Corpus Stylistics, Stylometry, and the Styles of Henry James
Author(s): David L. Hoover
Source: *Style*, Vol. 41, No. 2, Style in Fiction (Summer 2007), pp. 174-203
Published by: Penn State University Press
Stable URL: https://www.jstor.org/stable/10.5325/style.41.2.174

David L. Hoover
New York University

Corpus Stylistics, Stylometry, and the Styles of Henry James

1. Introduction

> They would dine together at the worst, and, with all respect to dear old Waymarsh—if not even, for that matter, to himself—there was little fear that in the sequel they shouldn't see enough of each other.

The beginning of Henry James's *The Ambassadors* is sufficient to explain why "There has never been any doubt that he had a 'style'" (Lodge 189). And the obvious artifice of sentences like the one above, from its famous first paragraph, have helped to establish James as an exceptionally self-conscious stylist. Furthermore, although change must be expected in the style of any author with a career spanning forty years, the stylistic differences between James's novels of the 1870's and those written after 1900 have long been considered extreme. Shortly after James's death, Carl Van Doren asserted that the style of early novels like *The American*, *The Europeans*, *Daisy Miller*, *Washington Square*, or *The Portrait of a Lady* is quite different from the "obscure" late style that some readers dislike (18). Traditionally, the distinctiveness of James's late style has been attributed primarily to his sometimes convoluted and self-interruptive syntax. R. W. Short notes that James's syntactic "distortions" often "obliterate the normal elements of connection and cohesion. When he has undone the usual ties, his meanings float untethered, grammatically speaking, like particles in colloidal suspension" (73–4). Ian Watt's well-known explication of the first paragraph of *The Ambassadors* also memorably discusses the syntax (442-55), as do Richard Ohmann (274–5), and Leech and Short (100–1) (see also my "Altered Texts" 110–13).

Other critics have pointed out some non-syntactic alterations James made in his revisions of his early novels for New York edition; for example, more explicit and precise lexis, more figures of speech, more varied and elaborate speech markers, more contractions and colloquialisms, and more adverbial modifiers. These alterations, as well as syntactic changes, tend to make the revised earlier novels more like the later ones (on non-syntactic differences between early and late James, see

Lee; Watt; Lodge; Gettmann; Krause; and especially Chatman). Here, however, I want to investigate the distinctiveness of James's style and the traditional division of his novels into early and late styles by applying methods of authorship attribution and stylometry, methods based not upon syntax, but simply upon the frequencies of words of all kinds.

A corpus approach that takes into account most of the words of James's novels seems especially appropriate for examining the lexical aspects of his style, and it at least partially addresses a problem perceptively discussed by Leech and Short more than twenty-five years ago: "While a condensed poetic metaphor, or a metrical pattern will jump to the attention as something which distinguishes the language of poetry from everyday language, the distinguishing features of a prose style tend to become detectable over longer stretches of text, and to be demonstrable ultimately only in quantitative terms" (2–3). Recent work has shown that stylometric techniques can successfully identify unusual sub-styles and multiple narrators within a novel, can distinguish parodies from originals, and can illuminate the styles of multiple translations of a novel (Stewart; Hoover, "Multivariate"; Hoover, et al.; Burrows, "Englishing"; Burrows, "Who Wrote Shamela"; Rybicki). Here both well-established and emerging stylometric techniques will be used to study the differences between Henry James's early and late styles. These techniques easily distinguish James from his contemporaries and also identify distinct sub-styles within the James's nineteen major novels. They confirm the traditional distinction between early and late James, identify an intermediate style and lay the groundwork for a fuller analysis of the linguistic and stylistic differences that define James's styles. They show that James's style can be demonstrated quantitatively.

One of the most promising characteristics of the new techniques is that they can identify words that display extremely variable frequencies across the entire James corpus. Such words clearly differentiate James's sub-styles, and the fact that they disproportionately increase or decrease steadily through time also marks James's career as remarkably unidirectional. The new techniques also expand the traditional stylometric focus on the most frequent words of the novels to include rare and moderately frequent words that characterize the three periods. The rare words, especially the *-ly* adverbs, are often quite novel and, in spite of their rarity, strike a vivid Jamesian note. The moderately frequent words are even more significant: they are frequent enough to be consciously noticed in early novels, but are rare or nonexistent in late novels, or vice versa. Stylometric techniques constitute a promising new avenue of research that exploits the power of corpus analysis and focuses attention on a manageable subset of the author's vocabulary.

2. Preliminary Tests: Distinguishing James from other Authors

Stylometric techniques assume that word frequencies are largely outside the author's conscious control because they result from habits that are stable enough to create a verbal fingerprint. Style variation within an author's works seems to threaten their validity. Fortunately, the threat is more apparent than real: an author's various styles can be similar enough to appear alike when compared with the styles of other authors and yet different enough to be distinguished from each other. A stylometric technique cannot reasonably be used to distinguish sub-styles in James unless it can distinguish his novels from those of his contemporaries, but preliminary tests on a specially-created corpus of seventy-one novels by James and his contemporaries shows that two popular stylometric techniques are very successful in doing so. (For more details on this corpus, see the Appendix.)

