Is the Continuation of *The Mystery of Edwin Drood* a Posthumous Work of Charles Dickens? A Multivariate Analysis

Katsumi Goto
Ph.D. Candidate at Chubu University, Japan
gotok@isc.chubu.ac.jp

Abstract
Three years after Dickens' death, Thomas Power James added a continuation to *The Mystery of Edwin Drood*, claiming that it was written by the 'spirit-pen of Charles Dickens, through a medium.' This study attempts to clarify whether the continuation can be considered a posthumous work of Dickens as James suggested, at least as far as the linguistic features are concerned. First, as a preparatory step, the effectiveness of authorship attribution techniques are assessed using six corpora of the leading Victorian novelists: two works by Dickens, including *The Mystery of Edwin Drood*, two works by William Thackeray and two works by George Eliot. Word preferences in these corpora are analyzed using two multivariate analysis techniques, multi-dimensional scaling (MDS) and hierarchical cluster analysis, to check the usefulness of these methods for authorship attribution. It is shown that these methods can successfully distinguish the works of each author from those of the others. Second, the continuation is added to the six corpora and analyzed using the same methods. It is demonstrated that the continuation is quite distinct from the works of Dickens. On the contrary, it is closer to those of Thackeray. The results suggest that James' claim regarding the continuation is not supported.

Keywords: *The Mystery of Edwin Drood,* Thomas Power James, continuation, spirit-pen, multivariate analysis

1 Introduction

The Mystery of Edwin Drood (henceforth *ED*) was published serially up to Dickens' sudden death in June 1870, so the work was left unfinished. In the same year, the first continuation, *The Cloven Foot,* was published in the United States by R. H. Newell under the pseudonym 'Orpheus C. Kerr' (Newell, 1870). This work is a parody and thorough rewrite, and was published at the same time as *ED.* After Dickens' death, Newell added four new chapters. Furthermore, he converted the proper names of characters, and scenes and incidents in *ED* to American style, so that American readers could become familiar with the story.[1]

In the following year, *John Jasper's Secret,* the second continuation, was published by Henry Morford (Morford, 1871). Although he did not alter the proper names, he rewrote the novel thoroughly, and included an original conclusion.

The third continuation, by Thomas Power James, who was an American printer (henceforth James), was published in Brattleboro, Vermont. Three years after Dickens' death, James 'completed' the original *ED* by adding as many as 23 continuation chapters (Dickens [James], 1873). Different from the preceding two continuations, he left the original chapters intact. Additionally, the volume of his continuation was so large that it surpassed the original slightly. In addition to these features, this continuation was notable because James asserted that it was written 'by the spirit-pen of Charles Dickens, through a medium;' that is, written by James, a medium, underpinned by the deceased Dickens' spirit.

The assertion led many critics to comment on the work. Soon after its publication, W.H.B.[2] criticized the continuation:

> The old tricks of style are weakly imitated, here and there bits of carefully arranged sentiment are duly thrown in; […] but the jest is flat, the sentiment is twaddle, and the language resembles that of Dickens as a penny-whistle in the hands of an ambitious urchin resembles a trumpet (B.,W. H., 1874, p. 219).

In addition to this harsh criticism of the work overall, W.H.B. indicated specific grammatical or lexical misuses of, for example, transitive and intransitive verbs (e.g., *lay/lie)*, false concordance in number between the subject and verb, and the past tense form of irregular verbs. He also indicated many 'Yankeeisms,' such as *realise [realize]*[3] used in the sense of *perceive*; *transpire* for *happen*; *located* for *situated/living*; and *directly* for *as soon as* (ibid. pp. 221–22).

In 1905, George F. Gadd commented:

> […], for there is little enough in the second part of the book [the continuation] which, in any degree, serves to recall the genius of the first [*ED*]. […] But those faults are as nothing in presence of the awful grammatical vagaries, […] (Gadd, 1905, p. 272).

Half a century passed, and Arthur Conan Doyle, the creator of the Sherlock Holmes series and also a spiritualist who had examined the posthumous works of Oscar Wilde and Jack London, quoted newspaper articles in *The Boston Post* on the course of events that led to James writing the conclusion, noting: 'The case is surely one which is either deliberate deception or truth' (Doyle, 1927, p. 343). He proceeded:

> The central point of the whole discussion must be the narrative itself. It seems to me to be like Dickens — but Dickens gone flat. The fizz, the sparkle, the spontaneity of it is gone. But the trick of thought and of manner remains (ibid.).

