35

Using corpus linguistics to explore the language of poetry: a stylometric approach to Yeats' poems

Dan McIntyre and Brian Walker

1 Poetry as the object of corpus linguistic analysis

Over the last 15 or so years, methods from corpus linguistics have become ever more common in the analysis of literary texts (see, for example, Semino and Short 2004; O'Halloran 2007; Mahlberg 2012; Mastropierro 2018; McIntyre and Walker 2019; O'Halloran 2007; Semino 2004). Much of this work has been largely stylistic in nature and has focused on the propensity of textual elements to function as triggers for literary effects. As a result, such work has come to be known generally as corpus stylistics. More recently, corpus stylistics has begun to turn its attention back to some of the earliest concerns of stylisticians, namely the characteristics of genre style (e.g. Montoro and McIntyre 2019) and authorial style (e.g. Evans 2018). In this endeavour, stylistics has started to draw on the methods of the closely related discipline of stylometry and its techniques of authorship attribution (e.g. Hoover and Hess 2009).

Our aim in this chapter is to demonstrate the value of combining corpus stylistic techniques with stylometric methods of analysis in order to generate stylistic insights into the language of poetry. We analyse a specially constructed corpus of William Butler Yeats' poetry containing 307 of his published poems to determine whether there is textual evidence of a change on Yeats' style over time. The poems in our corpus match those presented in *The Collected Works of W. B. Yeats* (Finneran 1997) and are available in electronic form at webpages hosted by California State University, Northridge (CSUN). We use our Yeats corpus of poems to answer a series of research questions which aim to show how combining corpus stylistics and stylometry can shed light on both authorial style and text style.

2 W. B. Yeats: literary critical responses to his poetry

W. B. Yeats was a prolific poet and dramatist who continued to write until his death in 1939. There is some agreement among literary critics that Yeats' style changed and

DOI: 10.4324/9780367076399-35

developed over his long writing career and that these changes fall into distinct writing phases. There is less agreement, however, about the number, timing and duration of these phases. Carter and McRae (2017), for instance, suggest that there are three: early (approximately 1889–99), middle (approximately 1904–28) and late (approximately 1930–39). Sarker (2002), on the other hand, says that it is "fashionable" to divide Yeats' style development into four phases: Celtic twilight (1889–99), middle (1904–14), transitional (1919–35) and last (1938–9). There is a consensus that Yeats' first writing phase comprises poems published during the late nineteenth century and was influenced by the pre-Raphaelites, romanticism and Irish mythology. His subsequent writing moved away from these influences and became shaped by modernism. Luebering (2011), for example, notes that Yeats' poems of the late 1800s have a "dreamlike atmosphere" and are heavily reliant on Irish folklore and legend, whilst, in the first of his twentieth-century volumes: *In the Seven Woods* (1904) and *The Green Helmet* (1910), Yeats 'discarded the Pre-Raphaelite colours and rhythms of his early verse and purged it of certain Celtic and esoteric influences' (Luebering 2011: 168; see also Davis 2015). For Jeffares (1968: 100), the poems of *The Green Helmet* (1910) are 'transitional' in the sense that Yeats was 'still forming a new style' and 'stripping off the decoration of the old' (see also Carpenter 1969: 51–9). Matthews (2014) and Sarker (2002) suggest that 1914 saw a notable change in Yeats' style, with Matthews (2014: 335) arguing that *Responsibilities* (1914) 'marked a radical shift in his style toward the techniques and values of Pound and the younger generation'. Our analysis in Section 4 aims to determine whether there is textual evidence for some of these literary critical claims concerning Yeats' changes in style over time.

3 Stylometry and corpus stylistics

Our aim in this chapter is to explore some of the claims regarding Yeats' changing style using a combination of two computational approaches to stylistics: stylometry and corpus stylistics. From stylometry we use cluster analysis and principal components analysis (PCA), and from corpus stylistics we use keyword analysis. Our stylometric analysis is aimed at establishing groupings of Yeats' volumes of poetry to determine whether there is any linguistic evidence for literary critical views concerning the constituent periods of Yeats' career as a poet. Our corpus stylistic analysis of keywords is aimed at investigating Yeats' style more closely to determine whether there are any language features that are common to particular periods of his writing life.

Stylometric analysis (sometimes referred to as computational stylistics) draws on a range of techniques, including cluster analysis and principal components analysis, which we use for our analysis. Both are data reduction methods that can be used to look for underlying patterns in the data. For example, while we might intuitively assume that Yeats' early poems share more lexical characteristics with each other than they do with his later work, stylometric analysis reveals whether this is actually the case. In effect, stylometry offers a data-driven approach to identifying similarities between texts. So, instead of asserting, say, three distinct periods to Yeats' poetry and then looking for textual evidence of this, stylometric methods group the target texts based on lexical frequencies in the texts. It is then up to the analyst to decide whether the outcome of the statistical analysis supports non-statistically derived claims.

Cluster analysis is a statistical procedure by which texts are grouped together into clusters based on similarities of word frequencies within the texts. One famous and ground-breaking method is that developed by the linguist John Burrows, known as the

Burrows method (see Burrows 1987, 2002, 2003). The basic principle is to establish the most frequent words (MFWs) and their frequencies across a group of texts and then assess how the frequencies of those words within individual texts measure up against the group frequencies. The number of MFW (written *n*MFW) can vary, but for the purposes of this explanation, we will assume 100 MFW (written *n*MFW = 100). To ensure comparability, normalised frequencies are used (for more on normalisation, see Chapters 10 and 39, this volume) and these are compared against the group frequencies using a distance measure (i.e. measuring "distance" between frequencies; Burrows (2002) uses a measure called Delta). The texts under analysis are thus grouped based on the 100 most frequent words and their distances from the frequencies established for the group. The process also involves applying a method of linkage, which is a mathematical technique for organising the results into groups based on the distance calculations (see Hoover 2003). The closest two texts in terms of similarity of MFW are grouped together into what is known as a cluster. The process then continues, with the next most similar texts being grouped together and so on. The process results in all the texts under analysis being grouped together into one cluster of multiple strands (see Hoover 2013 for further explanation). The visual output from a cluster analysis is called a dendrogram (see Figures 35.1 and 35.2). We discuss how to interpret dendrograms in the next section.

