IDENTIFYING DISTRIBUTIONAL PATTERNS IN EIGHTEENTH-CENTURY PERIODICAL ESSAYS

Paul Brocklebank

Abstract

This paper investigates the distribution of words and clusters within a single corpus and across a pair of related corpora. With a corpus containing Samuel Johnson's periodical essays as the target corpus and a corpus of Addison's essays as the reference corpus, it is shown how standard techniques for identifying keywords can be extended to identifying distributional tendencies within texts at the levels of sentence, paragraph and whole section/essay. Supplementing the investigation with collocational and concordance data, the main keywords, including TO at sentence, AND at paragraph, BY at essay level, and the main three-word clusters at the various levels, are discussed. It is argued that the methods described are useful additions to the corpus stylistic researcher's arsenal of techniques.

Key words

corpus stylistics, lexical distributions, eighteenth-century essays, keywords, key clusters, Samuel Johnson

1 Introduction

While in recent years the investigation of keywords and strings of words (also referred to as 'n-grams' or 'clusters') has become prominent in corpus stylistic analyses of literary texts (cf. e.g. Mahlberg 2007, 2009, Culpeper 2009, Fischer-Starcke 2009), less attention has been paid to patterns in the distribution of keywords and clusters across texts of the same genre. Taking the periodical essays of Samuel Johnson as the target of enquiry, the current paper outlines an approach to examining distributions of keywords and n-grams across a collection of his essays, by using the concept of keyness to examine patterns across segments of sentences, paragraphs and whole essays.

Section 2 of this paper outlines the methodological approach adopted in the study. The two sections that follow present and discuss the results of a keyness analysis across the sub-parts of Johnson's essays, dealing with single words in Section 3 and three-word clusters in Section 4. The final section comments on the methodology and results, and points to ways in which the research can be extended.

Discourse and Interaction 8/1/2015, pp. 5-19
ISSN 1802-9930
DOI: 10.5817/DI2015-1-5

2 Methodology

The investigation discussed in this paper is an outgrowth of a larger project that uses techniques from corpus linguistics to explore stylistic differences across three sets of eighteenth-century periodical essays, those of Joseph Addison, Jonathan Swift and Samuel Johnson. In brief, the main aim of the project is to compare one 'target' set of essays with the other two sets, which together act as the reference corpus for the analysis. The research examines lexical differences between sets of essays, with the focus on identifying and exploring the keywords, the key collocates of those keywords, and the main three-word or four-word clusters. The first stage comparing Johnson with Addison and Swift is discussed in Brocklebank (2013).

Where the previous research stage concentrated on keyness across the Johnson corpus treated as an undifferentiated whole, the current study compares the distribution of lexical occurrences at the sentence, paragraph and essay level with distributions of the same words in Addison's essays. Comparisons have often been made between the prose style of Johnson and Addison, either negatively in Bate's contrast between the 'elephantine, ponderous' style of the former and 'the light and carefree touch' of the latter (Bate 1955: 29), or with greater stress on similarities, as when Rogers talks of Johnson's 'direct emulation' of Addison (Rogers 1993: 120). The current study, and the project of which it is a part, is an attempt to throw some corpus linguistic light on this comparison.

The Johnson corpus consists of his contributions to *The Rambler* (1750-52), *The Adventurer* (1752-54) and *The Idler* (1758-60), and the Addison corpus is composed of his *Spectator* essays (in two series, 1711-12 and 1714). Although the *Adventurer*, *Idler* and *Spectator* essays were all available as downloads from Project Gutenberg (http://www.gutenberg.org/), the *Rambler* essays were sourced from the website of the Electronic Text Center at the University of Virginia in 2012 (these have since been removed from the site). In both cases, the essays were pre-edited to remove any material not written by the two authors, such as essay-initial mottos and quotations from other authors.