Although Doyle did not take a totally skeptical view of the continuation purportedly being written by a spirit-pen, he had a negative view of its being a posthumous work of Dickens himself (ibid. p. 345).

A century after the publication of the continuation, the issue of the authenticity of the spirit-pen had been minimally advanced. In 1973, Richard Wolkomir reviewed the continuation in the journal *Psychic* (Wolkomir, 1973). 'As befits the journal in which this article appeared, Wolkomir took a positive view of the continuation' (Cox, 1998, p. 569). Although he cited affirmative discourses for the spirit-pen from newspaper articles, the writings of Doyle, and an article about the Dickens-James affair in a Vermont historian's book (Hill, 1961, pp. 178 – 82), he concluded:

> Most investigators today react like the hardbitten reporters from New York and Boston papers who came to Brattleboro in 1873 intending to expose James as a fraud. They left Brattleboro, said the Springfield *Union*, 'absolutely stumped' (Wolkomir, 1973, p. 17).

These criticisms derived from the conventional attentive reading of both *ED* and the continuation, which was a mainstream method of literary criticism at that time. Corpus linguistics has grown in popularity since the latter twentieth century, but to date, there has been no attempt to apply such methods to this issue. This study attempts to clarify the authenticity of the continuation using corpus linguistics techniques.

2 Methods

2.1 Techniques

There is no doubt that the continuation was penned by James' hand. The research question for this study is whether the continuation is sufficiently close to be considered a posthumous work of Dickens, as James suggested, at least as far as the linguistic features are concerned. To settle the issue, word preferences in the continuation and in Dickens' works are analyzed.

Multivariate analysis techniques are used to classify these works according to their word preferences. They can be fundamentally classified into two groups: supervised and unsupervised learning methods. The former, such as support vector machine (SVM) and random forest (RF), require the categorization of data both for the works in question and reference works. Regarding Dickens, many works are available to enable reference to his word preferences, but there are none for James.[4] Therefore, supervised methods cannot be adopted in this study. Thus, unsupervised methods, specifically multi-dimensional scaling (MDS) and hierarchical cluster analysis (henceforth cluster analysis), are adopted.[5] They are considered as effective means for authorship attribution (Tabata, 2016; Hoover, 2003b).

2.2 Corpora

Regarding the data, to assist in evaluating the results of these analyses, another Dickens work and four corpora of other leading Victorian novelists are added to the two corpora (*ED* and the continuation) as third reference corpora: *Our Mutual Friend* (1864) also by Dickens, *Vanity Fair* (1848) and *The Virginians* (1857–59) by William Thackeray, and *Middlemarch* (1871–72) and *Silas Marner* (1861) by George Eliot.[6]

It is essential to exclude inappropriate words for authorship attribution from texts. Three types of words were removed. The first was dialogue, that is, characters' speech. The word usage in dialogue is very different from that in narrative (Burrows, 1987). Narratives are generally considered to reflect the author's language, whereas dialogue is considered to imitate the characters' distinctive language; therefore, all dialogue was removed[7] from the seven corpora. The sizes of the remaining narratives of these corpora are summarized in Table 1.

Table 1. Narrative sizes of seven corpora.

label	D1	D2	T1	T2	E1	E2	P
works	*ED*	*OMF*	*VF*	*Virginians*	*MM*	*Silas*	*Continuation*
Tokens	54,092	174,243	250,735	232,764	208,628	49,196	71,794
Types	5,439	8,743	10,234	9,782	9,234	4,546	4,040

Legend:		
	ED	*The Mystery of Edwin Drood* (1870)
	OMF	*Our Mutual Friend* (1864)
	VF	*Vanity Fair* (1848)
	Virginians	*The Virginians* (1857–1859)
	MM	*Middlemarch* (1871-72)
	Silas	*Silas Marner* (1861)
	Continuation	James' continuation to *ED*

Second, proper nouns and words that were peculiar to individual texts were excluded. These words have little connection with the author's word preferences. They definitely distinguish works, but are rarely words that characterize the author's language. Hence, if they remain, they could cause misleading results in the analysis of authorship attribution. To reduce the influence of changes in, for example, character, setting and subject matter, these words were 'culled' following the methods used by Hoover (2004): (i) proper nouns were removed, and (ii) words for which one of the corpora supplied more than 70% of the occurrences were removed, that is, 'culled at 70%.' Because the same proper nouns could occur in both *ED* and the continuation, the 'culled at 70%' criterion was applied in two steps: (i) in the six corpora, that is, *ED* plus the continuation and the remaining five corpora, considering the former two works as one corpus, and (ii) in the seven corpora considering *ED* and the continuation individually.