The first of these techniques is based on **Delta,** a measure of textual difference, recently developed by John F. Burrows, that is designed to identify which of a set of given authors is most likely to have written a text of uncertain authorship ("Delta"; "Englishing"; "Questions"; Hoover, "Testing"). Twenty-five analyses based on the 100-4000 most frequent words of the seventy-one novels show that Delta is very effective, attributing thirty-eight of forty novels to their correct authors in analyses based on the 2200-4000 most frequent words (see Appendix 1 for an explanation of Delta). **Delta-Lz**, a modification of Delta based only on words with widely divergent frequencies, correctly attributes thirty-nine of the novels over a wider range of analyses (for more details on Delta-Lz, see Appendix 2; see also my "Delta Prime?"; "Word Frequency"). The effectiveness of Delta and Delta-Lz in attributing all of James's novels to him, regardless of their diverse styles, confirms that they are appropriate tools for investigating his style.

The second technique is **Cluster Analysis**, which, as its name suggests, groups novels together based on the similarities and differences of the frequencies of the most frequent words (see Appendix 3 for more details). Cluster Analysis is also very effective, correctly attributing thirty-seven of the forty novels, as Figure 1 shows. Note how James's novels form one of two main clusters, reflecting the distinctiveness of his style. Cluster Analysis and Delta, though very different techniques, tend to fail for the same authors, and this suggests that both are capturing real similarities and differences. Furthermore, the novels of most of the authors show a strong tendency to appear in order of publication (the author-title for all authors with multiple texts ends in a two-digit year of publication). Although minor perturbations appear in the order of James's novels (especially for *Daisy Miller*,