The size of the two corpora is shown in Table 1:

Corpus	**Number of essays**	**Number of tokens**
Johnson	323 (*The Rambler* 203, *The Adventurer* 29, *The Idler* 91)	434,344 (*The Rambler* 295,625, *The Adventurer* 46,532, *The Idler* 92,187)
Addison	255	323,841

Table 1: Composition of the two corpora

It was decided to focus the current investigation on the distribution of those top keywords in Johnson's essays that were generated with just the Addison corpus as the reference corpus. These were identified in Brocklebank (2013) using *WordSmith* (cf. Scott 2007) as the corpus tool, and with log-likelihood as the statistical measure for determining keyness. The top twenty keywords in descending order were: BY, TO, YET, WITHOUT, CAN, AND, BE, OR, NO, ALWAYS, WHOSE, POWER, SCARCELY, EASILY, HAPPINESS, HOPES, LIFE, MYSELF, NOR and EQUALLY. Many of these words are functional in nature, and as such also occur relatively often in the Addison corpus. For example, BY occurs 5,291 times in Johnson and 1,999 times in Addison, and TO occurs 15,397 times in the former, and 8,384 times in the latter. This means that we have enough data to carry out a meaningful distributional comparison between the two corpora, even though the total sizes of the two corpora as shown in Table 1 are relatively small.

During the planning stages of the project, one corpus tool for analyzing distributions that was considered for adoption was *WordSkew*, which has been developed by Michael Barlow at the University of Auckland (cf. Barlow 2014 for an example of *WordSkew* in use). However, at the time of writing only a 'demo' version of this software is available (from http://www.wordskew.com/) with certain limitations. Although with *WordSkew* a researcher can investigate the distribution of words and phrases in a corpus at the sentence, paragraph and section level, in its currently accessible version the user must pre-select single words and phrases to be analyzed, and cannot analyze all of the words in a particular corpus. More importantly, there is no way to compare distributions across different corpora; the researcher is restricted to obtaining comparisons within a single corpus.

Therefore, I decided to write my own small programs and use other existing resources to overcome these restrictions. Using the R programming language, I composed three text splitters to divide each sentence, paragraph and essay into three equal parts. The text splitters treated strings of text delimited by full stops, exclamation marks and question marks as sentences, and stretches of text terminating in carriage returns as paragraphs. The essays were also collected together in a single file, but manually tagged to divide the end of one essay from the beginning of another. The R program worked through the corpus file counting the number of tokens in each unit (sentence, paragraph or essay), dividing the number by three – rounding up to the nearest whole number when necessary – and then assigning each 'start', 'middle' and 'end' to its own sub-corpus. The *WordSmith KeyWords* program was subsequently used to calculate the keywords for the various sub-corpora, with other necessary calculations carried out manually.

3 Distributions of single lexical items

Comparisons between sub-parts of texts can be 'internal' or 'external'. An 'internal' comparison involves calculating differences between sub-parts of a single text or single set of texts, e.g. the starts of Johnson's paragraphs compared with the middles and ends of his paragraphs. On the other hand, an 'external' comparison would look at differences between the same sub-parts of different texts – Johnson's sentence starts and Addison's sentence starts, for instance.

For the 'internal' comparison, therefore, two wordlists were generated: one for a corpus consisting of a collection of the sub-parts to be investigated, and another for a (reference) corpus consisting of the remaining text. Using the *WordSmith KeyWords* program keywords were calculated for each sub-part. Table 2 gives the top keywords for each position with their keyness scores.

Level	**Start (keyness score)**	**Middle (keyness score)**	**End (keyness score)**
Sentence	I (839.65)	AND (175.27)	OR (241.39)
Paragraph	THIS (288.56)	HE (16.88)	AND (198.75)
Essay	RAMBLER (62.80)	HE (26.73)	&C (97.09)

Table 2: Internal keyword distributions for the Johnson corpus

For the starts of sentences the topkeyword was I, which had the highest keyness score (839.65) for any position. For the middles and ends of sentences the top keywords were AND and OR respectively. HE was the top keyword for mid-position in both paragraphs and essays, albeit with very low keyness scores. The words RAMBLER and &C (for 'etc.') are the keywords for the starts and ends of essays; this is a reflection of the fact that many of Johnson's essays took the form of letters addressed to *The Rambler* (although actually written by Johnson himself), with 'To the Rambler' at the start of the letter and 'I am, &c' as part of the closing.

The results obtained from this investigation of internal differences in distributions were somewhat unrevealing. For example, the fact that the top keywords at the start of Johnsonian sentences include I, INDEED and THUS is unremarkable in that these words can be expected to favor this sentence position for any writer, not just Johnson. In addition, the keyness of RAMBLER and &C reflect genre-specific properties of the essays themselves rather than being signatures of a particularly Johnsonian style. More revealing would be a keyness

comparison of distributions with another writer, an examination of 'external' rather than 'internal' distributions.

