poetry on Chadwyk Healy's LION site, so the word can be well represented in the poetic genre. And in comparisons that will be made in the presentation in both vocabulary and themes between Gobineau and Balzac, the interest of both authors in divination will be highlighted. Brunet's database shows seven occurrences of divination/divinations in the Comédie Humaine (CH) and four of divinatoire, a typical total of the two words (11) for the 4,242,038 words in that corpus (CH) and for his time, 1825-1849: 35 occurrences in the 12,352,370 words contained in the ARTFL database. However, these words are only about one-half as common in French literature for the last quarter of the nineteenth century as in the second quarter. There are 18 occurrences in 9,548,198 words for the ARTFL database during the period 1875-1899 when the Nouvelles asiatiques were published.

Besides the few sample words above where the computer-assisted techniques have been applied by a human translator of French literature, the presentation will suggest how such frequency-based comparisons can assist in the translation of thematic groups of words representing the literary authors' symbolic universes. Often such clusters of associated words can be determined from methodically searching the secondary literature: the thematic areas where critics have focused their interest over the centuries. Furthermore, a complementary technique for using advanced search engines on the Internet to aid in solving translation problems will be illustrated or demonstrated.

As long as French and English databases remain available with tools that allow for the appropriate date-stamping, so to speak, of word usage, a methodology can be developed using frequencies and simple statistical tests such as z-scores for comparing the ranking of words across chronological gaps. Such resources offer new and useful tools to the translator both in aiding the development of metatranslations and justifying both them and final translations. Additionally, this process can facilitate greater detail and support for literary criticism that makes use of intertextual and intratextual linguistic materials.

Stylometry, Chronology and the Styles of Henry James

David L. HOOVER

Department of English, New York University

Contemporary stylistic and stylometric studies usually focus on an author with a distinctive style and often characterize that style by comparing the author's texts to those of other authors. When an author's works display diverse styles, however, the style of one text rather than the style of the author becomes the appropriate focus. Because authorship attribution techniques are founded upon the premise that some elements of authorial style are so routinized and habitual as to be outside the author's control, extreme style variation within the works of a single author seems to threaten the validity of the entire enterprise. This apparent contradiction is only apparent, however, for the tasks are quite different. Successful attribution of a diverse group of texts to their authors requires only that each author's texts be more similar to each other than they are to texts by other authors, or, perhaps more accurately, that they be less different from each other than from the other texts. The successful separation of texts or sections of texts with distinctive styles from the rest of the works of an author takes for granted a pool of authorial similarities and isolates whatever differences remain.

Recent work has shown that the same techniques that are able to attribute texts correctly to their authors even when some of the authors' styles are quite diverse do a good job of distinguishing an unusual passage within a novel from the rest of the text (Hoover, 2003). Other quite subtle questions have also been approached using authorship attribution techniques. Nearly 20 years ago, Burrows showed that Jane Austen's characters can be distinguished by the frequencies of very frequent words in their dialogue (1987). More recent studies have used authorship techniques to investigate the sub-genres and varied narrative styles within Joyce's *Ulysses* (McKenna and Antonia, 2001), the styles of Charles Brockden Brown's narrators (Stewart, 2003), a parody of Richardson's *Pamela* (Burrows, 2005), and two translations of a Polish trilogy made a hundred years apart (Rybicki, 2005). Hugh Craig has investigated chronological changes in Ben Jonson's style (1999a, 1999b), and Burrows has discussed chronological changes in the novel genre (1992a).

