Profiling Stylistic Variations in Dickens and Smollett through Correspondence Analysis of Low Frequency Words

Tomoji Tabata *(tabata@lang.osaka-u.ac.jp)*
Osaka University

The aim of this paper is to present the result of a corpus-driven, quantitative analysis of the style of Dickens in comparison with the style of Smollett. The particular problem discussed is the differing distribution of *-ly* adverbs in the texts written by the two authors. By applying a multivariate stylo-statistics model, this study illustrates how sharply the two authors differ in their uses of adverbs as well as how texts are differentiated according to genre and chronology within authorial groups.

On the relationship between linguistic registers and adverbs, Biber et al. (1999, 541) present interesting findings from a large-scale corpus:

> It is interesting to note that, overall, fiction ... uses many different descriptive *-ly* adverbs, although few of these are notably common (occurring over 50 times per million words). Rather, fiction shows great diversity in its use of *-ly* adverbs. In describing fictional events and the actions of fictional characters, writers often use adverbs with specific descriptive meanings.

In fact, *-ly* adverbs found in Dickens are quite diverse. In the 23 texts used in this study, the number of types amount to 1,728; Smollet employed 634 types. Among those, a few types are highly frequent, such as *really* and *certainly*, occurring more than one thousand times. Conversely, a large number of adverbs occur only once. Such *hapax legomena* include a few types which sound very much Dickensian, such as *evil-adverbiously*, *patientissamentally*, *Shakespearianly*. Although the number of tokens of *-ly* adverbs account for only a little more than 1% of total word-tokens in the texts, the findings by Biber et al. suggest that *-ly* adverbs deserve special attention in stylistic study of fiction.

This study deals with a corpus of texts comprising Dickens' and Smollett's major works. Dickens' set includes fifteen 'serial fictions', six 'sketches', one 'miscellany', and one 'history'. Smollett's contains six 'fictions' and one 'sketch'. The total word-tokens in the corpus amount to 5.8 million, with the Dickens component containing 4.7 million tokens and the Smollett component totalling 1.1 million word-tokens. The present project was initiated as a study based on a comprehensive collection, not a sample corpus, of texts by the targeted authors. Therefore, the imbalance in the number of texts as well as tokens is inevitable. However, due attention will be paid in the choice of variables to minimize a potential effect of the differences in the population of the two sets. All the texts in the corpus have been annotated with the *POS* tags, using Eric Brill's *Rule-Based Tagger* (also known as the *Brill Tagger*). Manual post-editing has been conducted to eliminate a number of ill-assigned tags.

In an early successful attempt at a computational description of literary style, Milic compared the style of Jonathan Swift with the writings of his contemporaries, with special reference to the relative frequencies of word-classes in the texts and to grammatical features such as seriation and connection. Cluett (1971 & 1976) adopted a similar approach to conduct a diachronic study of prose style across 4 centuries: from the16th to the 20th centuries. Brainerd's works (1979 & 1980) are ambitious attempts to apply discriminant analysis to the question of genre and chronology in Shakespeare plays. Takefuta's approach to text typology, or register variation, is among the first to successfully employ factor/cluster analysis to the lexical differences between registers. His pioneering work, however, is not widely acknowledged because it was written in Japanese. Since Burrows (1987) and Biber (1988), it has become popular practice to employ multivariate techniques in quantitative studies of texts. Biber carried out *factor analysis (FA)* on 67 linguistic features to identify co-occurring linguistic features that account for dimensions of register variation. A series of research projects based on Biber's *Multi-Feature/Multi-Dimensional* approach have been successful in elucidating many interesting aspects of linguistic variation, such as language acquisition, ESP, diachronic change of prose style, and differences between conversational styles in British and American English, to give a handful of examples (Biber & Finegan; Conrad & Biber eds.).

The Biber model is one of the most sophisticated approach by far. Yet it is not without its critics. Nakamura (1995) raises a major objection. He argues that Biber's variables are "quite arbitrarily selected with no definite criterion and mixed levels" (1995, 77-86). Further, Sigley (1997) notes that almost half of Biber's 67 linguistic features are too rare in texts of 2,000 words.

Burrows (1987), on the other hand, applied a *Principal Component Analysis (PCA)* to the thirty most common words in the language of Jane Austen. The method demonstrates that differing frequency patterns in these very common words show significant differentiations among Austen's characters, and that the statistical analysis of literary style may lead not only to a deeper understanding of the novel itself but may also contribute to our deeper appreciation of it. In this use of a PCA, the frequencies of common words are used as variables. The

Burrows method seems to have higher replicability and feasibility; since it focuses on common words, most of the variables are frequent enough to produce stable statistical results. In addition, it does not require a multi-layered tagging scheme optimised for Biber's MF/MD approach.