Like cluster analysis, principal components analysis is a data reduction method. However, rather than grouping texts based on the 100 MFW, PCA aims to simplify the dataset in order to reveal the cause of most of the variation in the data. A non-linguistic example may make this easier to follow. Imagine you are collecting statistics on motorbikes, particularly speed, colour, number of cylinders and manufacturer. As you look at your data, you realise that the relationship between speed and number of cylinders is so close that it is simpler to conflate these two variables into one new variable: power. By so doing, you have reduced the complexity of the dataset and made it easier to see what distinguishes certain motorbikes from others. When applied to language, PCA works on the same principle. The procedure aims to conflate information about the 100 MFWs into a smaller number of variables known as components. The visual output of PCA is a graph on which the values of the two most significant components are plotted (see Figures 35.3 and 35.4). These two components – principal component 1 (PC1) and principal component 2 (PC2) – are deemed to be the most significant because they account for most of the variation in the data (again, see Hoover 2013 for more details). PC1 is plotted on the horizontal axis, while PC2 is plotted on the vertical axis. At this stage, the analyst's task is to interpret what non-linguistic variable (e.g. author, genre, time period) the grouping of the texts on the PCA graph reflects.

To carry out our cluster analysis and principal components analysis, we used the *Stylo* package for *R* (Eder *et al.* 2016), developed by the Computational Stylistics Group at the Institute of Polish Language in Kraków. *Stylo* provides a convenient and well-documented way to perform Burrows-style analyses using *R*.

Turning now to our corpus stylistic methods, in addition to our stylometric analyses, we carried out keyword analysis of Yeats' poems. Keyword analysis is based on the notion of keyness, which refers to unusually high or unusually low frequencies of particular words within a source text or corpus when compared, using statistical tests to a reference corpus or text (see Chapter 9, this volume). Keyness analysis offers a principled way to discover avenues for further investigation in a text or corpus. Keyness, however, is not necessarily an indication that a word is interpretively significant. This needs to be established by qualitative analysis of each keyword in context.

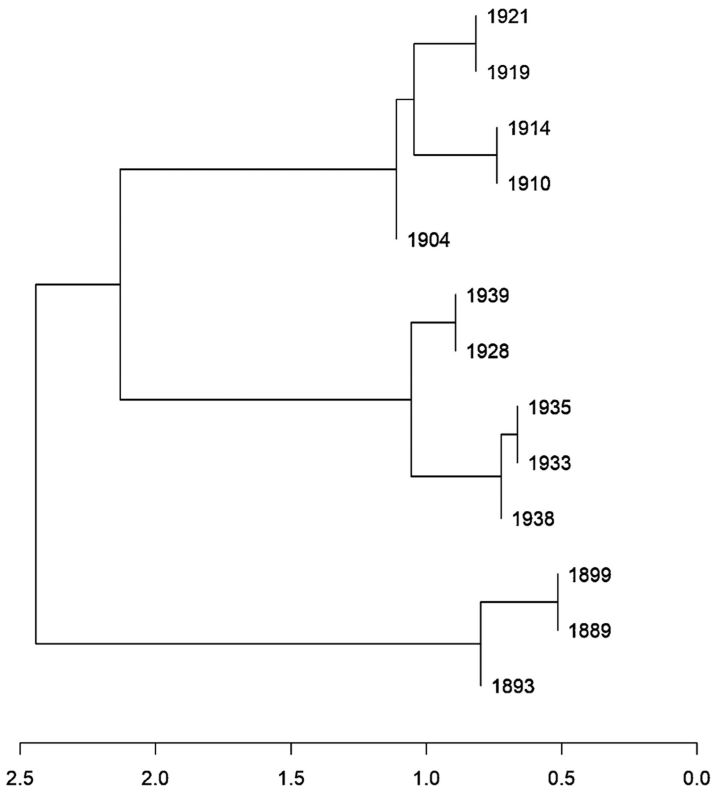

Figure 35.1 Cluster analysis of Yeats' poetry in 13 volumes based on *n*MFW = 100

In our study we made comparisons between the words and their associated frequencies in different volumes of Yeats' poetry using the keyword facility in *AntConc* (Anthony 2019), which is freely available corpus analysis software. Current thinking on the calculation of keywords (see Gabrielatos 2018) is that two statistical measures are required: statistical significance and effect size. The use of effect size as part of statistical testing has been argued for within social sciences more generally (see Levin and Robinson 1999; Thompson 1999). The current version of *AntConc* (Anthony 2019) offers several different statistics for this purpose. We used log likelihood for keyness and log ratio for effect size.

Log likelihood (LL) is an inferential statistic used to measure statistical significance (see Chapter 13, this volume). It indicates the confidence with which we can infer that any keywords are a result of the datasets being different in some way. Such statistical significance testing depends on the notion of a null hypothesis (usually denoted by H_0), which states that there is no difference between the two datasets or, more precisely, no difference in the populations from which the datasets were drawn (see McIntyre and Walker 2019: 154–8 for further explanation). Log likelihood provides an indication of the probability of getting the results if H_0 were true (i.e. no difference). The higher the log likelihood value, the lower the probability that the results would be obtained if H_0 is true. The probability is usually presented as a *p*-value, where the lower the number, the lower the probability. A *p*-value of less than 0.05 ($p < 0.05$) indicates that there is a 5 per

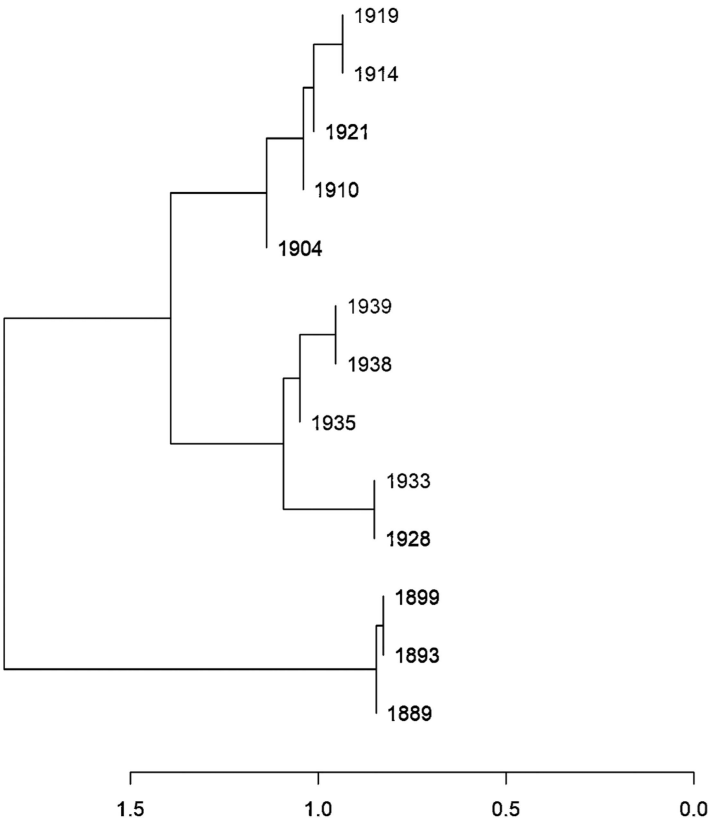

Figure 35.2 Cluster analysis of Yeats' poetry in 13 volumes based on *n*MFW = 1,000

cent chance that the results would be obtained if the null hypothesis were true; $p < 0.01$ indicates a 1 per cent chance, and $p < 0.001$ indicates a 0.01 per cent chance. The usual practice is to set a level of significance below which results are ignored. This means setting the level of confidence at which you are happy to reject the null hypothesis (i.e. that there is no difference between datasets). With LL the higher the score, the more confidently we can reject H_0. Using the settings in AntConc, we used a statistical cut-off of 10.83 (which equates to $p < 0.001$), so words with LL scores below 10.83 were ignored. This threshold is arbitrary but is a recognised cut-off in linguistics and social sciences more generally.