The approach that was adopted involved a 'like with like' comparison with *The Spectator* essays of Joseph Addison. Each sub-part of a sentence, paragraph or essay in Johnson was compared with the equivalent sub-part in Addison; for instance, and again using *WordSmith*, a list of keywords for Johnson sentence starts was generated with the reference corpus for the comparison consisting of Addison sentence starts. Similar lists were produced for sentence middles and ends, and the process repeated for paragraphs and essays. This yielded a total of nine keyword lists. The top thirty keywords for each position were then tabulated with their keyness scores for the other positions at the same level. In order to pinpoint the more extreme variations in the data, the standard deviation (SD) for each of the keywords at the relevant level was calculated and the words ranked according to this SD. Table 3 gives the top ten keywords for Johnson vis-à-vis Addison at the sentence level (for each word the position with the highest keyness score is underlined).

Rank	Word	Start Keyness	Middle Keyness	End Keyness	SD Score
1	TO	285.64	133.83	167.57	65.09
2	YET	188.32	156.89	52.62	58.00
3	OR	53.47	90.19	157.24	42.96
4	CAN	83.42	179.79	99.76	42.11
5	BY	258.51	194.16	289.15	39.59
6	AND	165.48	78.61	120.92	35.47
7	HE	73.35	4.57	5.09	32.30
8	MUST	-20.13	33.50	51.94	30.57
9	WITHOUT	88.58	135.08	161.00	29.96
10	WANT	1.31	31.53	71.29	28.66

Table 3: Top ten key word distributions for the Johnson corpus vis-à-vis Addison at the sentence level

The word with the highest standard deviation score for keyness across the three positions at the sentence level was TO. When compared with Addison this

word is key across all positions – start, middle and end – but it is particularly conspicuous at the start of sentences with a keyness score of 285.64. An example of a sentence with a heavy use of TO in its first third is the following from The Rambler 27, where TO appears four times in 26 words, and is used either as an infinitive marker or a preposition:

(1) ***[I had resolution to throw off the splendour which reproached me to myself, and retire to an humbler state, in which I am now endeavouring to]*** *recover the dignity of virtue, and hope to make some reparation for my crime and follies, by informing others, who may be led after the same pageants, that they are about to engage in a course of life, in which they are to purchase, by a thousand miseries, the privilege of repentance.*

One possible reason for TO appearing so conspicuously at the beginning of Johnsonian sentences is the length of the sentences themselves. Mean sentence length for the Johnson corpus is 71.18, whereas for Addison mean sentence length is less than half that at 34.13. Longer sentence length points to more clauses in a single sentence, and where shorter sentences would have infinitives appearing near the ends of sentences, longer sentences would be more likely to have them appearing at the ends of clauses contained within the starts of sentences.

A similar preference for sentence-initial position is also evident for the word with the second highest SD score, YET, although to a lesser extent than for TO, as its score of 188.32 is still within one standard deviation of the mean. When analyzed internally, sentence initial is a preferred position for YET in Johnson's sentences (with a keyness score of 31.39), and it does seem to be a stylistic preference of his to use this word to mark contrast with a preceding sentence:

(2) *(An oppressor) destroys the only plea that he can offer for the tenderness and patience of mankind.* ***Yet****, even this degree of depravity we may be content to pity, because it seldom wants a punishment equal to its guilt.* (The Rambler 11)

On the other hand, OR, also key across all positions, is most prominent at the end of sentences, as we might expect from a typical Johnsonian sentence with its parallelism and contrasts, such as the following, where 'others' are contrasted with the subject ('himself'), 'amusements' are contrasted with 'studies', and there is the parallelism of the prepositional phrases headed by 'to':

(3) *And if he finds, with all his industry, and all his artifices, that he cannot deserve regard, or cannot attain it, he may let the design fall at once, and, without injury* ***[to others or himself, retire to amusements of greater pleasure, or to studies of better prospect.]*** (The Rambler 1)

Turning now to the paragraph keyword distributions (Table 4), we see that while the two words with the top SD scores, AND and WITHOUT, are key in all positions, their most prominent position is towards the end of a paragraph. As mentioned above, Johnson's style is famous for its extensive use of parallelism (cf. e.g. Wimsatt 1941), and this is perhaps what is reflected here. YET is also used by Johnson to a greater extent than Addison, particularly in the first third of a paragraph. Therefore, the current analysis has revealed that Johnson's style is marked by an inclination to use YET both sentence initially and paragraph initially.