Third, personal pronouns were excluded. Their frequency in narratives varies depending on whether the narratives are first or third-person. Thackeray's two works, *Vanity Fair* and *The Virginians*, for example, are first-person narratives. Naturally, first-person pronouns are prominent in them. These characteristics result from the narrative style of the works and not the word preferences of the author.

For each corpus, the remaining text was divided into sections with approximately 10,000 tokens to make the similarity or difference between each corpus more visible in each analysis. The resultant sub-corpora consisted of 66 sections. For each section, word frequencies were counted in terms of lemmas[8] instead of word types, with the expectation that the word preferences in the section would be clearer. Table 2 shows a simplified structure of the word frequency list used in the analyses, which was arranged by first skewering the word frequencies for nearly 9,000 word types across the 66 sections, and then sorting the top 1,200 words in descending order of frequency. Incidentally, the 1,200 words accounted for 85% of tokens among the 9,000 words.

Table 2. Simplified structure of the word frequency list.

		Dicken 1			Dickens 2			Thackeray 1			Thackeray 2			Eliot 1			Eliot 2			James		
rank	types	D1A	···	D1D	D2A	···	D2M	T1A	···	T1P	T2A	···	T2N	E1A	···	E1R	E2A	···	E2C	PA	···	PD
1	**the**	786	···	759	870	···	713	945	···	1145	1229	···	1097	632	···	563	1054	···	838	847	···	933
2	**and**	456	···	507	491	···	505	706	···	869	752	···	812	403	···	384	532	···	518	432	···	493
3	**be**	347	···	358	378	···	396	536	···	595	579	···	638	511	···	462	582	···	594	420	···	504
4	**of**	436	···	382	433	···	375	492	···	613	593	···	646	507	···	387	537	···	400	301	···	363
5	**to**	271	···	351	324	···	326	395	···	493	484	···	548	405	···	387	472	···	472	410	···	479
6	**a**	335	···	330	421	···	323	401	···	425	366	···	367	349	···	324	372	···	368	281	···	302
7	**in**	260	···	247	309	···	278	316	···	341	317	···	386	284	···	250	331	···	264	201	···	265
8	**have**	115	···	207	165	···	204	238	···	296	310	···	300	256	···	288	335	···	322	240	···	280
9	**that**	111	···	169	139	···	202	140	···	226	130	···	167	199	···	214	255	···	259	237	···	309
10	**with**	161	···	197	184	···	211	163	···	213	189	···	191	185	···	147	129	···	201	122	···	150
···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···
···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···
···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···	···
1190	**prospect**	2	···	0	0	···	0	0	···	0	0	···	2	3	···	4	2	···	3	0	···	0
1191	**shoe**	0	···	3	2	···	0	1	···	0	1	···	1	2	···	0	1	···	4	0	···	0
1192	**surround**	1	···	0	2	···	1	0	···	2	2	···	1	0	···	0	0	···	1	2	···	2
1193	**contrast**	0	···	2	1	···	1	0	···	0	1	···	0	2	···	0	1	···	2	0	···	0
1194	**relate**	0	···	1	2	···	2	0	···	1	1	···	0	1	···	0	0	···	0	5	···	0
1195	**bar**	1	···	2	1	···	7	0	···	0	2	···	0	0	···	1	1	···	0	0	···	0
1196	**eagerly**	0	···	0	0	···	0	0	···	2	3	···	2	1	···	0	3	···	1	2	···	3
1197	**utterly**	0	···	0	0	···	0	0	···	1	3	···	4	0	···	0	0	···	2	0	···	1
1198	**disturb**	0	···	2	1	···	0	1	···	0	1	···	2	1	···	1	0	···	0	0	···	0
1199	**sooner**	1	···	1	0	···	1	0	···	0	1	···	0	1	···	0	5	···	0	1	···	2
1200	**merry**	0	···	0	0	···	0	1	···	0	0	···	2	1	···	1	3	···	0	3	···	2

The abbreviation system for the sub-corpus labels used in this study is as follows: For the first character, 'D' denotes Dickens, 'T' Thackeray, 'E' Eliot and 'P' the continuation, that is, the purported posthumous section. With

the exception of James' continuation, the second character, '1' or '2,' distinguishes the two works of each author. The third character, and the second character for the continuation, is assigned to distinguish each sub-corpus.