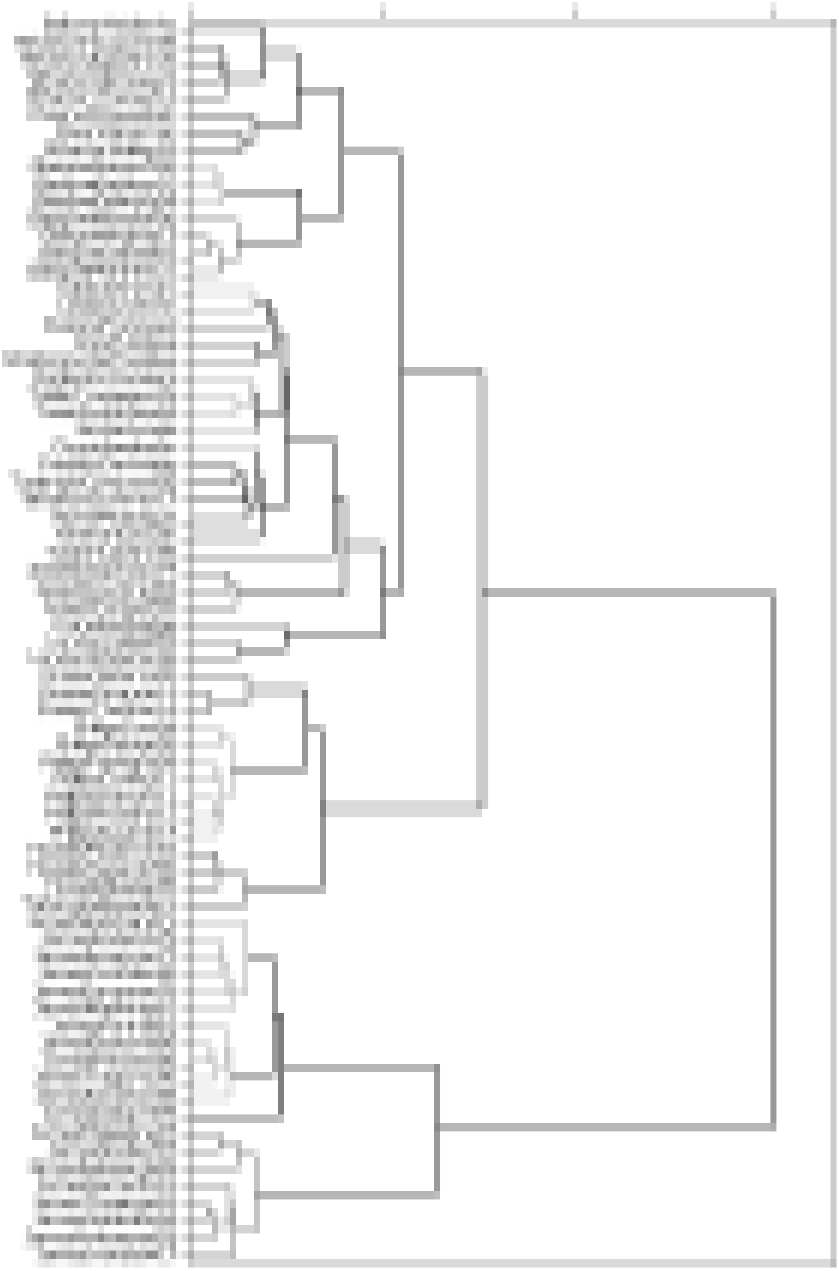

Fig. 1. Cluster Analysis: 71 American Novels, 800 Most Frequent Words

which is likely to be anomalous because it is so short), the strong tendency for his novels to form early and late groups suggests that Cluster Analysis should prove very useful in investigating chronological changes in James's style. When the word list is increased to the 4000 most frequent words, Cluster Analysis, like Delta-Lz, correctly clusters thirty-nine of the forty novels.

3 Chronological Stylometry: Documenting the Development of James's Style

We can now investigate the development of James's style by analyzing twenty of his novels, including all his major novels except three that are not available as electronic texts: *The Other House* (1896), *The Outcry* (1911), and *The Sense of the Past* (1917; an unfinished posthumous novel):

> *Watch and Ward*, 1871; *Roderick Hudson,* 1875; *The American,* 1877; *The American,* 1907 [1877]; *The Europeans,* 1878; *Daisy Miller,* 1878; *Daisy Miller,* 1909 [1878]; *Confidence,* 1880; *Washington Square,* 1881; *The Portrait of a Lady,* 1881; *The Portrait of a Lady,* 1908 [1881]; *The Bostonians,* 1886; *The Princess Casamassima,* 1886; *The Reverberator,* 1908 [1888]; *The Tragic Muse,* 1890; *The Spoils of Poynton,* 1897; *What Maisie Knew,* 1908 [1897]; *The Awkward Age,* 1899; *The Sacred Fount,* 1901; *The Wings of the Dove,* 1909 [1902]; *The Ambassadors,* 1909 [1903]; *The Golden Bowl,* 1909 [1904]; *The Ivory Tower,* 1917 [1]

I have already noted Van Doren's distinction between the early novels and the "obscure" later ones, and Joseph Warren Beach accepts this distinction while also suggesting that the novels of the 1890's can be considered transitional (x). Richard Poirier includes the novels through *The Portrait of a Lady* in his study of the early novels, Dorothea Krook-Gilead identifies *The Awkward Age* as the first late novel (135), and John F. Burrows suggests a similar division (*Computation* 159). With some variations, then, a consensus exists about which novels are early and late, and relatively large gaps in original publication dates between *The Portrait of a Lady* (1881) and *The Bostonians* (1886), and between *The Tragic Muse* (1890) and *The Spoils of Poynton* (1897) divide the novels reasonably, if artificially, into three groups that agree broadly with that consensus.

Delta tests treat early, intermediate, and late James as three different authors. *Roderick Hudson, The Bostonians,* and *The Ambassadors* have been chosen as representatives of the early, intermediate, and late styles on the basis of their publication dates and similar sizes, and the twenty remaining novels comprise the test texts. Delta categorizes the novels by period quite effectively, often giving completely correct results, and Delta-Lz is even more accurate, with completely correct results in most analyses. The following details come from analyses based on a word list created from twenty of the twenty-three novels, omitting only the New York edition versions of the three novels for which I have two versions. (Also omitting the brief *Daisy Miller* and the New York edition of *The Reverberator* produces even more accurate results, but the additional errors reported below are instructive.)

Delta analyses based on a word list that includes personal pronouns, omits words that do not occur in any of the primary novels, and omits any word for which a single text supplies ninety-nine per cent of its occurrences (to remove proper names and other words frequent in only one novel) produce five completely correct results, based on the 200-600 most frequent words. The New York editions of the early novels *Daisy Miller* and *The American* are identified as early and the intermediate novel *The Reverberator* is identified as intermediate in all of these analyses. In four of them, however, the New York edition of *The Portrait of a Lady* (1881) is identified as intermediate. The original version of this final novel of the early period is also identified as intermediate in some analyses that are not completely correct. In such analyses, *Washington Square* (1881) and *Confidence* (1880) are also occasionally identified as intermediate, and *The Spoils of Poynton* (1897), from the beginning of the late period, is once identified as intermediate. It seems encouraging that incorrect attributions identify these latest early and earliest late novels as intermediate, for James's style is more naturally conceived of as a continuum than as three separate periods. Whatever other differences characterize the early and late James, these broad and widely accepted critical categories are compellingly correlated with the frequencies of the most frequent words.