Log ratio (LR) is a descriptive statistic developed by Hardie (2014) for keyword research. It indicates the scale of the difference between word frequencies, which is known as effect size. A log ratio of 1 means that the frequency in the target corpus is twice as much as the frequency in the reference corpus; an LR of 2 means that the difference is four times as much, an LR of 3 means an eightfold difference and so on. We applied an (arbitrary) log ratio cut-off of 1 in order to eliminate keywords where the difference in frequencies was small (for further explanation of LR, see McIntyre and Walker 2019: 158–64).

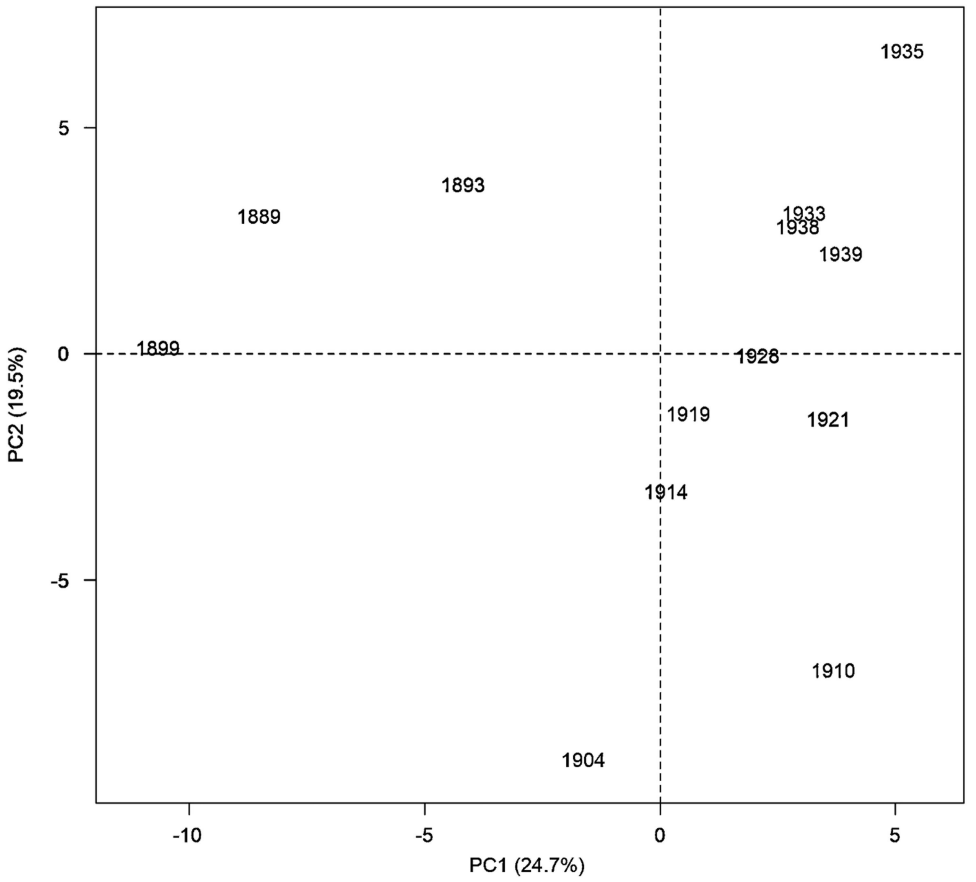

Figure 35.3 PCA of 13 volumes of Yeats' poetry *n*MFW = 100

4 A multivariate analysis of the style of W. B. Yeats

Cluster analysis and principal components analysis are methods of multivariate analysis, hence the title of this section. The research questions that we set out to answer in our multivariate analysis are:

1. Does multivariate analysis reveal lexical differences and/or similarities that support the literary critical claim that Yeats' style changes over time?
2. Does PCA group the collections of Yeats' poetry by style phase similar to those suggested by literary scholars?

To carry out our multivariate analysis of Yeats' style, we began by downloading the poems from the CSUN website in plain text format, one poem per file, using a web-scraper extension to *Google Chrome*. This gave us a total of 307 files spread across 13 volumes (see Table 35.1). To carry out the computational analysis, we merged the poems associated with each volume into one file so that there was just one file per volume, giving 13 files in total.