Hoover showed that cluster analyses based on the 800 most frequent words are usually more accurate for authorship attribution than those of smaller lists (2003a, p. 271). By contrast, he also reported that as few as the ten most frequent words correctly distinguish authors in cluster analysis (Hoover, 2003b, p. 343). Taking these findings into consideration, and to confirm the differences between the results analyzed using different frequent words, lists of the 25, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 1,100 and 1,200 most frequent words were used in the study.

3 Results

3.1 Assessment of the two methods

3.1.1 MDS plot

First, as a preparatory step to classify the continuation, the effectiveness of the two methods, MDS and cluster analysis, was assessed using six corpora, that is, James' continuation was excluded from the seven corpora. A typical MDS plot based on a 500-word list from the six corpora is shown in Figure 1.

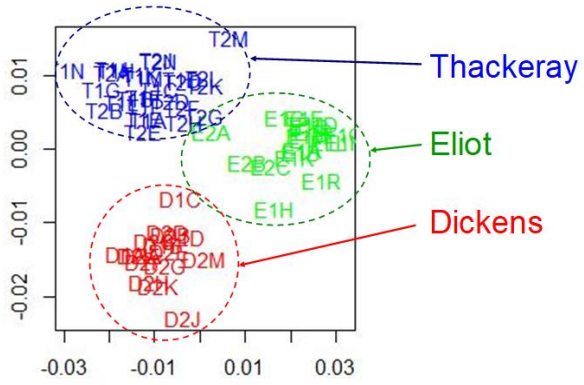

Fig. 1. MDS plot for six corpola: 500-word list.

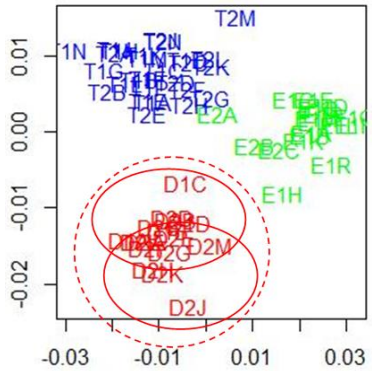

Fig. 2. Example of unclear sub-clustering in the MDS plot.

Figure 1 shows that all sections of the two works of each author are grouped into distinct clusters.[9] This suggests that the method properly distinguishes the authors of the six corpora. Incidentally, roughly the same clustering patterns are observed in all word lists, from 25 to 1,200 words.[10]

Note that the overall similarities or differences between the corpora are visually recognizable from the clustered plots in Figure 1. Hence, in this respect, MDS is very useful. By contrast, more detailed clustering information is difficult to determine. Figure 2 shows such an example: although Dickens' cluster seems to have two sub-clusters, it is barely recognizable. This is because MDS plots cannot exhibit their full information in limited dimensions, for example, two dimensions, that is, a plane.

3.1.2 Cluster analysis

Detailed clustering information is recognizable in cluster analysis. The dendrogram of cluster analysis based on the same 500 words in the previous MDS is shown in Figure 3. The height level at which the dendrogram is cut determines the number of clusters to focus on. When cut into three, as shown in Figure 3, where the number of cuts corresponds to that of authors, each cluster contains all the sections of a single author. Again, roughly the same clustering patterns are observed in all word lists, from 25 to 1,200 words.[11]

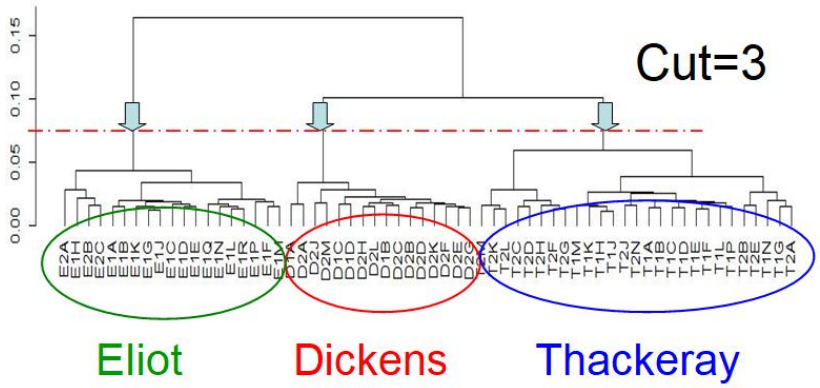

Fig. 3. Cluster analysis for six corpora: 500-word list cut into three.