Delta-Lz analyses using the same word list produce almost universally correct results. The only errors identify the intermediate novels *The Princess Casamassima* and *The Tragic Muse* as early. The New York editions of *The American, The Portrait of a Lady,* and *The Reverberator* are assigned to their original periods of publication except in one analysis in which *The Portrait of a Lady* is identified as intermediate. The New York edition of *Daisy Miller* is identified as early in all analyses except those based on the 3000-4000 most frequent words, where it is identified as late, suggesting that James's revisions have their greatest effect on less frequent words. Note that Delta produces its best results based on the 200-600 most frequent words and Delta-Lz produces completely correct results using still larger numbers of words. The distinctions among the three styles are quite stable even when words with a frequency rank of about 4000 are included, words that together account for more than ninety per cent of all the words in the novels. These distinctions thus rest on a very firm foundation.

Principal Components Analysis using the same word list also gives excellent results, and the analysis shown in Figure 2 reveals a remarkable trajectory in the development of James's style. (The nearer the novels are to each other in the graph, the more similar they are stylistically; see Appendix 4 for a brief explanation.) The three periods are clearly visible in the graph, and the novels tend to appear in publication order throughout, though this analysis suggests that the intermediate and early novels are similar and that the three earliest of the late novels are somewhat different from the later ones. The New York edition versions of the early novels move toward the later novels without leaving the early group, showing that James's extensive revisions make them more like the later novels without masking their dates of composition. These results confirm Burrows's work on the *The American,* in which he briefly (but nicely) discusses the variability of James's style ("Not Unless" 98). (The peculiar position of *Daisy Miller* in the graph is again presumably a result of its small size.)

Finally, Cluster Analysis based on the same word list produces striking results in which the novels fall in almost perfect chronological order (see Figure 3). This pattern is quite stable for the 400-995 most frequent words, confirming that James's style is characterized by an extraordinary unidirectional development over time. Cluster Analysis identifies five possible sub-styles, and, in contrast to Principal Components Analysis, suggests that the intermediate novels are more similar to the earliest of the late novels than to the early novels.

4. The Chronological Characterization of James's Vocabulary

Documenting the chronological development of James's style in this way characterizes it in broad terms, but the large word lists are difficult to discuss with any specificity. Mapping the development of James's style has pushed us farther from the individual words, farther from the texts. One way to limit the sheer volume of information and focus attention on words that best characterize James's style is to concentrate on words with very different frequencies in the early, intermediate, and late novels, using the **Distinctiveness Ratio** (see the Appendix 5 and Hoover, et al., chapter four, for details). This simple but effective measure of variability, introduced by Ellegård, is defined as the rate of occurrence of a word in one text divided by its rate of occurrence in another (Kenny 69–70). Though it is obviously designed to compare just two texts, a modified version that divides the maximum by the minimum and then by the median frequency identifies words with the widest range in frequency across the three periods of James's style.

I will concentrate here on two other methods of selecting characteristic words for discussion, but the 100 words with the largest Distinctiveness Ratios among those that increase or decrease steadily from the early to the late styles deserve comment. Even this severely limited list of words displays several important pat-