I am using authorship attribution techniques to study the often-remarked differences between Henry James's early and late styles.[1] I begin by analyzing a corpus of 46 American novels of the late 19th and early 20th century (12 by James and 34 by eight other authors) to determine the extent to which multivariate authorship attribution techniques based on frequent words, such as principal components analysis, cluster analysis, Burrows's Delta, and my own Delta Primes, successfully attribute James's early and late novels to him and distinguish them from novels by eight of his contemporaries.[2] Because all of these techniques are very effective in this task, all are appropriate for further investigation of the variation within James's style, but DeltaLz produces especially accurate results, correctly attributing all 40 novels by members of the primary set in eleven analyses based on the 2000-4000 most frequent words. All of the results also reconfirm recent findings that large numbers of frequent words are more effective than the 50-100 that have been traditionally used, and that the most accurate results (for novel-sized texts) often occur with word lists of more than 1000 words (see Hoover, 2004a, 2004b). The PCA analysis in Fig. 1, based on the 1001-1990 most frequent words, clusters the novels quite well–better, in fact, than analyses that include the 1000 most frequent words.

When cluster analysis, PCA, Delta, and Delta Prime techniques are applied to nineteen novels by Henry James, they show that the early (1971-1881) and late styles (1897-1904) are very distinct indeed, and that an "intermediate" style (1886-1890) can also be distinguished. DeltaLz again produces especially accurate results, correctly identifying all early, intermediate, and late novels in 24 analyses based on the 200-4000 most frequent words. These results paint a remarkable picture of an author whose style was constantly and consistently developing, a picture that is congruent with James's reputation as a meticulous craftsman who self-consciously transformed his style over his long career. A comparison with Charles Dickens and Willa Cather shows that Dickens's early and late novels tend to separate, but do not fall into such neat groups as James's do, and that Cather's novels form consistent groupings that are not chronological. These authors seem not to have experienced the kind of progressive development seen in James. It is dangerous, then, simply to assume chronological development of authorial style.

Finally, these same techniques show that the heavily revised versions of *The American* (1877), *Daisy Miller* (1878), and *The Portrait of a Lady* (1881) that appear in the New York edition of James's novels (1907-09) are consistently and dramatically closer to the style of the later novels. Yet even his notoriously detailed and extensive revisions do not allow PCA to group the revised early novels with the late novels. Instead, the revised versions fall at the border between the early and intermediate novels in PCA graphs (see Fig. 2), and consistently join with their original versions in cluster analyses. The results obtained using Delta show that even the errors make sense. In analyses that are not completely accurate, *The Portrait of a Lady* and *Washington Square,* the latest of the early novels (both 1881) are sometimes identified as intermediate. The other errors involve the identification of *The Spoils of Pointon,* the first of the late novels (1897), as intermediate; no analyses incorrectly identify an early novel as late or a late novel as early. In the analyses that are completely correct for the 19 unambiguously early, intermediate and late novels, the New York edition versions of earlier novels are identified as follows: the revised early novels *The American* and *Daisy Miller* are universally identified as early, the revised intermediate novel *The Reverberator* is universally labeled intermediate, and the revised early *The Portrait of a Lady* is usually labeled intermediate, but sometimes early. This is an intuitively plausible result, with the latest of the early novels pulled far enough toward the late novels to appear intermediate, but, interestingly enough, DeltaLz, which produces much more accurate results overall, labels all of the novels according to their original publication dates in eighteen analyses based on the 200-2800 mfw. In the remaining six analyses, the early *Daisy Miller* is identified as late, and in one analysis (based on the 4000mfw) *The Portrait of a Lady* is identified as intermediate. Further investigation of the implications of these results is ongoing.

Authorship attribution techniques thus confirm the traditional distinction between early and late James, establish the existence of an intermediate style, and lay the groundwork for a fuller analysis of the linguistic and stylistic differences upon which they rest. Based as they are on a very large proportion of the text of the novels (the 4000 most frequent words typically comprise more than 94% of a novel), these results provide a wealth of material for stylistic analysis. The huge numbers of words involved will, however, require new methods of selection, analysis, and presentation if they are not to prove overwhelming and incomprehensible. Meeting these challenges will advance and refine authorship attribution techniques, and, at the same time, further illuminate the linguistic bases of James's style and his process of revision.

