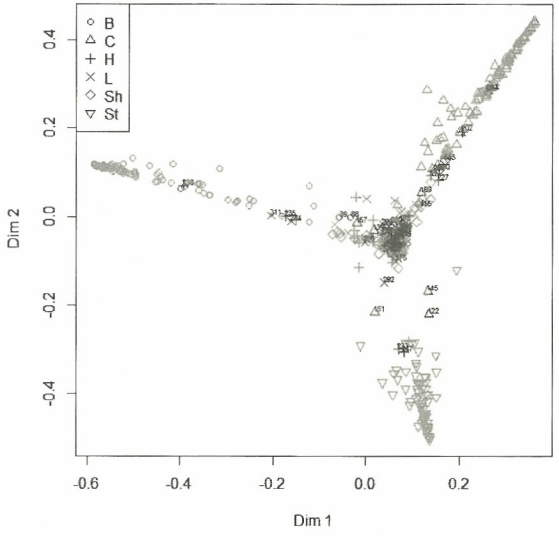
Table 4. Classification Accuracy

	SVM	Random Forests
Unigram	98.66	89.62
Bigram	99.12	87.42
Trigram	99.12	81.90

random forests. The results showed a classification accuracy of 89.62%. When we used POS bigrams and POS trigrams as variables, the classification accuracy was 87.42% and 81.90%, respectively. From these results, we concluded that the supervised learning methods could successfully discriminate between the styles of Alice Sheldon and the others.

Figure 2 shows an MDS plot of the results for random forests based on proximity. The circles, triangles, crosses, second-type of crosses, diamonds, and second-type of triangles represent works by Butler, Clarke, Hemingway, Le Guin, Alice Sheldon, and Sturgeon, respectively. Works that were misclassified are represented in red and include their IDs. The plot shows evidence of six clusters for Butler, Clarke, and Sturgeon. Thus, by employing an MDS plot, inter-author variations can be detected. However, according to this plot, the cluster for Sturgeon is a little farther from the other five clusters. This means that works written by Sturgeon are completely dissociated from the other five authors. We next take a closer look at works written by Alice Sheldon. Works written by Sheldon are located in the origin of

Figure 2. Multi-dimensional scaling plot (variables: unigrams of POS).

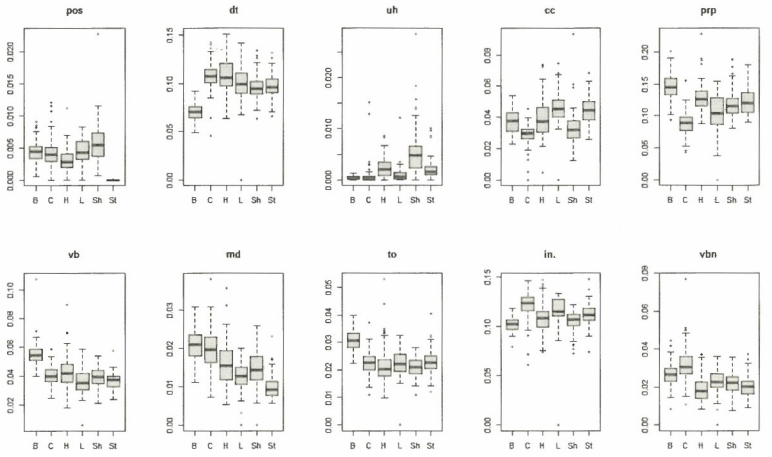


this MDS plot. In this set of data, the style of Alice Sheldon's writing yields a neutral status. Therefore, many works written by Sheldon were misclassified to other classes.

To capture the manner in which the authors use these syntactic variables, I drew 30 boxplots. These plots show the relative frequency of these syntactic variables. Features that I used in these plots are the unigrams of POS. First, the syntactic feature "POS," which represents the possessive ending, is underused in texts by Sturgeon. The syntactic feature "uh," which represents interjections, is overused in texts by Alice Sheldon. The syntactic feature "cc," which represents coordinating conjunctions, is overused in texts by Le Guin and underused in texts by Sheldon. In this sense, Sheldon's writing style is defined by interjections and sentence length in both an exploratory and quantitative sense.

Next, features that I used in Figure 4 are the bigrams of POS. First, the syntactic feature "in dt," which represents the bigram for prepositions and determiners, is underused in texts by Octavia Butler. The syntactic feature "to vb," which represents to-infinitives and

Figure 3. Box plots (variables: unigrams of POS).



verbs, is overused in texts by Octavia Butler. The syntactic feature "md vb," which represents modal auxiliaries and verbs, is overused in texts by Clarke and underused in texts by Sturgeon. The syntactic feature "nn cc," which represents nouns and coordinating conjunctions, is underused in texts by Sturgeon and underused in texts by Sheldon. In addition to these features, "dt jj (determiner and adjectives)," and "jj nn (adjectives and nouns)" are effective for discrimination. These features are not seen in Figure 3, which depicts the inter-authorial variations when using the distribution of POS.