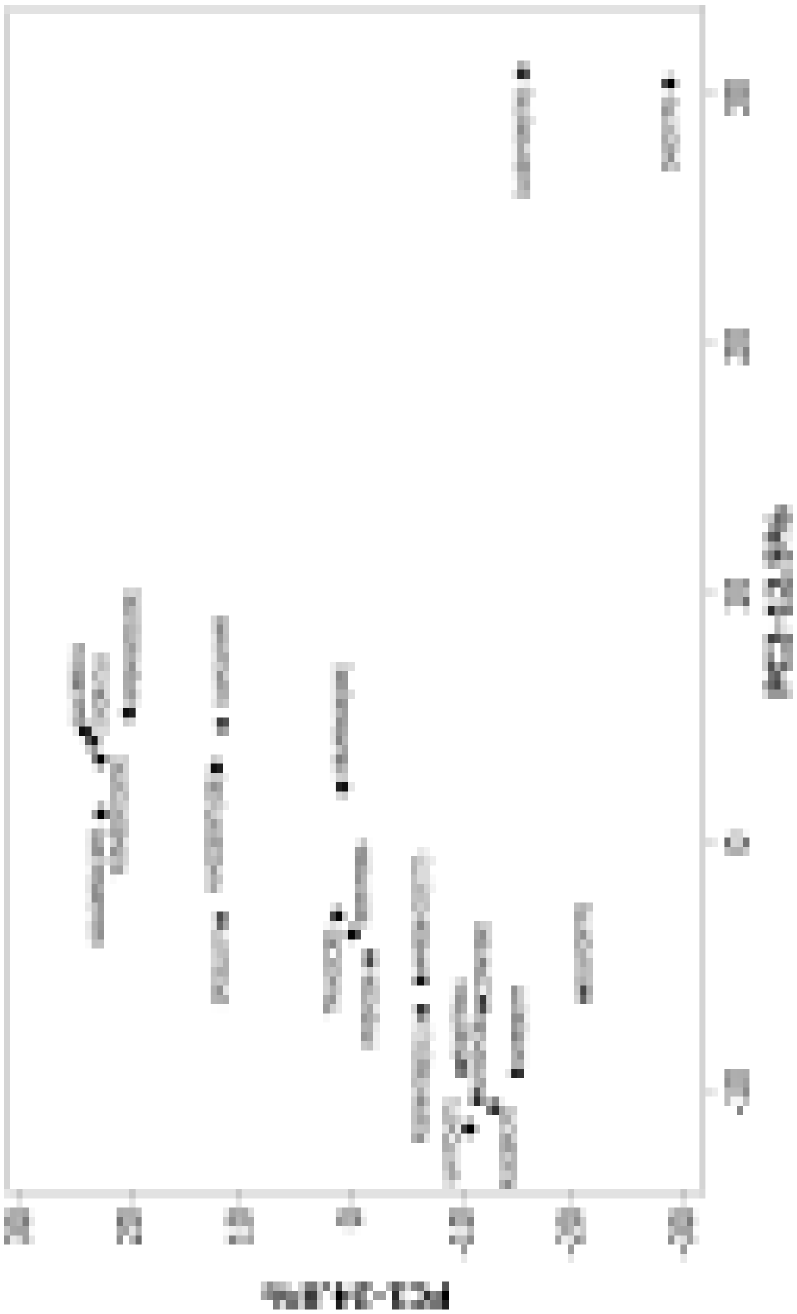

Fig. 2. Principal Components Analysis: 23 James Novels, 995 Most Frequent Words

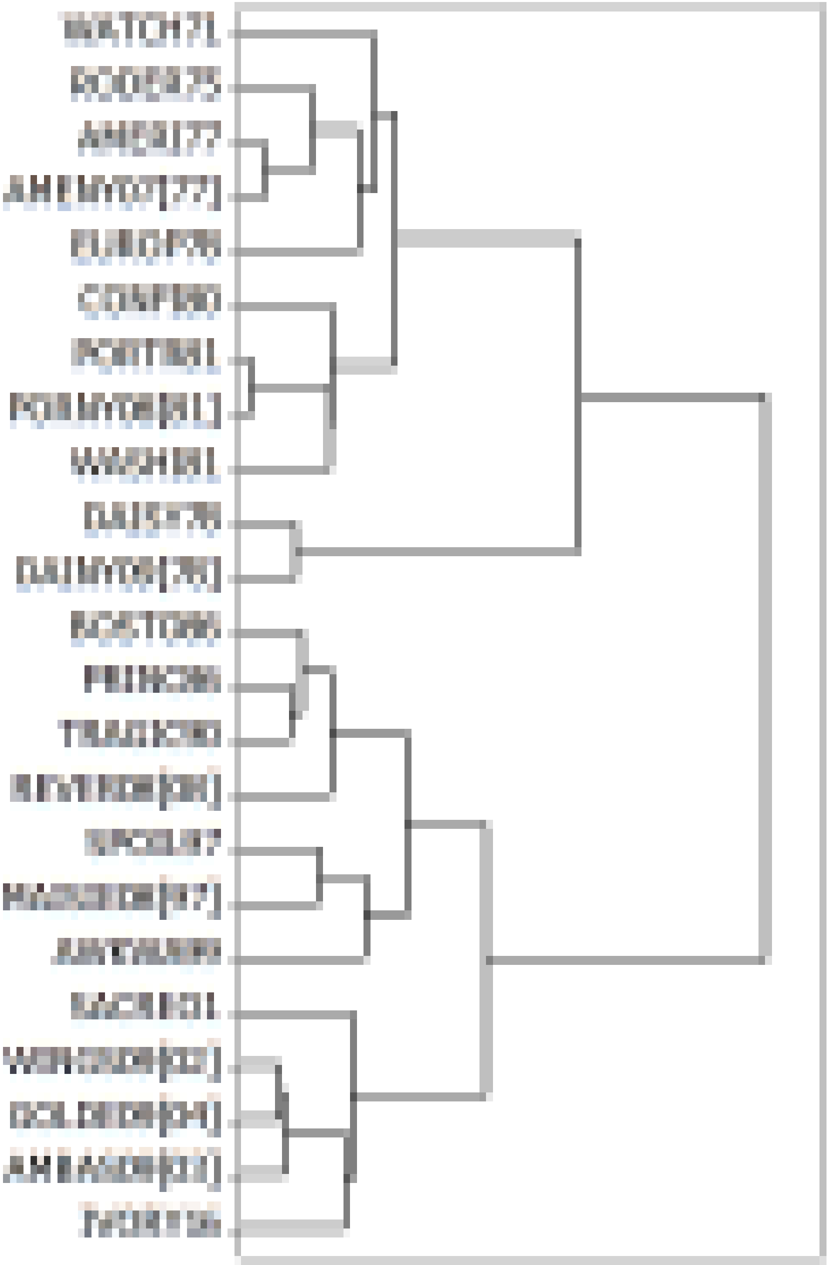

Fig. 3. Cluster Analysis: 23 James Novels, 995 Most Frequent Words

terns. There are six *-ly* adverbs in the increasing list, but only two in the decreasing list, confirming the often-remarked fondness of the late James for *-ly* adverbs. Adjectives are slightly more frequent among the decreasing than the increasing words, and four increasing adjectives are comparatives and superlatives, adjective forms that reverse the general trend of fewer adjectives in the late novels. Of the remaining thirty-eight increasing words, twenty are abstract nouns and only six are concrete nouns, while seven of the remaining twenty-four decreasing words are abstract nouns and eight are concrete. Thus the trend toward more abstract diction in the later James is visible even within this narrow focus.

Another significant feature of these 100 words is the presence of families of words that show the same pattern: *frowned/frowning* decrease in frequency, and *gasp/gasped, precaution/precautions, sharper/sharpest*, and *clearer/clearest/clearness* increase. Including about 120 more of the increasing or decreasing words with the highest Distinctiveness Ratios adds *frown* to the *frown-* family, adds *cleared* and *clearly* to the *clear-* family, and uncovers additional families of both kinds:

> Decreasing: *coquetry/coquette, displease/displeasure, offend/offended, shrewd/shrewdness*
>
> Increasing: *disconcerted/disconcerting, oddest/oddly, passage/passages, precaution/precautions, wail/wailed, wince/winced*

Examining all words that occur at least twice in the twenty novels reveals that sixty-five of the 100 words above form families. Sixteen of these consistently increase or decrease, and most of the remaining families are nearly consistent (typically only rare words fail to follow the pattern). Just fourteen of the sixty-five deviate significantly from the increasing or decreasing pattern within the 100 words above. Further research will be needed to determine whether the adoption and abandonment of word families is a result of James's carefully crafted style or is true of most or many other authors.

Zeta and **Iota**, two recently introduced stylometric measures that focus attention on words that are characteristic of an author, seem appropriate for studying the development of James's style (Burrows, "All the Way Through"). Zeta identifies characteristic words of moderate frequency and Iota identifies characteristic words of low frequency (for further explanation of these measures and the word lists below, see Appendix 6). Zeta, especially, should be effective in narrowing our focus and returning us to the text, for the words it identifies are frequent enough to be noticed easily by a reader. Zeta analyses that treat early and late James as two authors identify words that are characteristic of one of the two but very rare or non-existent in the other, so characteristic, in fact, that the following lists of 122 early and 137 late words can each alone distinguish early and late novels (word families in bold):

