Teasing Out Authorship and Style with T-tests and Zeta

Hoover, David L.
david.hoover@nyu.edu
New York University, USA

Most computational stylistics methods were developed for authorship attribution, but many have also been applied to the study of style. Investigating Wilkie Collin's *Blind Love* (1890), left unfinished at his death and completed by Walter Besant from a long synopsis and notes provided by Collins, requires both authorship attribution and stylistics. External evidence indicates that Besant took over after chapter 48 (Collins 2003), which provides an opportunity to test whether Besant was successful in matching Collins's style and to investigate the styles of Collins and Besant. This divided novel also facilitates the comparison of two computational methods: the T-test and Burrows's Zeta.

The t-test is a well-studied method for determining the probability of a difference between two groups arising by chance (a classic use in authorship and stylistics is Burrows 1992.) Here I use t-tests to identify words used very differently by Collins and Besant. After showing that those word frequencies accurately identify the change of authorship, I examine the words themselves for stylistically interesting characteristics.

I created a combined word frequency list for four novels by Besant and three by Collins, then deleted words occurring only once or twice, personal pronouns (too closely related to the number and gender of characters), all words with more than 90% of their occurrences in one text (almost exclusively proper names), and words limited to one author (required for t-testing). I divided the novels into 167 4,000-word sections, and performed t-tests for the remaining 6,600 words (using a Minitab macro). I cleaned up the results and sorted them on the p value in Excel (with another macro), and retained only the 1719 words with $p < .05$, about 1,000 for Collins and 700 for Besant (see `https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html` for detailed instructions and the macros).

I tested these words on six new texts for each author, a novel and five stories for Besant and six novels for Collins. Beginning with the 500 most distinctive words for each author, I deleted a few words that were absent from these texts and used the remaining 993 words to perform a cluster analysis (Fig. 1). (To keep the graph readable, I divided the novels into 10,000- word sections, retaining only half the sections.) Obviously, these marker words are quite characteristic of the authors.

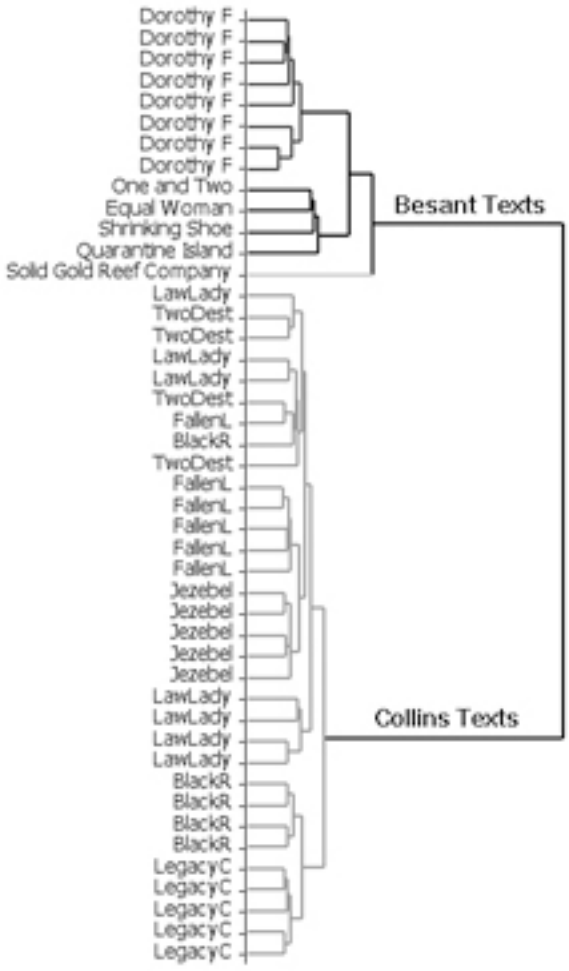

Fig. 1. *Besant versus Collins: Cluster Analysis*

When sections of *Blind Love* are tested along with the texts above, the authorship change after chapter forty-eight is starkly apparent (Fig. 2). This graph is based on the sums of the frequencies of the 500 most distinctive words for each author in each section. (The texts are divided into 1,000-word sections; only a few sections of the novels are shown; the frequencies

of Collins's marker words are multiplied by -1 for clarity.) Although Besant was working from extensive notes, his style is distinctly different. Had we not known which was Besant's first chapter, these t-tested marker words would have easily located it.