A particular strength of the Burrows methodology is in testing cases of disputed authorship and national differences in the English first-person retrospective narrative, known as 'history'. Among the most successful applications are Burrows (1989, 1992 & 1996), Craig (1999a, b, & c). The Burrows approach or similar methodology has been applied to Bible stylometry. Some scholars like Linmans, Merriam, and Mealand use *Correspondence Analysis (CA)* instead of PCA. In the context of text typology, Nakamura (1993) applied CA to the frequency distribution of personal pronouns to visualize association between personal pronouns and 15 text categories in the LOB corpus.

My earlier work (Tabata) also used CA to analyse the distribution patterns of Part-of-Speech in Dickens's 23 texts and identified a contrast between serial fiction and sketches. The present study is different from the Burrows model in that it extends the range of variables to include low-frequency words, or rare words, by applying CA in the analysis of *-ly* adverbs. CA is one of the techniques for data-reduction alongside PCA and FA. Unlike PCA and FA, however, CA does not require intervening steps of calculating correlation matrix or covariance matrix, and can therefore process the data directly to obtain solution. CA allows examination of the complex interrelationships between row cases (i.e., texts), interrelationships between column variables (i.e., adverbs), and association between the row cases and column variables graphically in a multi-dimensional space. It computes the row coordinates (word scores) and column coordinates (text scores) in a way that permutes the original data matrix so that the correlation between the word variables and text profiles are maximized. In a permuted data matrix, adverbs with a similar pattern of distribution make the closest neighbours, and so do texts of similar profile. When the row/column scores are projected in multi-dimensional charts like Figures 1 to 4, relative distance between variable entries indicates affinity, similarity, association, or otherwise between them. One advantage CA has over PCA and FA is that PCA and FA cannot be computed on a rectangular matrix where the number of columns exceeds the number of rows, a concern of the present study. Yet CA can handle such types of a data table with, for example, the row cases consisting of thirty texts and the column variables consisting of hundreds of adverbs.

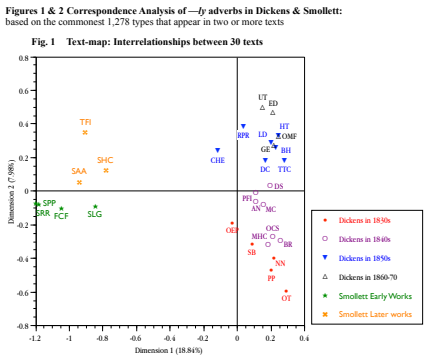

Figure 1: Correspondence Analysis of —ly adverbs in Dickens & Smollett based on 1,278 types that appear in two or more texts: Text-map showing interrelationships between 30 texts

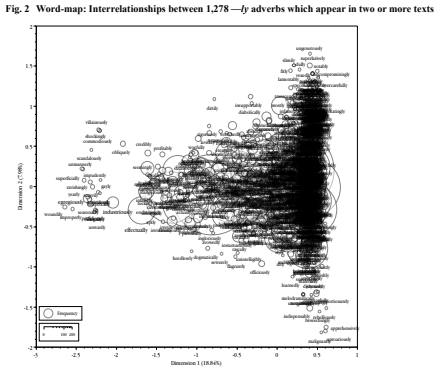

Figure 2: CA: Word-map showing interrelationships between 1,278 types of —ly adverbs

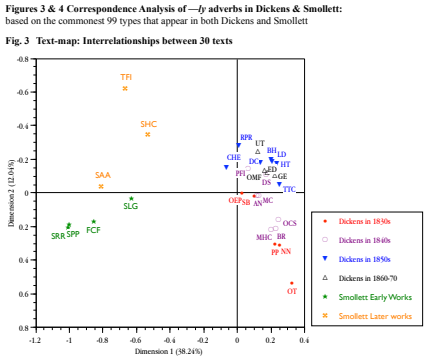

Figure 3: Correspondence Analysis of —ly adverbs in Dickens & Smollett based on the most common 99 types: Text-map showing interrelationships between 30 texts

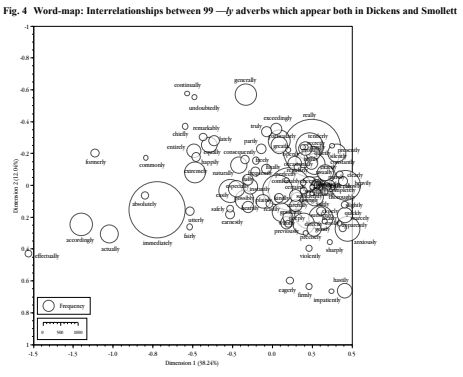

Figure 4: CA: Word-map showing interrelationships between 99 —ly types of adverbs