122 characteristically early words:

affectation, afterward, angels, angrily, arm-chair, aroused, ascended, asleep, audacious, beheld, benignant, bouquet, bravado, buying, caressingly, Catholic, chiefly, chronic, churches, classic, complex, confectioner's, constancy, convent, conversing, ***coquetry, coquette****, dated, deportment, detest, diminutive, dresses, dwelling, easel, eastern, enchanting, eyed, farther, footstep, foreground,* ***frown, frowning****, gayety, glances, graciously, grievous, grudge, habitual, handkerchief, headache, heartless, heavens, hence, humbly, imperious, Jove, joyous,* ***keenly, keenness****, ladies', lain, longed, matrimony, melting, misfortunes, mock, moonlight, motioned,* ***mountain, mountains****, narrative, ornamented, outright, paces, penniless, peremptory, petition, picturesqueness, pious, portico, prayers, pregnant, prosperous, quoted, repent, resume, riddle, roads, roses, rugged, salt, sarcasm, scold, seventeen,* ***shrewd, shrewdness****, shrinking, sketching, skill, slippers, smelling, sober, squarely, starlight, sternly, stooped, sturdy, suitor, sunset, tresses, unkind, unmarried, upwards, valley, vexed, virtues, waist, wanderings, wedding, weeping, whispered, witty*

137 characteristically late words:

acquisition, afar, alien, ambiguity, amenity, anticipated, anybody, arrest, aspects, assault, ***assert, asserted****, attestation, austerity, bears, blandness, blest, blinked, brim, brisk, cage, causing, certitude, chap, chuck, circled, clutch, competence, concurred, consequent, consistently, contacts, controlled, corrective, debated, designed, dig, diplomacy, discernible, disconcerting, discouragement, discrimination, dodge, doom, drops, enclosed, entertainer, everyone, everything's, expert, extravagantly, facial, flag, flare, flaunt, flooded, foretaste, four-wheeler, frock, funny, fuss,* ***gasp, gasped****, happening, helpful, hitch, impunity,* ***impute, imputed****, incontestably, ironic, ivory, lighting, likewise, linked, loathes, logically, lounge, loveliest, madness, magnificently, markedly, mayn't, momentarily, Mummy, nobody's, nurses, person's, pertinence,* ***pleasantly, pleasantness****, porter, precipitation, previously, prodigy, proportionately, reckoned, recognitions, recovery, reported, rueful, scarce, seal,* ***sharper, sharpest****, shift, shriek, signed, smoking-room,* ***somebody, somebody's****, spaces, spreading, squared, straightway, strayed, stress, sufficiency, surge, surrounding, terrific, thereby, token, tucked, twenty-five, uncanny, unconsciousness, unnaturally, untouched, unutterable, upshot, wail, wan, warranted, weigh, whisked, wince*