Rank	Word	Start Keyness	Middle Keyness	End Keyness	SD Score
1	AND	52.17	95.61	221.14	71.64
2	WITHOUT	69.72	141.70	178.92	45.33
3	YET	172.71	145.08	77.52	39.98

Table 4: Top three key word distributions for the Johnson corpus vis-à-vis Addison at the paragraph level

To get a clearer idea of how the keywords are being used by Johnson, we can examine the collocates of the keywords, and for an even more detailed picture the concordances. By way of example, Table 5 presents collocational information for AND paragraph-finally in Johnson's essays.

+R1 Cluster (no. of hits)	+R2 Examples (3 or more hits)
AND THE (425)	other (5), last, next (4), danger, eye, French, general, happiness, heart, love, mind, most, only, same, soul (3)
AND THAT (209)	the (26), he (20), nothing, they, to (7), I, we (6), a, his, she (5), every, in, when (4), as, by, it, most, no, none, of, our, there (3)
AND TO (141)	the (8), have, leave, shew (4)
AND THEREFORE (67)	I (4), could, easily, no (3)

Table 5: The four main collocates of AND paragraph-finally in Johnson's essays

The total number of hits for AND in this position is 6147. On 425 occasions AND was followed by THE (i.e. in +R1 position), indicating coordination with a following noun phrase or a clause that contains a noun phrase initially. There seems to be quite a wide range of noun phrases in this position as the word that

occurs most often after AND THE (at +R2) – OTHER – only appears five times. The +R1 collocates of AND THAT (209 hits), such as THE and HE, indicate the coordination of a complement clause with THAT as a complementizer, rather than a noun phrase with THAT as a demonstrative. AND TO (141 hits) could involve prepositional (THE at +R2) or verbal coordination (HAVE, LEAVE, SHEW at +R2), although which use predominates is impossible to tell from just looking at the main +R2 collocates. The right collocates of AND THEREFORE (67 hits) also point to coordination of full clauses (with I and NO at +R2) or verb phrases (with COULD and EASILY at +R2), but again a clearer picture would require a detailed look at all of the +R2 collocates.

In addition to AND, another word that Johnson favors for coordinating noun phrases is WITHOUT, which appears 580 times towards the end of a paragraph and is very much a keyword in this position. In fact, AND is also one of the main collocates of WITHOUT, filling the +R2 position on 76 occasions, as in (4).

(4) *... the seriousness of Wit was* ***without*** *dignity* ***and*** *the merriment of Learning without vivacity.* (The Rambler 22)

Also identified with help from a concordance of paragraph-final examples with WITHOUT is the following example of a single paragraph, which gives a good illustration of AND and WITHOUT being used in tandem to produce Johnsonian parallelism.

(5) *In the time when Bassora was considered as the school of Asia, and flourished by the reputation of its professors and the confluence of its students, among the pupils that listened round the chair of Albumazar was Gelaleddin, a native of Tauris, in Persia, a young man amiable in his manners and beautiful* ***[in his form, of boundless curiosity, incessant diligence, and irresistible genius, of quick apprehension and tenacious memory, accurate without narrowness, and eager for novelty without inconstancy.]*** (The Idler 75)

Here, the paragraph moves towards its conclusion with a series of modified abstract nouns conjoined with AND, first a group of three ('boundless curiosity, incessant diligence AND irresistible genius'), then a group of two ('quick apprehension AND tenacious memory'). Finally, the paragraph ends with two conjoined phrases that consist of two adjectives post-modified with phrases beginning with WITHOUT ('accurate WITHOUT narrowness, AND eager for novelty WITHOUT inconstancy').

The final keyword analysis to be discussed is that for the starts, middles and ends of whole essays. The top three keywords ranked according to SD score are given in Table 6.