Figures 1-4 summarise the results of applying a CA model in the frequency analysis of *-ly* adverbs in texts. Figures 1 and 2, based on 1,278 *-ly* adverbs which occur in more than one text, clearly differentiate between the Dickens and Smollett sets. The pattern along the horizontal axis allows quite straightforward interpretation. A more sceptical mind, however, might attribute it to the imbalance in the number of texts between the authorial sets as well as in the number of types of adverbs with the Dickens corpus at 4 times the size of the Smollett corpus. One might be able to respond to such a scepticism with Figures 3 and 4, which are based on the most common 99 *-ly* adverbs used by both Dickens and Smollett. Despite the decrease in the number of variables from 1,278 to 99, the configuration of Figure 3 is remarkably similar to that of Figure 1. Of further interest is that, in each of the two authors' sets, earlier works tend to be found towards the bottom of the chart with later works in the upper half of the diagram. Additionally, in the Dickensian territory of Figures 1 and 3, serial fiction texts occupy the right end while other genres, such as sketches and history, are located slightly towards the left. The series of results seems to illustrate how the authorial difference, text genre, and chronology are reflected in the frequency pattern of *-ly* adverbs in the texts written by Dickens and Smollett. This pilot study might suggest the effectiveness of the stylo-statistical approach based on correspondence analysis of lower frequency words in texts.

Bibliography

Biber, D. *Variation across speech and writing.* Cambridge: Cambridge University Press, 1988.

Biber, D., and E. Finegan. "The Linguistic Evolution of Five Written and Speech-Based English Genres from the 17th to the 20th Centuries." *History of Englishes: New Methods and Interpretation in Historical Linguistics.*. Ed. M. Rissanen. Berlin/New York: Mouton de Gruyter, 1992. 668-704.

Biber, D., S. Johansson, G. Leech, S. Conrad , and E. Finegan. *Longman Grammar of Spoken and Written English.* Harlow: Pearson Education Ltd, 1999.

Brainerd, B. "Pronouns and Genre in Shakespeare's Drama." *Computers and the Humanities* 13.3 (1979): 3-16.

Brainerd, B. "The Chronology of Shakespeare's Plays: A Statistical Study." *Computers and the Humanities* 14 (1980): 221-230.

Burrows, J.F. *Computation into Criticism: A study of Jane Austen's novels and an experiment in method.* Oxford: Clarendon Press, 1987.

Burrows, J.F. "'A Vision' as a revision?" *Eighteenth-Century Studies* 22 (1989): 551-65.

Burrows, J.F. "Computers and the Study of Literature." *Computers and Written Texts.* Ed. C.S. Butler. Oxford: Blackwell, 1992. 167-204.

Burrows, J.F. "Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative." *Research in Humanities Computing 4.* Ed. S. Hockey and N. Ide. Oxford/New York: Oxford University Press, 1996. 1-33.

Cluett, R. "Style, Precept, Personality: A Test Case." *Computers and the Humanities* 5 (1971): 257-274.

Cluett, R. *Prose Style and Critical Reading.* New York: Teachers College Press, 1976.

Conrad, S., and D. Biber, eds. *Variation in English: Multi-Dimensional Studies.* Harlow: Pearson Education Ltd, 2001.

Craig, D. H. "Johnsonian chronology and the styles of A Tale of a Tub." *Re-Presenting Ben Jonson: Text Performance, History.* Ed. M. Butler. London: Macmillan, 1999a. 210-232.

Craig, D.H. "Authorial Attribution and Computational Stylistics: If You Can Tell Authors Apart, Have You Learned Anything About Them?" *Literary and Linguistic Computing* 14 (1999b): 103-113.

Craig, D.H. "Contrast and Change in the Idiolects of Ben Jonson Characters." *Literary and Linguistic Computing* 33.3 (1999c): 221-240.

Linmans, A.J. "Correspondence Analysis of the Synoptic Gospels." *Literary and Linguistic Computing* 13 (1998): 1-13.

Mealand, D.L. "Style, Genre, and Authorship in Acts, the Septuagint, and Hellenistic Historians." *Literary and Linguistic Computing* 14 (1999): 479-505.

Merriam, T. "Heterogeneous Authorship in Early Shakespeare and the Problem of Henry V." *Literary and Linguistic Computing* 13 (1998): 15-28.

Milic, L. T. *A Quantitative Approach to the Style of Jonathan Swift.* The Hague: Mouton, 1967.

Nakamura, J. "Text Typology and Corpus: A Critical Review of Biber's Methodology." *English Corpus Studies* 2 (1995): 75-90.

Nakamura, J. "Statistical Methods and Large Corpora: A New Tool for Describing Text Types." *Text and Technology: In Honour of John Sinclair.* Ed. M. Baker, G. Francis and E. Tognini-Bonelli. Amsterdam: John Benjamins, 1993. 293-312.

Sigley, R. "Text Categories and Where You Can Stick Them: A Crude Formality Index." *International Journal of Corpus Linguistics* 2.2 (1997): 199-237.

Tabata, T. "Investigating Stylistic Variation in Dickens through Correspondence Analysis of Word-Class Distribution." *English Corpus Linguistics in Japan.* Ed. T. Saito et al. Amsterdam: Rodopi, 2002. 165-182.

Takefuta, Y. *コンピューターの見た現代英語：ボキャブラリーの科学 ('The Computer Analysis of the Contemporary English Language: a quantitative study of vocabulary').* Tokyo: Educa, 1981.