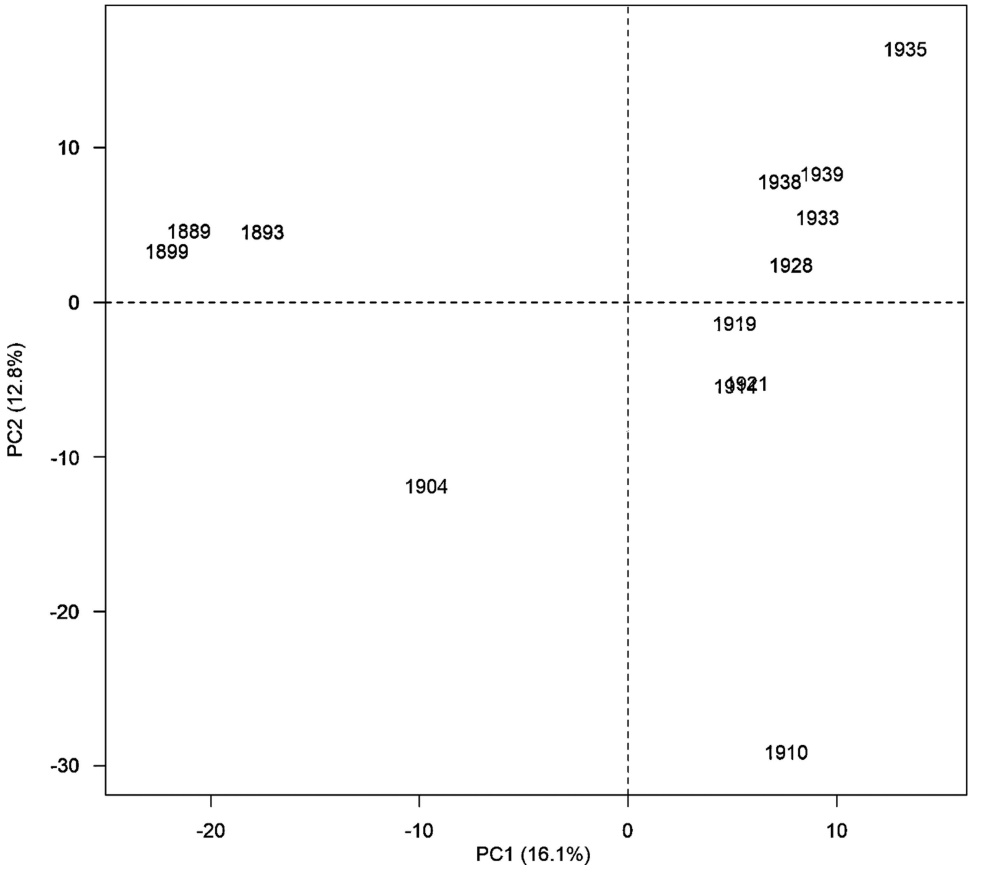

Figure 35.4 PCA of 13 volumes of Yeats' poetry *n*MFW = 1000

Table 35.1 The *Yeats Corpus of Poetry*

Volume no.	*Year of publication*	*Volume title*	*No. of poems*	*Total words*
1	1889	*Crossways*	16	4193
2	1893	*The Rose*	23	4511
3	1899	*The Wind Among The Reeds*	37	4194
4	1904	*In The Seven Woods*	14	2231
5	1910	*The Green Helmet and Other Poems*	21	1925
6	1914	*Responsibilities*	31	5443
7	1919	*The Wild Swans at Coole*	40	9024
8	1921	*Michael Robartes and The Dancer*	15	3258
9	1928	*The Tower*	20	7644
10	1933	*The Winding Stair and Other Poems*	30	7664
11	1935	*Parnell's Funeral and Other Poems*	6	1951
12	1938	*New Poems*	35	5529
13	1939	*Last Poems*	19	4634
Totals			**307**	**62201**

Using *Stylo* for *R* we performed a cluster analysis and PCA based on the MFWs across the volumes, ranging from 100 to 1,000 in increments of 50. To calculate distance between volumes, we used Smith and Aldridge's (2011) cosine distance variant of Burrows's (2002) Delta statistic because this statistic has been found to provide more accurate and reliable results (see Smith and Aldridge 2011; Evert et al. 2017). Two of the results of the cluster analyses are shown in the form of dendrograms (see Figures 35.1 and 35.2). The dendrograms are for *n*MFW = 100 and *n*MFW = 1000, thereby showing two snapshots of the results at the extremes of the range used. The general patterns indicated by these two dendrograms are very similar. The first branch, which is indicated by a vertical line known as a clade (labelled X), separates out the 1889, 1899 and 1893 volumes from the rest. The second major branch (labelled Y) separates the 1904, 1910, 1914, 1919 and 1921 volumes from the 1928, 1933, 1935, 1938 and 1939 volumes. The horizontal lines on the dendrogram indicate distance between branches. So, there is a relatively large distance between the volumes from the 1800s and the rest. Where there is a large difference between cluster branches, it is fairly safe to assume that the clusters are showing a reliable view of differences and similarities in the data. Small differences are indicative of less reliable groupings. Notice that the relative distances (indicated by the scale on the x-axis) are smaller when *n*MFW = 1,000 compared with *n*MFW = 100. This suggests that greater difference lies in the first 100 MFWs.

The dendrograms suggest that there is a reliable separation between Yeats' early nineteenth-century romantic poems and the twentieth-century poems that followed them. They also suggest that the twentieth-century poems, while part of the same family, differ in style, forming two sub-groups.

Two of the results for the PCA analysis (*n*MFW = 100 and *n*MFW = 1,000) can be found in scatterplots shown in Figures 35.3 and 35.4. As with the dendrograms, the plots present the start and end points of the range of experiments we performed and indicate the general patterns associated with the data across the range of MFW used (100–1,000). The PCA scatterplots provide visualisations of the data from a different analytical perspective and help to clarify the patterns observed in the dendrograms.

The scatterplots plot the volumes of poetry based on their scores along principal component 1 (PC1), which is plotted on the x-axis, and principal component 2 (PC2), which is plotted on the y-axis. Distances based on *n*MFW 100 to 1,000 are therefore plotted in two dimensions (horizontal and vertical). The twentieth-century volumes are largely separated by PC2 (indicated by vertical distance). The noticeable exception is the 1904 volume, which differs from the nineteenth century and the rest of the twentieth century on both PC1 and PC2. Another noticeable outlier is 1910, which aligns with the rest of the twentieth-century volumes in relation to PC1 but differs in relation to PC2.

The graphs shown in Figures 35.3 and 35.4 are divided into quadrants. In both graphs, the 1889, 1893 and 1899 volumes occupy the top left quadrant, although this pattern is more definite and the volumes closer together in Figure 35.4 (where *n*MFW = 1,000). The 1928–39 volumes occupy the top right quadrant, although when *n*MFW = 100, 1928 sits on the dividing line between the top right and bottom right quadrants. The 1935 volume is distanced vertically from other volumes in that quadrant. The bottom right quadrant contains the 1910, 1914, 1919 and 1921 volumes. The 1910 volume is at some vertical distance from the other volumes in that quadrant, with the latter being closer to top right quadrant. The 1904 volume sits alone in the bottom left quadrant.

Figure 35.4, which is the scatterplot for *n*MFW = 1,000, shows that the distances between volumes are closer than when *n*MFW = 100 (Figure 35.3). The 1914 and 1921

volumes are the most similar, since they occupy the same space on the graph. Closely related is the 1919 volume. The 1933, 1938 and 1939 volumes form a reasonably close grouping, with the 1928 volume falling in between but being closer to the 1914, 1919 and 1921 volumes. These seven volumes differ mostly in relation to PC2 and are similar in relation to PC1, meaning that they are fairly well aligned horizontally, but less well aligned vertically.

The graph shows three outliers: 1904, 1910 and 1935. All three are shown to be a large distance from other volumes of poetry. All three volumes are drastically different from the rest in terms of PC2. The 1910 and 1935 volumes are similar to the other twentieth-century volumes in terms of PC1, while the 1904 volume falls in between the nineteenth and twentieth century.