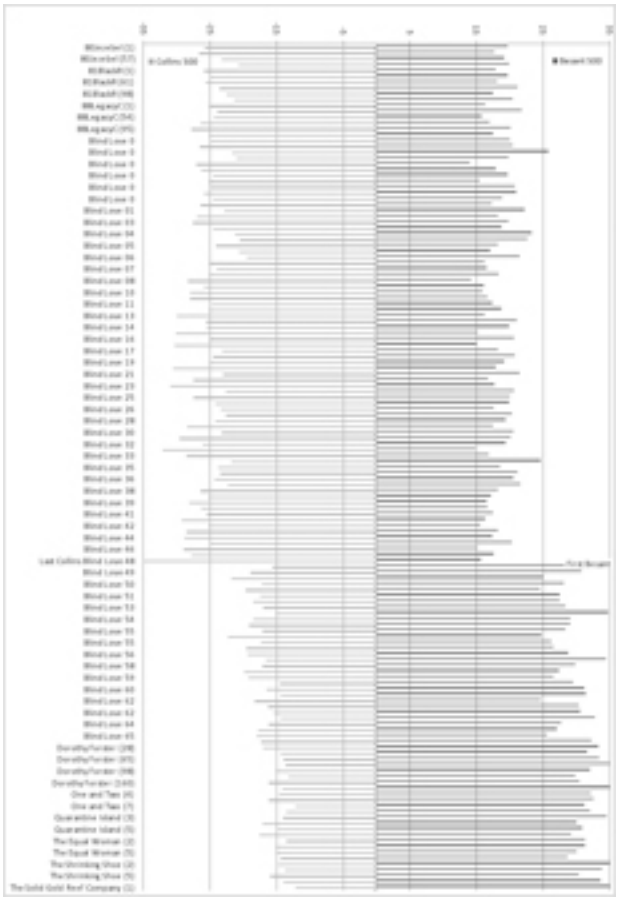

Fig. 2. *Besant, Collins, Blind Love: T-tested Marker Words*

Because the styles of Collins and Besant are so distinct, these marker words should also characterize them. Consider the twenty most distinctive words for each author:

Besant: *upon*, *all*, *but*, *then*, *and*, *not*, *or*, *very*, *so*, *because*, *great*, *thing*, *things*, *much*, *every*, *there*, *man*, *everything*, *is*, *well*

Collins: *answered*, *to*, *had*, *Mrs*, *on*, *asked*, *in*, *Miss*, *mind*, *suggested*, *person*, *resumed*, *excuse*, *left*, *at*, *reminded*, *creature*, *inquired*, *reply*, *when*

Obviously, more of Besant's words are high frequency function words, and many Collins words are related to speech presentation (*answered*, *asked*, *inquired*, *resumed*, *suggested*, *reply*, and *reminded*). The presence of *added*, *begged*, *declared*, *exclaimed*, *explained*, *expressed*, *muttered*, *rejoined*, and *said* as likely speech markers among the other Collins marker words, but only *gasped*, *groaned*, *murmured*, *replied*, and *stammered* for Besant, suggests they have different ways of presenting speech.

Sorting all of each author's marker words alphabetically immediately reveals word families that each author favors, as *thing*, *things*, and *everything* among the twenty most distinctive Besant words already suggests (*anything* and *nothing* are also Besant markers). His *every* and *everything* are joined by *everybody* and *everywhere*; *anything* by *any* and *anywhere*; *nothing* and *not* by *never*, *no*, *nobody*, *none*, and *nor*; and *much* by *more*, *moreover*, *most*, and *mostly* among his markers. Collins's *answered* is joined by *answer*, *answering*, and *unanswerable*; and five of his twenty words are joined by two others: *ask*, *asked*, *asks*; *inquired*, *inquiries*, *inquiry*; *leave*, *leaving*, *left*; *person*, *personally*, *persons*; *suggest*, *suggested*, *suggestion*.

About 600 of the 1,700 distinctive words form groups favored by one author, but only about 175 form split groups, many of which fall into intriguing patterns. Collins uses more contractions, so *didn't*, *doesn't*, and *don't* are Collins words, but *did* and *does* are Besant words, and similarly for *must*, *need*, *should*, and *would* and their negative contractions. The singular and possessive forms of *brother*, *friend*, *sister*, and *son* are Collins's words and the plural forms are Besant's; the singular vs. plural pattern continues almost without exception in split noun groups. Verbs in *-ing* are Collins words and 3rd singular present forms Besant's. Finally, all nineteen cardinal number marker words are Besant's, including the numbers *one* to *ten* (note that Besant's preferred plural nouns often follow numbers). This extraordinary patterning may not seem particularly surprising, but, so far as I know, it has never been noticed before, and cries out for investigation.