None of the words in either list is very frequent. All occur at least six times in James's novels, but only two occur more than 100 times–the late words *scarce* (232) and *funny* (102), and only a handful occur more than fifty times. These lists are too large and diverse for easy characterization, but they differ in many ways: for example, in the proportions of various parts of speech. (All figures are approximations; many word forms belong to more than one part of speech and only an exhaustive examination of all their occurrences would allow precise counts.) The early list has many more adjectives than the late one (25% versus 16%). Not many of those in either list describe physical characteristics or qualities, and even in the few that might qualify the physical seems attenuated: *asleep, diminutive, Eastern, joyous, rugged,* and *sturdy* (early); *brisk, facial, rueful, wan*, and *loveliest* (late). If the early list has more adjectives, the late list has many more verbs (30% versus 18%). In both lists, many verbs relate to perception, language, and mental and emotional processes, and very few to physical actions or movement. Discovering why only the late list contains third person singular present tense verbs (*bears, loathes*) would require further research.

Although there are roughly equal percentages of singular nouns in the two lists, plural nouns are about three times as frequent among the early words as among the late words (13% versus 4.4%), and these plural nouns point to a further difference. As in the list of 100 words above, concrete nouns are much more prevalent among the early than the late words, but this is overwhelmingly true for the plural nouns. In the early list, *churches, angels, dresses, glances, heavens, ladies, mountains, paces, prayers, roads, roses, slippers, tresses,* and *wanderings*, might be broadly construed as concrete, while only *virtues* and *misfortunes* are clearly abstract. In the late list, *drops, nurses*, and *spaces* might be seen as concrete, while *aspects, contacts*, and *recognitions* are abstract. The presence of six pronouns in the late list and none in the early list also seems significant, though the pronouns would have to be examined in context to determine what the difference means. The adjectives, verbs, and nouns among these 259 words are thus consistent with James's focus on social and psychological interaction rather than the physical world or plot.

Besides these variations in the frequencies of parts of speech, some other differences seem significant. Consider, for example, the characteristically early words loosely related to courtship and marriage: *bouquet, coquetry, coquette, matrimony, suitor, unmarried, wedding, deportment, enchanting, heartless, tresses.* Among these, *coquetry, coquette, suitor, deportment,* and *tresses* seem old fashioned and almost precious in tone. In the late list, it is difficult to suggest any words that more than remotely to fit into this category. Though marriage and courtship are certainly important in late novels, the later James seems to have found less overt ways of writing about them. Finally, words that can be loosely categorized as slangy or informal are much more frequent in the late list (*blinked, brisk, chap, chuck, dodge, funny, fuss, hitch, shift, shriek, squared, tucked, upshot, wail, whisked, wince*) than in the early list (*grudge, mock, outright, riddle, witty*).

If these lists are too large for thorough analysis, they are too small to characterize the development of James's style adequately. An expanded Zeta analysis identifies 446 characteristic early and 673 characteristic late words that can each easily distinguish the early and late novels, and Iota analysis identifies 4952 characteristic early and 4696 characteristic late words that are equally accurate (see Appendix 6 for details). Because there are so many words to be considered in these tests, I have distinctively marked each one in a typical early novel and a typical late novel, *Roderick Hudson* and *The Ambassadors*, so that they can be examined easily in context. The early Iota words do not occur in the later novels and vice versa, so that there are about 5400 possible words to mark in each novel: the early and late Zeta words in both texts, the early Iota words in *Roderick Hudson* and the late Iota words in *The Ambassadors*.