Stylo for *R* can also create a scatterplot that includes the MFW used in the analysis. The position of the words on the plot indicates loadings, which is their ability or power to differentiate texts. Figure 35.5 shows a scatterplot with loadings for *n*MFW = 200 (any more words than this and the plot starts to become unreadable). It is still possible to

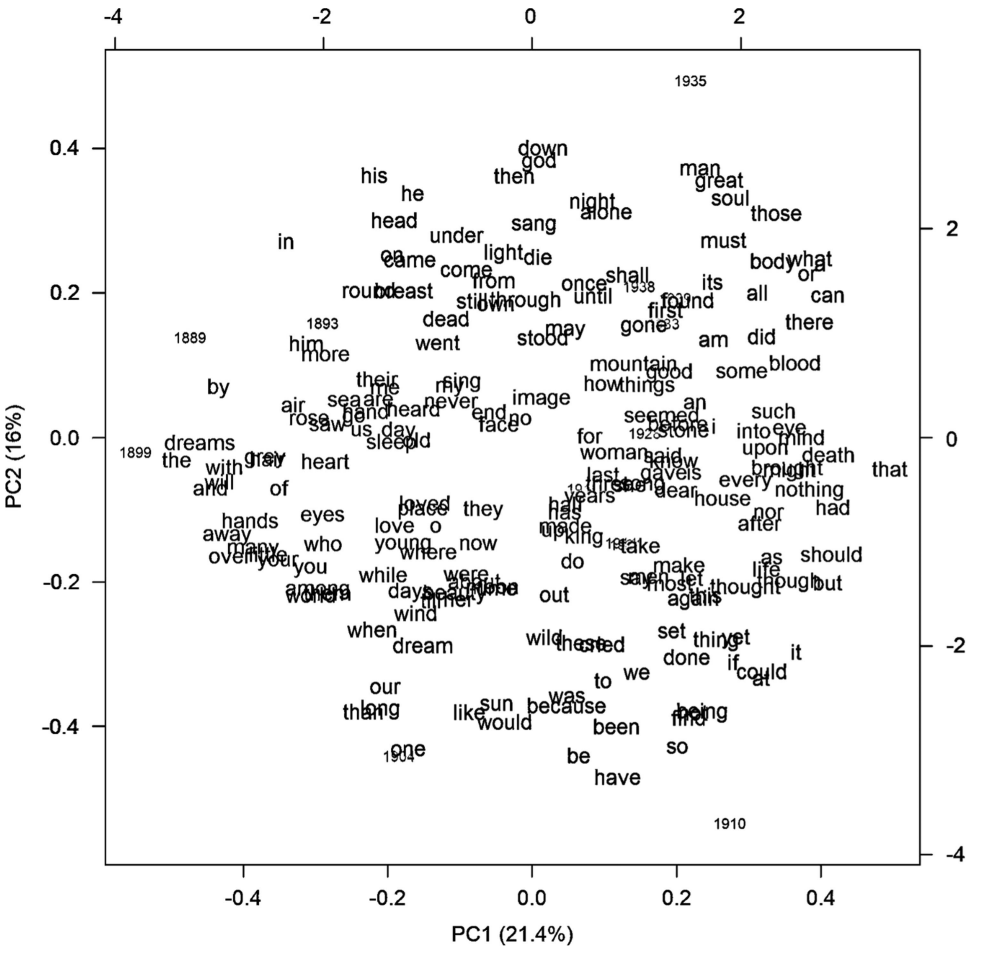

Figure 35.5 PCA of 13 volumes of Yeats' poetry *n*MFW = 200 with loadings

see in Figure 35.5 the general pattern of Yeats' volumes and to discern the words that exert a strong influence on the principal components. So, we can see that PC1 is influenced by *the* and *dreams* on the left-hand side of the chart and *that* on the right-hand side. (It is interesting to note that Meir 1974: 82 claims that 'Yeats' use of *that* and *what* most consistently marks his later style as distinctively Anglo-Irish'. It is indeed the case that Yeats on several occasions in his late works uses *that* instead of *who* and *what* instead of *which* and *whom*). PC2 is influenced by *down* at the top of the chart and *have* at the bottom.

As we mentioned at the start of this section, the cluster analysis and PCA experiments we carried out used a range of *n*MFW from 100 to 1,000 in increments of 50. Our results and discussion focussed on the start and end points of that range. However, as Eder (2017) points out, the choice of the number of MFW is completely arbitrary, so we could just as easily have looked at the results for *n*MFW = 101 and *n*MFW = 997. Eder also notes that small variances in the number of MFW can produce unpredictable and inexplicable differences in results (e.g. *n*MFW = 100 could produce different clusters from, say, *n*MFW = 103). This presents the researcher with the problem of working out which *n*MFW to use and which results to report, which might lead to the dendrograms that best fit the hypothesis being "cherry picked" and those that do not being ignored (see Rudman 2003).

In order to address some of these problems, Eder and his colleagues developed bootstrap consensus analysis which compiles the results of multiple PCA analyses across a range of MFW into a single tree diagram. Following the recommendations of Eder, we used *Stylo* to carry out such an analysis and generated a consensus tree for *n*MFW = 100 to 1000 in increments of 50. The resulting consensus tree, which can be seen in Figure 35.6, shows the general patterns for clustering across the range of MFW for the 13 volumes of poetry.

The tree reaffirms the general picture of there being three main groupings of Yeats' poetry: [1889, 1889, 1893]; [1904, 1910, 1914, 1919, 1921]; and [1928, 1933, 1935, 1938, 1939]. There are sub-groupings within these major groups, and some volumes within the groups are closer than others. With the exception of the 1928 volume, the groups roughly align with Carter and McRae's (2017) early, middle and later periods. In this analysis the 1928 volume is closer to the 1933 volume and is placed within the later period more generally. This particular result suggests that the style of the 1928 volume is a potential area for further research.

Eder (2017: 51) suggests that these kinds of visualisations 'speak for themselves', meaning that simply eye-balling, say, the dendrograms shows that Yeats' early poems in his first three volumes are substantially different from the rest (in some linguistic way or another based on the most frequent words). The visualisations, informative as they are, nevertheless leave unanswered questions about the nature and development of Yeats' style over time. Other than frequencies of the most frequent words in Yeats' poetry, we are no nearer to understanding what linguistic features and structures might be important in, for example, Yeats' lyrical style. The MFW analyses and subsequent visualisations do, however, offer us some sufficiently reliable groupings that allow us to confirm Carter and McRae's (2017) delimitation of the distinct periods of Yeats' writing life. Furthermore, these groupings (and, of course, the outliers) point to potentially profitable avenues of qualitative investigation. This is what we turn to in the next section.