These results suggest that MDS and cluster analysis distinguish all sections correctly for each author. This suggests that the methods are effective for authorship attribution.

Furthermore, a dendrogram also shows the degree of difference among clusters. When cut into two, as shown in Figure 4, Dickens' sections and Thackeray's sections are grouped into one cluster according to their similarities, and we can observe that the word preferences of Dickens are more similar to those of Thackeray than those of Eliot. Incidentally, such clustering might reflect the difference of literary style among authors.

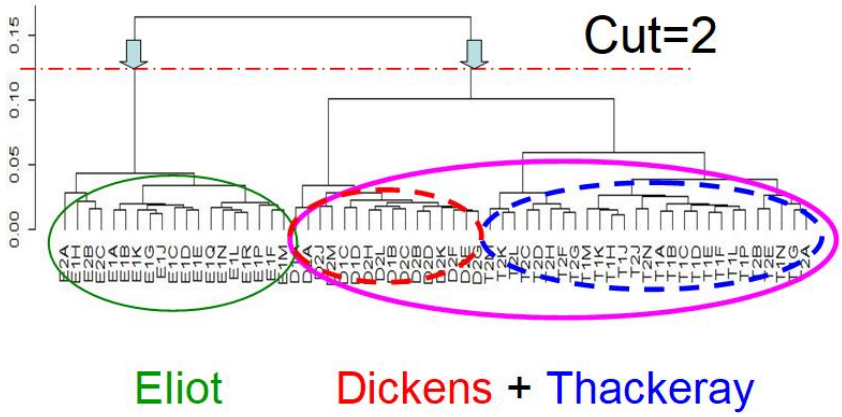

Fig. 4. Cluster analysis of six corpora: 500-word list cut into two.

By contrast, when cut into four, as shown in Figure 5, Thackeray's sections split into two sub-clusters, which reflects the internal differences. Thus, cluster analysis seems to provide more detailed information, and, in this respect, is superior to MDS.

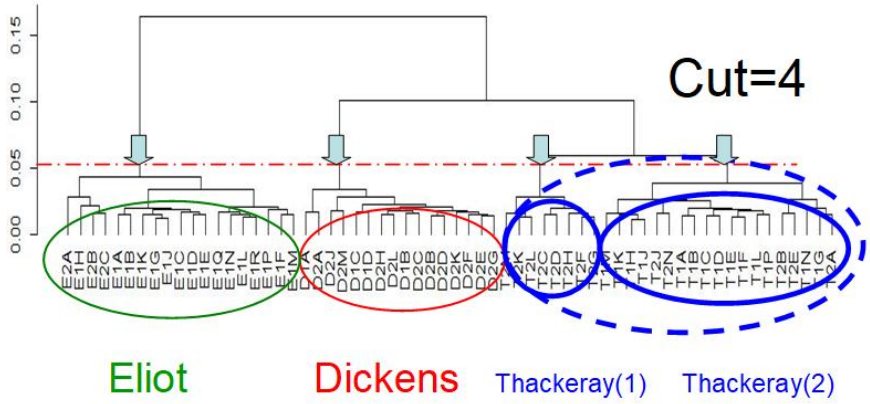

Fig. 5. Cluster analysis of six corpora: 500-word list cut into four.

3.2 Assessment of James' continuation

In this section, James' continuation is included in the word list and analyzed using the same methods. Figure 6 shows the clustered plots of the sections of the continuation. Although in the two-dimensional plot on the left, four

sections of the continuation, PA, PB, PC and PD, seem to be grouped with those of Thackeray or Eliot, they are actually distinct, as shown in the more elaborate three-dimensional plot on the right. The two-dimensional plot corresponds to that viewed from the X-Y plane, which is shown by the big arrow in the three-dimensional plot.