Two problems with t-testing are its privileging of relatively uninteresting high-frequency words and its inability to cope with words absent from one author. John Burrows's Zeta addresses both of these problems (Burrows 2006). (The specific form used here was developed by Hugh Craig (Craig and Kinney, 2009); for an automated spreadsheet and instructions for performing

Zeta analysis see `https://files.nyu.edu/dh3/public/ TheZeta&IotaSpreadsheet.html`).

Zeta's simple calculation begins with the same novels and the same word frequency list used for the t-test, except that personal pronouns and words present in only one author are now included. Zeta is simply the sum of the proportions of Collins sections in which each word occurs and Besant sections in which it does not. Here *answered*, the most distinctive Collins word (as in the t-tests), has a Zeta score of 1.8, and is present in 89 of 90 Collins sections and absent from 65 of 77 Besant sections. The most distinctive Besant word is again *upon*, present in all 77 Besant sections and absent from 25 of 90 Collins sections, with a Zeta of 0.28. Below are the twenty most distinctive Zeta words (those also identified by t-testing in bold):

Besant: ***upon***, *fact*, *presently*, *therefore*, *however*, ***everything***, *real*, *whole*, *cannot*, *though*, *rich*, *none*, *thousand*, *except*, *fifty*, *ago*, ***because***, *papers*, *also*, *twenty*

Collins: ***answered***, ***Mrs***, ***Miss***, ***excuse***, ***suggested***, ***resumed***, ***reminded***, *doctor*, ***inquired***, ***creature***, *notice*, *circumstances*, *tone*, *idea*, *temper*, *object*, *sense*, *feeling*, *governess*, *impression*

As noted above, Zeta marker words are less frequent than t-tested words. Only two Zeta marker words rank in the top 100 in the novels, compared to 20 of the t-tested words. About 3/4 of the 1000 t-tested marker words are also among the 1000 Zeta markers. Among the 2000 Zeta words are 275 words occurring in only one author; 59 form new single-author families, 27 join existing single-author families, and only 21 form split families.

The Zeta words also effectively detect the change of authorship in *Blind Love*. In the scatter graph in Fig. 3, the axes show the percentages of all the word types (unique words) in each section that are Besant or Collins marker words (longer texts are divided into 4000-words sections; the labels for even-numbered Collins sections of *Blind Love* are removed; only a few sections of other novels are included). Note how distinct Besant's chapters of *Blind Love* are from Collins's, though many of them are pulled toward Collins. This graph also includes *The Case of Mr. Lucraft* (Case in bold), jointly written by Besant and James Rice; it suggests, as has been argued (Boege 1956: 251-65), that Besant did most of the actual writing.

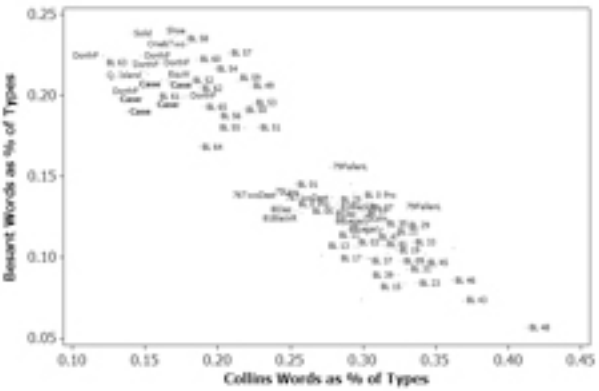

Fig. 3. *Besant, Collins, Blind Love: Zeta Analysis*

T-tests and Zeta analysis are both effective authorship attribution methods that produce lists of characteristic vocabulary for the authors being compared. Both identify morphological and semantic families of words and uncover extraordinarily consistent patterns and puzzling inconsistencies that suggest new directions for literary and stylistic analysis.

References

Boege, F. (1956). 'Sir Walter Besant: Novelist. Part One'. *Nineteenth-Century Fiction*. **10**: 249-280.

Burrows, J. F. (1992). 'Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information'. *LLC*. **7**: 91-109.

Burrows, J. F. (2006). 'All the Way Through: Testing for Authorship in Different Frequency Strata'. *LLC*. **22**: 27-47.

Collins, W. (2003). *Blind Love*. Bachman, M., Cox, D. (eds.). Peterborough, Ont.: Broadview Press.

Collins, W. (2009). *Blind Love*. London: Chatto & Windus. `http://ia311528.us.archive.org/0/items/blindlove00colluoft/blindlove00colluoft.pdf` (accessed 18th March 2009).

Craig, H., Kinney, A., (eds.) (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.