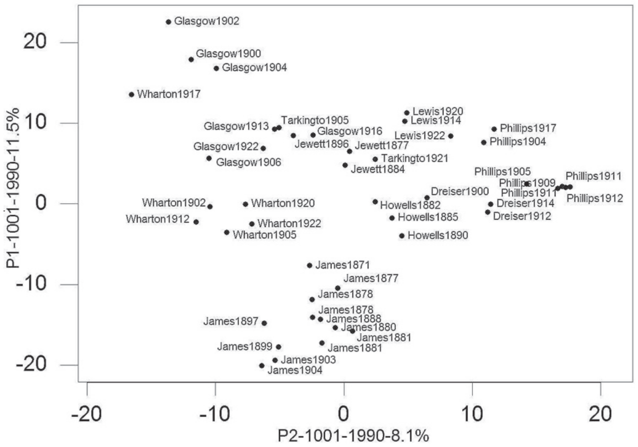

Fig. 1. 46 Novels by Henry James and 8 Other Authors

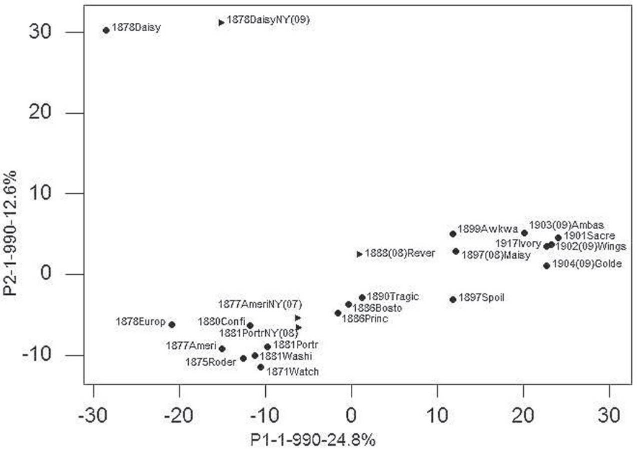

Fig. 2. Twenty-Three Novels by Henry James (revised versions marked with triangles)

References

Burrows, J. F. (1987). *Computation into Criticism. Oxford:* Clarendon Press.

Burrows, J. F. (1992a). "Computers and the study of literature." In Butler, C. S., ed. *Computers and Written Texts*. *Oxford: Blackwell,* 167-204.

Burrows, J. F. (1992b). "Not unless you ask nicely: the interpretative nexus between analysis and information," *LLC 7:* 91-109.

Burrows, J. F. (2002a). "'Delta': a measure of stylistic difference and a guide to likely authorship". LLC 17: 267-287.

Burrows, J. F. (2002b). "The Englishing of Juvenal: computational stylistics and translated texts." Style 36: 677-99.

Burrows, J. F. (2003). "Questions of authorship: attribution and beyond." *CHUM* 37: 5-32.

Burrows, J. F. (2005). "Who wrote Shamela? Verifying the authorship of a parodic text," *LLC* 20: 437-450.

Craig, H. (1999a). "Contrast and change in the idiolects of Ben Jonson characters," *CHUM* 33: 221-240.

Craig, H. (1999b). "Jonsonian chronology and the styles of a Tale of a Tub. In Butler, M. (ed.). *Re-Presenting Ben Jonson: Text, History, Performance*. Macmillan. St. Martin's, Houndmills, England, 210-32.

Hoover, D. L. (2003). "Multivariate analysis and the study of style variation." *LLC* 18: 341-60.

Hoover, D. L. (2004a). "Testing Burrows's Delta." *LLC* 19: 453-475.

Hoover, D. L. (2004b). "Delta prime?" *LLC* 19: 477-495.

McKenna, C. W. F. and A. Antonia. (2001). "The statistical analysis of style: reflections on form, meaning, and ideology in the 'Nausicaa' episode of *Ulysses," LLC* 16: 353-373.

Rybicki, J. (2005). "Burrowing into translation: character idiolects in Henryk Sienkiewicz's trilogy and its two English translations," *LLC* Advance Access published on March 24, 2005. doi:10.1093/llc/fqh051

Stewart, L. (2003). "Charles Brockden Brown: quantitative analysis and literary interpretation," *LLC* 18: 129-138.