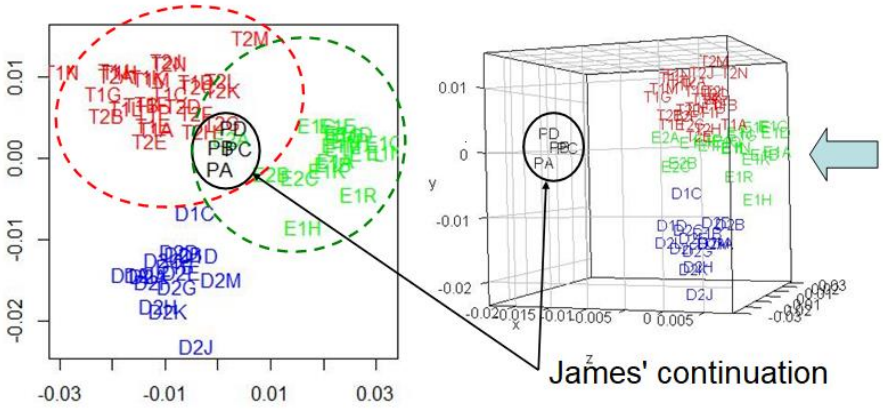

Fig. 6. MDS plot for seven corpora, including James' continuation: 500-word list.

The dendrogram of cluster analysis for the same seven corpora is shown in Figure 7. When cut into four, where the number of cuts corresponds to that of authors including James, each cluster contains all the sections of a single author.

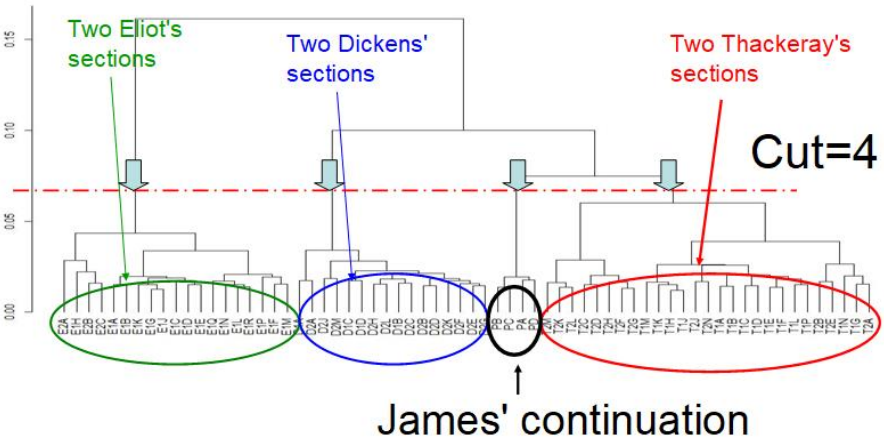

Fig. 7. Dendrogram for seven corpora: 500-word list cut into four.

Interestingly, when cut into three, as shown in Figure 8, the sections of the continuation are grouped with Thackeray's sections. This indicates that if the sections of the continuation are to be grouped with other sections, the first candidates will be those of Thackeray and never those of Dickens.

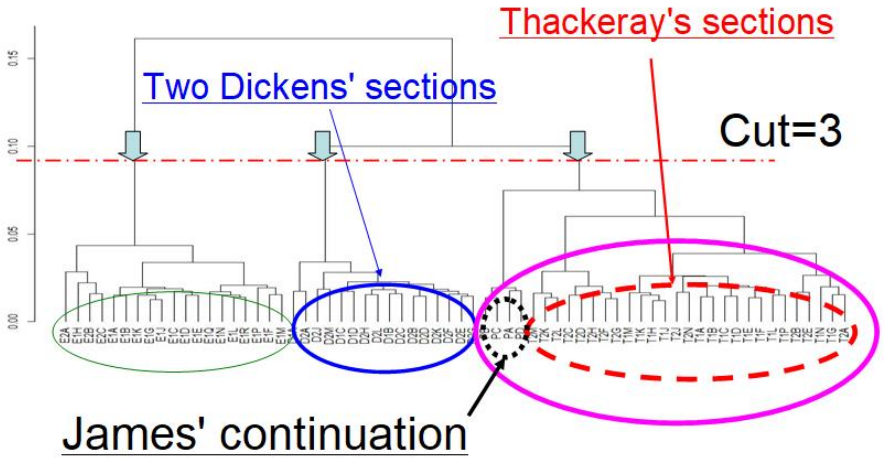

Fig. 8. Dendrogram for seven corpora: 500-word list cut into three.