Rank	Word	Start Keyness	Middle Keyness	End Keyness	SD Score
1	BY	308.84	190.66	235.73	48.70
2	RAMBLER	88.08	11.14	26.75	33.21
3	CAN	117.42	82.41	154.73	29.36

Table 6: Top three keyword distributions for the Johnson corpus vis-à-vis Addison at the essay level

As mentioned at the beginning of this paper, the overall keyword for the Johnson essays identified in Brocklebank (2013) was BY. In fact, while scoring heavily for all three subdivisions of the essays, BY is used most heavily in the first of the three subdivisions. Compare this with the next most prominent keyword RAMBLER, which makes the top three only because it has a high keyness score for its use in the starts of the essays. Finally, the reason for CAN having the third highest SD score in this distributional analysis is that it appears relatively often towards the end of the essays.

An examination of the +R1 collocates of BY (cf. Table 7) points to conspicuous use either as the head of a prepositional phrase (BY THE, BY A, BY HIS, BY THIS) or as part of the complex relative pronoun BY WHICH.

+R1 Cluster (no. of hits)	+R2 Examples (3 or more hits)
BY THE (318)	same (6), death, general (4), desire, experience, name, sight, various (3)
BY WHICH (117)	the (27), he (9), all, I (5), a, it, we (4), some, they (3)
BY A (90)	long (4), thousand, very (3)
BY HIS (41)	own (9)
BY THIS (41)	method (4)

Table 7: The five main collocates of BY essay-initially in Johnson's essays

Closer inspection of the concordance lines for BY in the Johnson and Addison sub-corpora reveals a further difference between the two authors in the use of this word. Results of this concordance analysis are summarized in Table 8.

Author	Total Hits	Total No. Hits Per 1,000/w	BY in Long Passives	Other Uses of BY
Johnson	1855	12.79	1131 (60.97%)	724 (39.03%)
Addison	649	6.00	319 (49.15%)	330 (50.85%)

Table 8: BY in long passives essay-initially in Johnson's and Addison's essays

Unsurprisingly, Johnson uses BY much more often than Addison at 1,855 times compared to 649, and if this is converted to a score per 1,000 words we see that the use of BY in Johnson is more than double the frequency in Addison. Working through the concordances manually, I counted whether each instance of BY in the Johnson and Addison essay starts marked the agent of a passive clause – in other words, whether it was part of a long passive – or whether it was being used in some other way, for example, as a marker of instrumentality in an active clause, with a temporal meaning, with a locative meaning, or as part of a fixed phrase such as 'by degrees'.

The results show that Johnson uses BY to mark passive agents much more often than Addison, with 1,131 or nearly 61 per cent of Johnson's essay-initial BYs marking agents in long passives, as compared with 319 or just over 49 per cent for Addison. The log-likelihood score for BYs in Johnson's essays starts is 12.01, which is highly significant. Thus, the prominence of BY seems to be going hand in hand with the prominence of long passives in essay starts. If stylistically the use of the passive voice is seen as a marker of greater formality and writtenness, and if, as Biber et al. (1999: 943) claim, "long passives should be considered as competing with the corresponding active constructions rather than with short passives", then Johnson's penchant for long passives reflects the greater writtenness of his style when juxtaposed with that of Addison.

4 Distributions of three-word clusters

This method of exploring distributions across texts can also be used to investigate strings of words or clusters. There are, however, two complicating factors. Firstly, as sentences, paragraphs and essays are split into segments, clusters that lie across the boundaries between segments are themselves going to be split and will be invisible in the final results. Secondly, for smaller corpora such as the ones that we are dealing with here, the number of hits is going to be considerably smaller for clusters than for single words, and so the results are likely to be less impressive from the point of view of statistical significance.

Notwithstanding these potential drawbacks, the distributions of the main three-word clusters in Johnson's essays were examined in an attempt to draw out further patterns from the data, and in what follows I briefly discuss the results of this analysis. The *kfNgram* software tool (Fletcher 2002) was used to generate a list of three-word clusters and their frequencies. The forty clusters that appeared fifty or more times in the essays were selected, and log-likelihood scores were calculated for these clusters across the whole, non-segmented Johnson and Addison corpora. The twenty clusters that scored positive log-likelihood scores of 10.83 or more (i.e. those significant at the $p < 0.001$ level) are listed in (6) below.