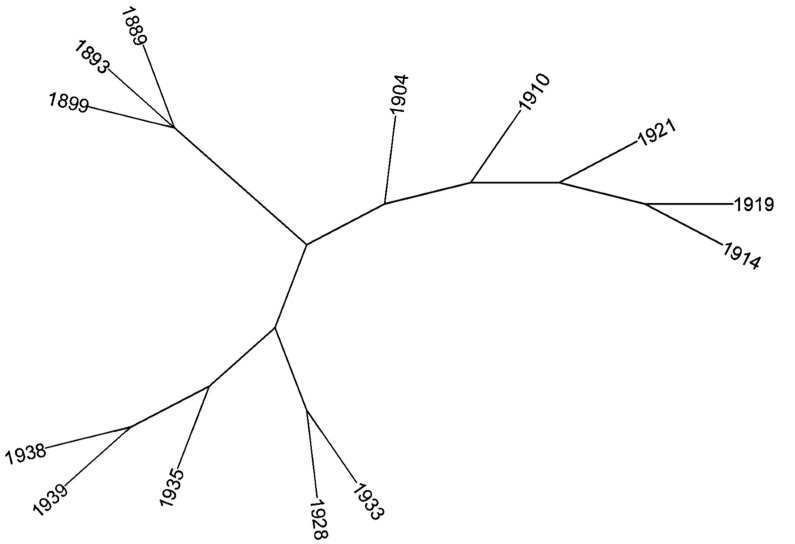

Figure 35.6 Bootstrap consensus tree of Yeats' poetry in 13 volumes based on *n*MFW = 100 to 1,000 in increments of 50

Table 35.2 The *Early Yeats Corpus*: volumes included and totals

Volumes included	*Year of publication*	*No. of poems*	*No. of words*
Crossways	1889	16	4193
The Rose	1893	23	4511
The Wind Among The Reeds	1899	37	4194
Totals		**76**	**12898**

5 Keywords and style

On the basis of the PCA results, we performed keyword comparisons on Yeats' poetry from the 1800s (early Yeats) and from 1928 to 1939 (late Yeats). We created two corpora containing the relevant collections, the details of which are shown in Tables 35.2 and 35.3.

We generated two sets of keywords by comparing the two corpora against each other using *AntConc* (Anthony 2019). We used a log likelihood cut-off 10.83 ($p < 0.001$) and a log ratio lower limit of 1. We also dismissed any keywords with a frequency less than 20 on the grounds that low-frequency keywords are less likely to reveal general style features across the period in question. The results are shown in Tables 35.4 and 35.5.

The keywords in both tables can be divided into different types based, to some extent, on grammatical class. There are those which carry meaning relating to content or theme (typically nouns, adjectives and lexical verbs) and those which relate to structuring of content and themes (typically grammatical words such as pronouns, prepositions, conjunctions and auxiliary verbs). Both can be said to be important for discerning style. That is, the style of Yeats' romantic poems is likely to concern a combination of themes and the manner in which those themes are presented.

Table 35.3 The *Late Yeats Corpus*: volumes included and totals

Volumes included	*Year of publication*	*No. of poems*	*No. of words*
The Tower	1928	20	7644
The Winding Stair and Other Poems	1933	30	7664
Parnell's Funeral and Other Poems	1935	6	1951
New Poems	1938	35	5529
Last Poems	1939	19	4634
Totals		**110**	**27422**

Table 35.4 Keywords in the *Early Yeats Corpus*

		Early Yeats		*Late Yeats*			
Rank	*Keyword*	*Freq.*	*%*	*Freq.*	*%*	*LL*	*LR*
1	sad	20	0.16	0	0.07	45.61	6.410
2	stars	33	0.26	1	0.12	67.03	6.133
3	sorrow	22	0.17	2	0.08	37.95	4.548
4	heavy	20	0.16	2	0.07	33.75	4.410
5	leaves	30	0.23	4	0.11	46.88	3.995
6	grey	22	0.17	4	0.08	30.93	3.548
7	little	27	0.21	8	0.10	30.12	2.843
8	rose	33	0.26	10	0.12	36.34	2.811
9	white	24	0.19	8	0.09	24.91	2.673
10	will	35	0.27	12	0.13	35.68	2.633
11	while	23	0.18	9	0.08	21.36	2.442
12	wandering	20	0.16	8	0.07	18.27	2.410
13	hair	34	0.26	14	0.12	30.39	2.368
14	dreams	22	0.17	11	0.08	16.64	2.088
15	you	104	0.81	56	0.38	73.39	1.981
16	go	22	0.17	12	0.08	15.27	1.963
17	dream	25	0.19	14	0.09	16.88	1.925
18	your	62	0.48	40	0.23	35.65	1.721
19	world	31	0.24	20	0.11	17.80	1.721
20	away	36	0.28	26	0.13	17.81	1.558
21	who	43	0.33	33	0.16	19.46	1.470
22	with	141	1.09	115	0.51	58.24	1.382
23	him	41	0.32	36	0.15	14.83	1.276
24	me	60	0.47	60	0.22	16.73	1.088

The keywords that carry semantic content (keywords 1–9, 12–14, 17, 19) potentially reflect thematic attention on particular emotions and flora. It is possible that the over-representation of the lexical items *dream* and *dreams* in the Early Yeats corpus is partly what motivated Luebering's (2011: 168) comment about the poems having a 'dreamlike atmosphere'. Here, for reasons of space, we will look more closely at just a small number of grammatical keywords. Focusing on grammatical keywords, highlighted in grey on the tables, is motivated by the fact these are more likely to suggest general language

Table 35.5 Keywords in the *Late Yeats Corpus*

	Late Yeats		*Early Yeats*			
Keyword	*Freq.*	*%*	*Freq.*	*%*	*LL*	*LR*
mountain	32	0.12	0	0.00	24.68	4.912
such	51	0.19	1	0.01	31.75	4.584
back	20	0.07	0	0.00	15.42	4.234
tomb	20	0.07	0	0.00	15.42	4.234
nothing	32	0.12	1	0.01	18.00	3.912
should	31	0.11	1	0.01	17.29	3.870
run	27	0.10	1	0.01	14.48	3.667
might	25	0.09	1	0.01	13.08	3.556
blood	47	0.17	2	0.02	24.11	3.466
keep	22	0.08	1	0.01	11.02	3.371
dear	38	0.14	2	0.02	17.99	3.160
soul	61	0.22	4	0.03	26.13	2.843
house	28	0.10	2	0.02	11.46	2.719
what	231	0.84	17	0.13	93.37	2.676
every	36	0.13	3	0.02	13.45	2.497
death	33	0.12	3	0.02	11.64	2.371
those	82	0.30	8	0.06	27.52	2.269
can	110	0.40	12	0.09	33.81	2.108
mind	51	0.19	6	0.05	14.66	1.999
did	42	0.15	5	0.04	11.94	1.982
thought	65	0.24	8	0.06	17.91	1.934
found	40	0.15	5	0.04	10.85	1.912
could	46	0.17	6	0.05	11.96	1.850
body	44	0.16	6	0.05	10.92	1.786
or	274	1.00	40	0.31	63.34	1.688
an	92	0.34	14	0.11	20.14	1.628
if	75	0.27	13	0.10	13.79	1.440
had	144	0.53	25	0.19	26.45	1.438
it	166	0.61	29	0.22	30.23	1.429
that	714	2.60	130	1.01	123.89	1.369
great	81	0.30	15	0.12	13.46	1.345
man	205	0.75	43	0.33	27.46	1.165
s	315	1.15	74	0.36	33.3	1.002

patterns across corpora that relate to style, since they do not of themselves indicate aboutness (that is, they are markers of style rather than of theme). In this respect they constitute what Enkvist (1973) terms style markers.