4 Conclusion

The results of the analyses that focus on the similarity of James' continuation to *ED* are shown in Figures 9 and 10, where *ED* is included in Dickens' works.

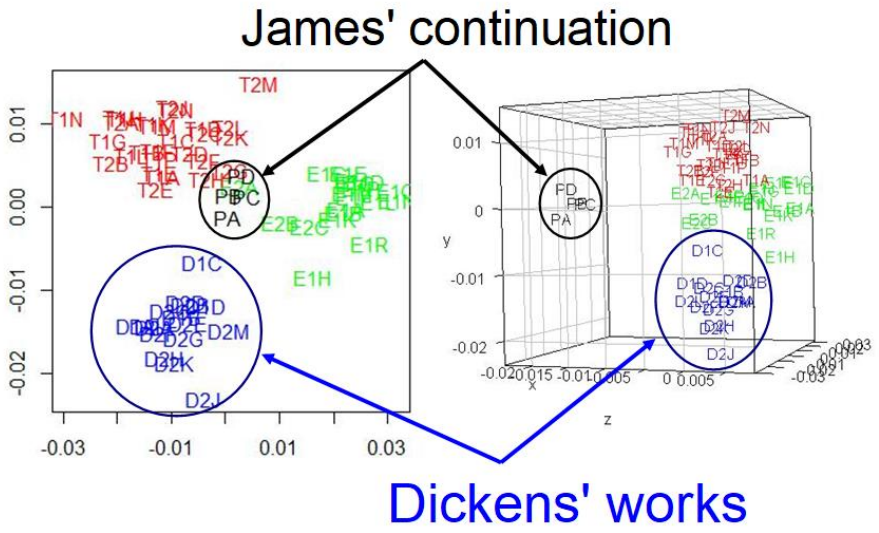

Fig. 9. Dickens' works vs James' continuation (MDS plot for a 500-word list).

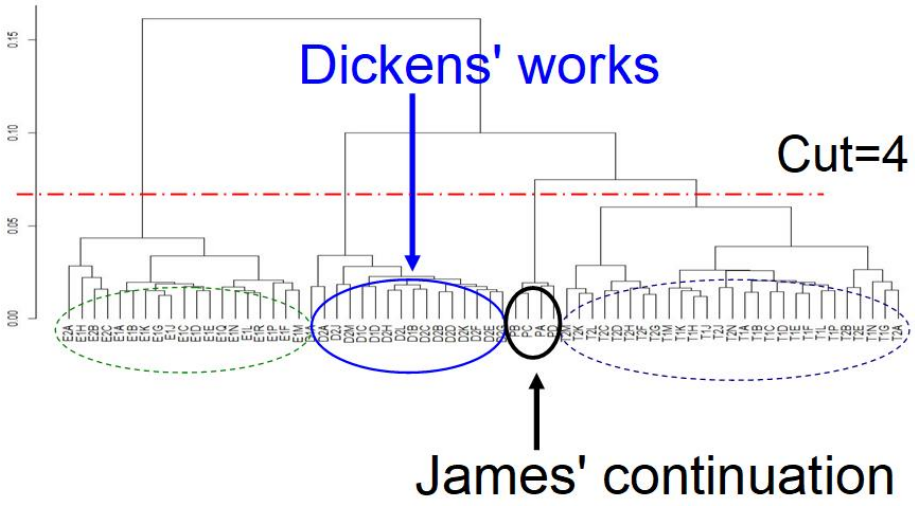

Fig. 10. Dickens' works vs James' continuation (cluster analysis for a 500-word list).

These figures display clear and distinct dissimilarities in word preferences between the continuation and the two Dickens corpora. Hence, it is concluded that the style of the continuation is different from that of *ED*. Genuine works may exist that are written through mediumship, in which the styles of the works are considered indistinguishable from those of the spirit him- or herself while on earth. The analyses described in this paper suggest that different sets of words are favored by different authors, and that James' assertion that the continuation was written by the spirit-pen of Dickens, through a medium, is not supported.

The century-old mystery, that is, the authenticity of the spirit-pen, was only vaguely concluded by authors such as W.H.B. and Gadd (negative conclusions), Wolkomir (affirmative) and Doyle (noncommittal). Although the perspective here is limited to word preferences, this study contributes to solving the mystery with a conclusion based on objective, and hence, more conclusive, evidence obtained from quantitative analysis techniques and comprehensive high-frequency words.