In Figure 5, features that I used in these plots are the trigrams of POS. Similar to the results shown in Figure 4, the feature "md rb vb (modal auxiliaries, adverbs, verb base form)" is overused in texts written by Butler and underused in texts written by Sturgeon. The feature "dt nn cc (determiners, nouns, and coordinating conjunctions)" is overused in texts written by Hemingway and underused in texts written by Sheldon. In this sense, works written by Sheldon do not contain nouns and conjunctions. Therefore, the length of these texts might be shorter than those of any other authors. The feature "dt jj nn (determiners, adjectives, and nouns)," the feature "jj nn in (adjectives, nouns, and preposition/subordinating conjunctions)", and the feature "in

Figure 4. Box plots (variables: bigrams of POS)

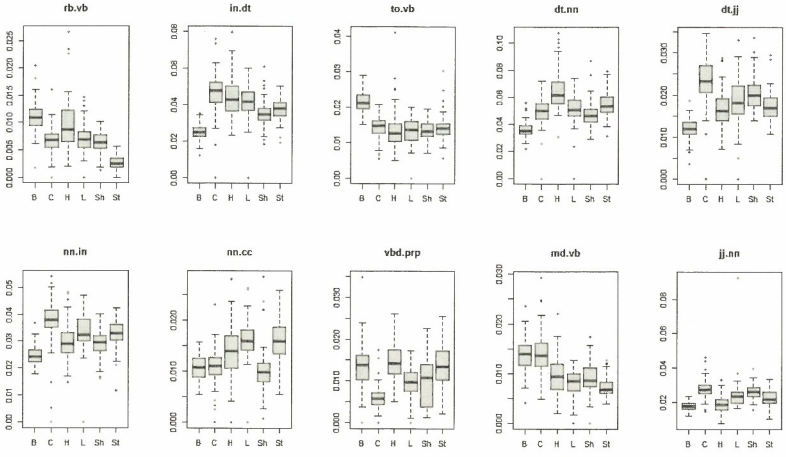
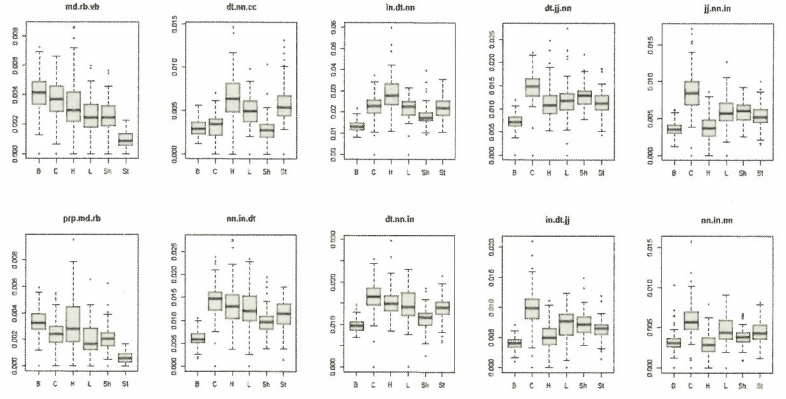


Figure 5. Box plots (variables: trigrams of POS).



dt jj (preposition/subordinating conjunctions, determiners, and adjectives)" are overused in texts written by Hemingway. Adjectives, in particular, which modify nouns, are considerably overused in texts written by Hemingway. In this manner, we can define the authors' style in a quantitative and exploratory manner.

4. Conclusion

My analysis considered the two kinds of supervised statistical

analyses (i.e., SVM and random forests), which did detect inter-author variations between the writing styles of the six authors. Many literary critics suggested that James Tiptree's manner of writing is somewhat masculine, similar to that of Hemingway. Based on the results derived from random forests, six works written by Alice Sheldon were misclassified as being in the Ernest Hemingway group, suggesting some similarities between Sheldon's writing style and that of Hemingway in this small data set and as suggested by literary critics.

This case study inspected the writing styles of only six authors. For future research, we want to apply the same supervised and unsupervised methods to other female authors of science fiction and fantasy from the 1960s and '70s. Apart from statistical methods we used in this study, a machine learning method called "topic modeling" will be employed to inspect the content of texts qualitatively.

WORKS CITED

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

Hou, Renkui, and Minghu Jiang. "Analysis on Chinese quantitative stylistic features based on text mining." *Digital Scholarship in the Humanities* 31, 2 (2014): 357–367.

Kotani, Mari. "Hazama no shisen: Mary Hastings Bradley & James Tiptree, Jr. oyako ni miru passing no seijigaku [Politics of passing as seen in . . .]." *Amerika Kenkyu* 33 (1999): 79–95. Print.

Lefanu, Sarah. "Who Is Tiptree, What Is She?: James Tiptree, Jr . . . " *Feminism and Science Fiction*. Bloomington: Indiana University Press, 1989. Print.

Le Guin, Ursula K. "Dearest Tree." *Letters to Tiptree*. Ed. Alexandra Pierce and Alisa Krasnostein. Yokine: Twelfth Planet Press, 1976. 192–195. Print.

Silverberg, Robert. "Who Is Tiptree, What Is He?" *Warm Worlds and Otherwise*. New York: Ballantine Books, 1975. iv–xviii. Print.

——. *Reflections and Refractions: Thoughts on Science-Fiction, Science, and Other Matters*. California: Underwood Books, 1997. Print.

Yuichiro Kobayashi & Mariko Abe "A machine learning approach to the effects of writing task prompts." *Learner Corpus Studies in Asia and the World*, 2, (2014): 163–175.