The keyword *you* is one of four personal pronouns in the keywords list. In order to assess whether the over-representation of *you* in the Early Yeats corpus was interpretatively significant in comparison to Late Yeats, we exported the concordance lines for *you* from both corpora into an Excel spreadsheet and analysed the context of use and referent of the pronoun. Forty-four per cent of the 104 occurrences of *you* in Early Yeats are part of interactions within poems (as in, for example, "Anashuya and Vijaya", "Fergus and the Druid"), as demonstrated by example 1. Most of the remaining

occurrences (55 per cent) are of the poetic persona (or poet) addressing someone or something via the poem (see example 2). Just 1 per cent of occurrences are of *you* being used generically (example 3).

(1) The sick man's wife opened the door:
'Father! you come again!'
('The Ballad of Father Gilligan')

(2) Were you but lying cold and dead,
And lights were paling out of the West,
You would come hither, and bend your head,
('He Wishes His Beloved Were Dead')

(3) Gay bells or sad, they bring you memories
('The Dedication To A Book Of Stories Selected From The Irish Novelists')

By contrast, 21 per cent of the occurrences of *you* in the poems in the Late Yeats corpus are in interactions in the poems (for example "The Three Bushes" and "Colonel Martin"), with 75 per cent being used to address someone or something via the poem. The remaining 4 per cent are generic *you*.

A further difference in the use of *you* between the poems in the two corpora is that in the Early Yeats corpus there are more occurrences of *you* used when the poetic persona (or poet) apparently addresses directly non-sentient objects and supernatural beings (for example Ireland, a fish, a deer, the heart, a valley, an angel) and unnamed individuals (potential lovers) than in the Late Yeats corpus. By contrast, in the Late Yeats corpus the poetic persona apparently directly addresses named people (Anne Gregory, Dorothy Wellesley, Von Hugel) and groups of people (*you, that have grown old* "The New Faces"; *Parnellites* "Come gather Round Me Parnellites"; *you that would judge me* "The Municipal Gallery Revisited"; *You that Mitchel's prayer have heard* "Under Ben Bulben") and also the reader or people generally, as example 4 demonstrates:

(4) Swift has sailed into his rest;
Savage indignation there
Cannot lacerate his breast.
Imitate him if you dare,
World-besotted traveller; he
Served human liberty.
('Swift's Epitaph')

This quantitative analysis of the keyword *you* shows that as well as an over-representation of the pronoun in the Early Yeats corpus, there is a change in the poetic stance of the poet in relation to his addressees. This does not go unnoticed by Ellman (1964), who suggests that 'Yeats altered the position of the reader' from the reader being 'almost an intruder on the poet's contemplations' to one of over-hearer or direct addressee.

As for the keywords *while* and *who* and *with*, these can all be used to connect additional information in the form of modifying phrases or clauses to other parts of the

sentence in which they occur. In both the Early and Late Yeats corpora all but one instance of *while* acts as a subordinating conjunction as demonstrated in example 5:

(5) come near me while I sing the ancient ways
('The Rose Upon The Rood Of Time')

The preposition *with* is used to provide additional information to verb phrases in the form of adverbials, as in (6):

(6) And blame you with many bitter words.
('The Fish')

Example 7 demonstrates that *with* is also used in the post-modification of noun phrases, as is *who*; both being used to post-modify 'a glimmering girl':

(7) It had become a glimmering girl
With apple blossom in her hair
Who called me by my name and ran
And faded through the brightening air.
('The Song Of Wandering Aengus')

Who is also used in questions, which is a topic we address in more detail later.

These sorts of uses also occur in the Late Yeats poems, but the frequency is less. This points to the more frequent use of adjuncts in poems in the Early Yeats corpus than in the Late Yeats corpus.

Turning now to the keywords in the Late Yeats corpus, Table 35.5 shows that the modal auxiliary verbs *should*, *might*, *can* and *could* are all key in the Late Yeats corpus and occur only a handful of times in the Early Yeats corpus. These keywords, along with *if* and *what*, relate to a comparative prevalence of evaluations and judgements of possibility, probability, beliefs and knowledge and present a view of how the poet sees the world. For example, the use of *can* in example 8 (emphatically) presents strong possibility. The choice of modal form (and indeed, the decision to use a modal form at all) nevertheless makes explicit the poetic persona's opinion and judgement.

(8) Nothing but stillness can remain when hearts are full
Of their own sweetness, bodies of their loveliness
('Meditations In Time Of Civil War')

The same is true of the modal auxiliary verb *should* which, even though prototypically associated with deontic modality, is used epistemically 18 times out of 31. Of the 13 deontic uses, 4 are in questions where the poetic persona asks 'why should […]?', apparently questioning existing states of affairs, as examples 9 and 10 demonstrate:

(9) Why should the imagination of a man
Long past his prime remember things that are
Emblematical of love and war?
('A Dialogue Of Self And Soul')

(10) Why should he think me cruel
Or that he is betrayed?
('A Woman Young And Old')

If is used in the construction of hypothetical scenarios in poems and for expressing possibility (e.g. examples 11). It is used 16 times, in combination with *what*, in questions (e.g. example 12).

(11) If Folly link with Elegance
No man knows which is which,
('The Old Stone Cross')

(12) What if I bade you leave
The cavern of the mind?
('Those Images')

The keyword *what* is predominantly used in interrogatives (62 per cent) in the Late Yeats poems, with the other occurrences being pronouns (28 per cent) or determiners (10 per cent). Questions in general are more prevalent in the Late Yeats corpus than in Early Yeats. We determined this by counting the question marks in each corpus and found 179 in the former and 33 in the latter. Even when the difference in the size of the corpora is taken into account (the Late Yeats being just over twice the size of Early Yeats), this five-fold difference suggests a difference in rhetorical style in Yeats' late poetry.

While the keyword analysis is brief, it does begin to suggest indicative linguistic characteristics of early and late style in Yeats' poetry. For instance, the comparative over-representation of modality and mood in the Late Yeats corpus suggests a development in one aspect of Yeats' style, whereby the poet is expressing doubts, suggesting what is possible or probable and questioning. Further keyness analysis, combined with other standard corpus linguistic techniques such as collocation and n-gram analysis, could be carried out to further explore the claims we have made in this section.

This chapter has aimed to demonstrate the value of combining stylometric methods of analysis with corpus stylistics in order to shed light on the changing nature of Yeats' poetic style. We first carried out a multivariate analysis of Yeats' poems using cluster analysis and PCA. In so doing, we focused on answering two questions:

1. Does multivariate analysis reveal lexical differences and/or similarities that support the literary critical claim that Yeats' style changes over time?
2. Does PCA group the collections of Yeats' poetry by style phase similar to those suggested by literary scholars?