Acknowledgements

The author wishes to thank Masayuki Ohkado for his invaluable comments on the study.

Notes

[1] E.g., John Bumstead (John Jasper), Flora (Rosa), Flowerpot (Rosebud), Bumsteadville (Cloisterham) and clove eater (opium-smoker), where the original words in *ED* are in parentheses.
[2] The author's full name is not known and the initials indicated here were used to sign his work.
[3] Spelled 'realize' in Dickens [James] (1873).
[4] Other than the continuation, a short text of *The Life and Adventures of Bockley Wickleheap* was reportedly found, which, James asserted, was also written in a trance. As this text is only a fragment, 600 words or so, which is less than 0.5% of the continuation, it seems inappropriate as a reference work.
[5] All of the analyses in this study are performed using standardized word frequencies and analytic functions, 'dist', 'cmdscale' and 'hclust', in the R (version 3.4.4) environment. The Euclidean distance as a similarity measure and, in cluster analysis, 'Ward's' as a linkage method are adopted following usual practices in text mining.
[6] Dickens [& James] (1873) was scanned and cleaned up for the corpora of *ED* and the continuation. The other five corpora were downloaded from Project Gutenberg.
[7] Also removed in Hoover (2003a, 2003b).
[8] AntBNC Lemma List (antbnc_lemmas_ver_001)
Laurence Anthony's Website, URL: http://www.laurenceanthony.net/software/antconc
[9] Although a few sections may seem to overlap, they are actually distinct. This can be confirmed using a three-dimensional plot.
[10] The characteristics of high frequency words are evaluated more than the low frequency words in an analysis using the Euclidean distance. In this study, the clustering patterns seem to be formed by as few as the 100 most high frequency words.
[11] The same remarks as those in note 10 hold here.

References

B., W. H. (1874). "Review. Part Second of *The Mystery of Edwin Drood*." *The Southern Magazine*, 14 (February 1874, p. 219–23).

Burrows, John (1987). *Computation into Criticism – A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Cox, Don Richard (1998). *Charles Dickens's The Mystery of Edwin Drood： An Annotated Bibliography.* New York: AMS Press, Inc.

Dickens, Charles [& T. P. James] (1873). *Part Second of the Mystery of Edwin Drood. By the Spirit-Pen of Charles Dickens, Through a Medium. Embracing, Also That Part of the Work Which Was Published prior to the Termination of the Author's Earth-Life*. Brattleboro, VT.:T. P. James.

Doyle, Arthur Conan (1927). "The Alleged Posthumous Writings of Great Authors." *The Bookman 66,* New York. [Online]. URL: https://www.unz.org/Pub/Bookman-1927dec-00342

Gadd, George F. (1905). "The History of a Mystery. A Review of the Solutions to 'Edwin Drood' (Chs. 3 & 4)." *The Dickensian*, 1: 270–273.

Hill, Ralph Nading (1961). *Contrary Country: A Chronicle of Vermont.* Brattleboro, Vermont: The Stephen Greene Press.

Hoover, David L. (2003a). "Frequent Collocations and Authorial Style." *Literary and Linguistic Computing*, 18(3): 261–286.

Hoover, David L. (2003b). "Multivariate Analysis and the Study of Style Variation." *Literary and Linguistic Computing*, 18(4): 341–360.

Hoover, David L. (2004). "Testing Burrows's Delta." *Literary and Linguistic Computing*, 19(4): 453–475.

Morford, Henry (1871) *John Jasper's Secret: Being a Narrative of Certain Events Following and Explaining "The Mystery of Edwin Drood."* London: Wyman and Sons.

Newell, R. H. [Orpheus C. Kerr] (1870) *The Cloven Foot: Being an Adaptation of the English Novel "The Mystery of Edwin Drood," to American Scenes, Characters, Customs, and Nomenclature.* New York: Carleton, London: S. Low, Son & Co.

Tabata, Tomoji (2016). "Kyocho sakuhin niokeru Dickens no Buntai. (Dickens' Style in Collaborative Writings.)" *Corpus to Eigo Buntai. (Corpus and Style of English.) Eigo Corpus Kenkyu Series. (English Corpus Study series.)* Ed. Masahiro Hori. vol.5. Tokyo: Hitsuji Shobo.

Wolkomir, Richard (1973). "Charles Dickens' Great Mystery." *Psychic*, 4: 16–17.