(6) *OF THOSE WHO, THE POWER OF, IS TO BE, IT MAY BE, IT HAS BEEN, THE REST OF, FOR WANT OF, I KNOW NOT, THE HAPPINESS OF, THE NECESSITY OF, THE ART OF, TO THE RAMBLER, THAT I WAS, THE PLEASURE OF, BY WHICH THE, FOR A TIME, THAT HE WHO, THE GREATER PART, SUCH IS THE, THE DIGNITY OF*

The analysis then proceeds as before: the distributions of the clusters across the Johnson and Addison sub-corpora are compared using the *WordSmith KeyWord* program, and standard deviation scores are used to pick out which clusters diverged most from their average log-likelihood score. The figures for the three-word clusters with the highest SD scores for all three levels are given in Table 9.

Level	Key Cluster	Start Keyness	Middle Keyness	End Keyness	SD Score
Sentence	FOR WANT OF	-0.06	2.72	34.75	15.80
Paragraph	IT MAY BE	39.96	2.92	0.37	18.09
Essay	TO THE RAMBLER	61.31	1.11	1.11	28.38

Table 9: Top three key three-word cluster distributions for the Johnson corpus vis-à-vis Addison

The FOR WANT OF cluster scored highly at the ends of sentences, marginally positive for the middles, and negatively for the starts. By contrast, for paragraphs IT MAY BE was prominent in the starts, but was not significantly positive for middles and ends. The story was similar for essays: TO THE RAMBLER received the highest keyness score overall for the starts, but scored negligible positive scores for the two other positions.

Below are some examples of these clusters to demonstrate their use in context. First, FOR WANT OF is used in the sense of 'because there is not' or 'because somebody did not', as can be seen from the two examples (7) and (8). In both sentences Johnson begins by stating that something happened, and ends the sentences by giving the lack of something as reasons for these events.

(7) *Horace tells us with more energy that there were brave men before the wars of Troy, but they were lost in oblivion* ***for want of*** *a poet.* (The Rambler 143)

(8) *…; he was prudent, but suffered his affairs to be embarrassed* ***for want of*** *regulating his accounts at stated times.* (The Rambler 201)

The cluster IT MAY BE often appears early in a paragraph, as part of a passive structure followed by an infinitive ('it may be presumed to have') or a complement clause ('it may be observed that', 'it may be laid down that'), as in the following examples:

(9) *When an opinion to which there is no temptation of interest spreads wide, and continues long,* ***it may be*** *reasonably presumed to have been infused by nature or dictated by reason…* (The Idler 52, start of 4th paragraph)

(10) ***It may be*** *observed, perhaps without exception, that none are so industrious to detect wickedness, or so ready to impute it, as they whose crimes are apparent and confessed…* (The Rambler 76, start of 4th paragraph)

(11) ***It may be*** *laid down as an axiom, that it is more easy to take away superfluities than to supply defects…* (The Rambler 25, start of 4th paragraph)

Here Johnson softens the assertion of his presumption, observation and axiom by putting them forward as (impersonal) possibilities, which are then elaborated on in the remainder of the paragraph.

Finally, and as noted earlier, TO THE RAMBLER appears 55 times at the beginning of essays to mark letters addressed to *The Rambler*, although these letters are actually written by Johnson himself. As the formulaic introduction to such letters, this is only to be expected. The only other occasions in which the phrase appears are in *The Rambler* 10, where the writer quotes a missive from a 'Lady Racket' who sends her 'compliments to The Rambler', and at the end of *The Rambler* 12, where at the end of a letter 'Zosima', discussing her search for a position as a maid, comes under the protection of a woman called Euphemia, to whom she expresses her gratitude 'by giving this account to the Rambler.'

5 Conclusions

Standard corpus analytic techniques allow us to identify keywords in a text or set of texts, but tell us nothing about how these keywords are distributed across the text(s). In this paper we have explored an approach to corpus stylistic text analysis that not only works with segments of language at different textual levels (sentence, paragraph, essay), but also incorporates standard statistical methods of comparison to bring to the surface the distinctive lexical properties of the text, either by comparison within text(s) or, perhaps more fruitfully, by comparisons between different sets of texts. This method can be applied to single words or, with some limitations, to strings of words (clusters).