In answer to these questions, the multivariate analysis provided some support for Carter and McRae's (2017) division of Yeats' poetry into three distinct stylistic periods. The secondary value of the cluster analysis and PCA was to motivate our corpus stylistic comparison of Yeats' poetry from the 1800s and that of 1928–39. By so doing, we were able to make some inroads into identifying some of the potential lexical characteristics of early and late Yeats. Of course, the keyword analysis presented here is only the beginning of what would need to be a more detailed and involved study, but it begins to

illustrate how corpus stylistic methods can be used in association with computational stylistic techniques to generate insights into style that would be impossible to uncover otherwise.

Further reading

Hoover, D., Culpeper, J. and O'Halloran, K. (2014) *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*, London: Routledge. (This book covers both corpus and computational approaches to stylistics and includes numerous case studies exemplifying some of the techniques discussed in this chapter.)

McIntyre, D. and Walker, B. (2019) *Corpus Stylistics: Theory and Practice*, Edinburgh: Edinburgh University Press. (This book outlines a theoretically informed approach to corpus stylistic analysis; chapter 7 deals with the corpus stylistic analysis of poetry particularly.)

Murphy, S. (2015) 'I Will Proclaim Myself What I Am: Corpus Stylistics and the Language of Shakespeare's Soliloquies', *Language and Literature* 24(4): 338–54. (This article demonstrates the value of keyword analysis in differentiating texts according to genre.)

References

Anthony, L. (2019) *AntConc 3.5.8*, Tokyo, Japan: Waseda University, Available at: https://www.laurenceanthony.net/software.

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

Burrows, J. (2002) '"Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship', *Literary and Linguistic Computing* 17(3): 267–87.

Burrows, J. (2003) 'Questions of Authorship: Attribution and Beyond', *Computers and the Humanities* 37: 1–26.

Carpenter, W. M. (1969) 'The "Green Helmet" Poems and Yeats' Myth of the Renaissance', *Modern Philology* 67(1): 50–9.

Carter, R. A. and McRae, J. (2017) *The Routledge History of Literature in English: Britain and Ireland*, 3rd edn, London: Routledge.

Davis, A. (2015) 'Edwardian Yeats: *In the Seven Woods*', *Études Anglais* 68(4): 454–67.

Eder, M. (2017) 'Visualization in Stylometry: Cluster Analysis Using Networks', *Digital Scholarship in the Humanities* 32(1): 50–64.

Eder, M., Rybicki, J. and Kestemont, M. (2016) 'Stylometry with R: A Package for Computational Text Analysis', *R Journal* 8(1): 107–121.

Ellman, R. (1964) *The Identity of Yeats*, New York, NY: Oxford University Press.

Enkvist, N. E. (1973) *Linguistic Stylistics*, The Hague: Mouton.

Evans, M. (2018) 'Style and Chronology: A Stylometric Investigation of Aphra Behn's Dramatic Style and the Dating of *The Young King*', *Language and Literature* 27(2): 103–32.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., & Vitt, T. (2017). 'Understanding and Explaining Delta Measures for Authorship Attribution', *Digital Scholarship in the Humanities* 32(2): 4–16. https://doi.org/10.1093/llc/fqx023

Finneran, R. J. (1997) *The Collected Works of W. B. Yeats. Volume 1: The Poems*, 2nd edn, New York, NY: Scribner.

Gabrielatos, C. (2018) 'Keyness Analysis: Nature, Metrics and Techniques', in C. Taylor and A. Marchi (eds) *Corpus Approaches to Discourse: A Critical Review*, London: Routledge, pp. 225–58.

Hardie, A. (2014) 'Log Ratio: An Informal Introduction', (Blog post), *ESRC Centre for Corpus Approaches to Social Science (CASS)*, Available at: http://cass.lancs.ac.uk/log-ratio-an-informal-introduction.

Hoover D. (2003) 'Frequent Collocations and Authorial Style', *Literary and Linguistic Computing* 18(3): 261–86.

Hoover, D. (2013) 'Textual Analysis', in K. Price and R. Siemens (eds) *Literary Studies in the Digital Age: An Evolving Anthology*, New York, NY: MLA Commons, Available at: https://dlsanthology.mla.hcommons.org/textual-analysis/.

Hoover, D. and Hess, S. (2009) 'An Exercise in Non-Ideal Authorship Attribution: The Mysterious Maria Ward', *Literary and Linguistic Computing* 24(4): 467–89.

Jeffares, N. A. (1968) *A Commentary on the Collected Poems of W. B. Yeats*, London: Palgrave Macmillan.

Levin, J. R. and Robinson, D. H. (1999) 'Rejoinder: Statistical Hypothesis Testing, Effect-Size Estimation, and the Conclusion Coherence of Primary Research Studies', *Educational Researcher* 29(1): 34–6.

Luebering, J. E. (ed.) (2011) *English Literature from the 19th Century Through Today*, New York, NY: Britannica Educational Publishing.

Mahlberg, M. (2012) *Corpus Stylistics and Dickens's Fiction*, London: Routledge.

Mastropierro, L. (2018) *Corpus Stylistics in Heart of Darkness and its Italian Translations*, London: Bloomsbury.

Matthews, S. (2014) 'W. B. Yeats', in D. E. Chinitz and G. McDonald (eds) *A Companion to Modernist Poetry*, Oxford: Wiley Blackwell, pp. 335–47.

McIntyre, D. and Walker, B. (2019) *Corpus Stylistics: Theory and Practice*, Edinburgh: Edinburgh University Press.

Meir, C. (1974) *The Ballads and Songs of W. B. Yeats: The Anglo-Irish Heritage in Subject and Style*, London: Macmillan.

Montoro, R. and McIntyre, D. (2019) 'Subordination as a Potential Marker of Complexity in Serious and Popular Fiction: A Corpus Stylistic Approach to the Testing of Literary Critical Claims', *Corpora* 14(1): 275–99.

O'Halloran, K. (2007) 'The Subconscious in James Joyce's "Eveline": A Corpus Stylistic Analysis that Chews on the "Fish Hook"', *Language and Literature* 16(3): 227–44.

Rudman, J. (2003) 'Cherry Picking in Nontraditional Authorship Attribution Studies', *CHANCE* 16(2): 26–32.

Sarker, S. K. (2002) *W. B. Yeats: Poetry and Plays*, New Delhi: Atlantic Publishers and Distributors.

Semino, E. and Short, M. (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*, London: Routledge.

Smith, W. H. and Aldridge, W. (2011) 'Improving Authorship Attribution: Optimizing Burrows' Delta Method', *Journal of Quantitative Linguistics* 18(1): 63–88.

Thompson, B. (1999) 'Improving Research Clarity and Usefulness with Effect Size Indices as Supplements to Statistical Significance Tests', *Exceptional Children* 65(3): 329–37.