We can take AND as an illustrative example. A standard keyword analysis would allow us to identify AND as a keyword in Johnson's texts. But, if we then adopt a distributional perspective, we find that at the paragraph level the word becomes more prominent the nearer we are to the end of the paragraph. Table 10, which includes frequency data per 1,000 words of text, shows that while the frequency of AND increases like this in both Johnsonian and Addisonian paragraphs, the trend is more pronounced in the former. Parallelism, which has already been identified as a feature of Johnson's essays, often (but not always) involves the use of AND as the hinge which links parallel instances. The statistical patterns provide evidence, therefore, that this parallelism is likely to increase as we approach the end of a Johnsonian paragraph.

Corpus	Start Hits Per 1000/w	Middle Hits Per 1,000/w	End Hits Per 1,000/w
Johnson	30.88	37.05	42.64
Addison	26.04	29.97	31.31

Table 10: Frequency of AND at the paragraph level in Johnson and Addison

Almost all of the top keywords identified by the distributional analysis belonged to the list of top twenty keywords for the essays treated as a single, undifferentiated corpus. If we restrict ourselves to the top three keywords at the sentence, paragraph and essay level and ranked by standard deviation scores, only RAMBLER at the essay level is new. However, what the addition of a distributional perspective adds to an analysis is that it permits the investigator to refine the standard analysis by clarifying at a different level of detail the writer's usage of the keywords, particularly if this is supplemented with information on collocations and an examination of concordances.

If a writer's preference for certain lexical items and strings of words can be considered as part and parcel of that writer's style, the way in which a writer tends to order these items in a sentence, paragraph or longer textual element may also be important when investigating that style. Be that as it may, there are two ways in which the data obtained from such an investigation could be extended. First, the quantitative nature of the presentation leaves considerable room for further investigation of a more qualitative nature. A closer look at the concordances and the keywords as they appear in context would help contribute to this. Another way in which the research could be developed is to turn the spotlight away from the functional words that dominate the general and distributional keyword lists, and to focus on investigating the distribution of content words, an extension of the original research which has the potential to reveal other distinctive aspects of Johnson's periodical essays.

Note
This work was supported by JSPS KAKENHI Grant Number 25370563.

References

Barlow, M. (2014) 'Ordering of elements in learner corpora.' *Learner Corpus Studies in Asia and the World 2*, 127-136.

Bate, W. J. (1955) *The Achievement of Samuel Johnson*. New York: Oxford University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *The Longman Grammar of Spoken and Written English*. London: Longman.

Brocklebank, P. (2013) 'Johnson and the eighteenth-century periodical essay: A corpus-based approach.' *English Language Overseas Perspectives and Enquiries 10*, 21-32.

Culpeper, J. (2009) 'Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet.' *International Journal of Corpus Linguistics* 14/1, 29-59.

Fischer-Starcke, B. (2009) 'Keywords and frequent phrases of Jane Austen's Pride and Prejudice: A corpus-stylistic analysis.' *International Journal of Corpus Linguistics 14 /4*, 492-523.

Fletcher, W. H. (2002) *kfNgram*. Available at: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html (accessed November 2014).

Mahlberg, M. (2007) 'Clusters, key clusters and local textual functions in Dickens.' *Corpora 2/1*, 1-31.

Mahlberg, M. (2009) 'Corpus stylistics and the Pickwickian watering-pot.' In: Baker, P. (ed.) *Contemporary Corpus Linguistics*. London: Continuum. 47-63.

Rogers, P. (1993) 'The Rambler and the eighteenth-century periodical essay: A dissenting view.' In: Downie, J. A. and Corns, T. N. (eds) *Telling People What to Think: Early Eighteenth-Century Periodicals from The Review to The Rambler*. London: Frank Cass & Co. Ltd. 118-129.

Scott, M. (2007) *WordSmith Tools 4.0*. Oxford: Oxford University Press.

Wimsatt, W. K. (1941) *The Prose Style of Samuel Johnson*. New Haven: Yale University Press.

Paul Brocklebank is Associate Professor in the Department of Liberal Arts at Tokyo University of Technology, Japan. He does research in corpus linguistics and stylistics, and is particularly interested in applying techniques from the former in work on the latter. Current projects all involve stylistic investigations of media, whether historical (eighteenth-century periodical essays), traditional (English language print media in the Far East), or new media (the language of podcasts).

Address: Paul Brocklebank, Department of Liberal Arts, Tokyo University of Technology, Katakura-machi 1404-1, Hachioji-shi, Tokyo-to 192-0982, Japan. [e-mail: cpaulb@stf.teu.ac.jp]