Paging through the marked novels is instructive. Burrows suggests that the failures he observes in some Zeta analyses may be an artifact of the middle of the word frequency spectrum where "the demands of subject and occasion might be expected to prevail over the effects of authorial habit" ("All the Way Through" 17). Yet Iota words, especially the rarest ones, sometimes also relate closely to the subject matter, as in Passage A from near the beginning of *Roderick Hudson* (1875) (**early Zeta words**; ***early Iota words***):

> A. He had sprung from a rigid Puritan stock, and had been brought up to think much more intently of the duties of this life than of its privileges and pleasures. His progenitors had submitted in the matter of **dogmatic theology** to the relaxing influences of recent years; but if Rowland's youthful consciousness was not chilled by the menace of long punishment for brief ***transgression***, he had at least been made to feel that there ran through all things a strain of right and of wrong, as different, after all, in their **complexions**, as the texture, to the spiritual sense, of Sundays and ***week-days***. His father was a ***chip*** of the primal Puritan block, a man with an icy smile and a stony frown. He had always bestowed on his son, on principle, more ***frowns*** than smiles, and if the **lad** had not been turned to stone himself, it was because nature had blessed him, inwardly, with a well of vivifying waters. Mrs. Mallet had been a Miss Rowland, the daughter of a retired ***sea-captain***, once famous on the ships that sailed from ***Salem*** and ***Newburyport***. He had brought to port many a ***cargo*** which crowned the edifice of fortunes already almost colossal, but he had also done a little sagacious ***trading*** on his own account, and he was able to retire, prematurely for so ***sea-worthy*** a ***maritime*** organism, upon a **pension** of his own providing. He was to be seen for a year on the ***Salem*** wharves, smoking the best tobacco and **eying** the seaward horizon with an inveteracy which superficial minds interpreted as a sign of repentance.

Iota words like *sea-captain, Salem, Newburyport, cargo, trading, sea-worthy,* and *maritime* obviously particularize Mallet's background. Although the same might be said of the Zeta words *dogmatic* and *theology*, the other Zeta words are not clearly linked to content. Passage A is relatively dense in early Zeta and Iota words, but the frequencies of such words in this novel suggest that about one Zeta word and three or four Iota words should appear in every passage this size. This helps to explain how such relatively infrequent words contribute to the recognizable style of a novel.

The marked words are not the only ones with a characteristic stylistic valence, however. For example, of the 162 different words in Passage A, sixteen appear only in early novels, including *Rowland, Rowland's*, and *Mallet*, and *cargo,* which only appear in *Roderick Hudson,* and *transgression, trading, sea-worthy, sea-captain, Newburyport,* and *chip,* which appear only once in the twenty novels. These six words provide a useful reminder that more such words exist than might be expected. In the twenty James novels, for example, more than 11,000 of the 32,000 different words appear only once, yet they constitute almost half a percent of the 2.4 million words in the corpus. One or more such words can be expected in any passage the size of Passage A; thus six is a high density of rare words, but not outrageously so. Besides the sixteen words that appear only in the early novels, another thirty-eight decrease steadily from the early to intermediate to late novels, and thirty-four more are more frequent in early than late novels. Together these account for more than half of the different words and almost half of the total words, so that James's lexical style is quite pervasive.

Consider next Passage B, also from *Roderick Hudson* (**early Zeta words**; ***early Iota words***):

> B. Some of Mrs. Light's courtesies were very low, for she had the happiness of receiving a number of the social potentates of the Roman world. She was rosy with triumph, to say nothing of a less **metaphysical** cause, and was evidently vastly contented with herself, with her company, and with the general promise of destiny. Her daughter was less overtly jubilant, and distributed her greetings with impartial ***frigidity***. She had never been so beautiful. Dressed simply in ***vaporous*** white, relieved with half a dozen white roses, the perfection of her features and of her person and the mysterious depth of her expression seemed to glow with the white light of a splendid pearl. She recognized no one individually, and made her courtesy slowly, gravely, with her eyes on the ground. Rowland fancied that, as he stood before her, her **obeisance** was slightly exaggerated, as with an intention of irony; but he smiled philosophically to himself, and reflected, as he passed into the room, that, if she disliked him, he had nothing to reproach himself with. He walked about, had a few words with Miss Blanchard, who, with a ***fillet*** of ***cameos*** in her hair, was leaning on the arm of Mr. Leavenworth, and at last came upon the Cavaliere Giacosa, modestly stationed in a corner. The little gentleman's ***coat-lappet*** was decorated with an enormous **bouquet** and his neck encased in a voluminous white handkerchief of the fashion of thirty years ago. His arms were folded, and he was **surveying** the scene with contracted **eyelids**, through which you saw the glitter of his intensely dark, ***vivacious*** pupil.

As in Passage A, the highlighted words here tell only part of the story. Of the 160 different words, fourteen appear only in early novels, including six proper names *(Rowland, Cavaliere, Blanchard, Leavenworth, Light's, Giacosa)* and four other words *(fillet, vivacious, cameos, coat-lappet)*, that appear only in *Roderick Hudson,* the last two of them only once in the twenty novels. Another forty-four of the 160 decrease steadily from the early to the late novels and fifty-two more are more frequent in early than late novels. In all, more than half the words of Passage B, representing more than two-thirds of the different words it contains are more frequent in early than late novels. So far as I am aware, no one has studied how such a current of characteristic vocabulary is processed by a reader, but it may help to account for the intuition of a sensitive reader that this is an early novel.

Consider now Passage C, from *The Ambassadors* (**late Zeta words**; ***late Iota words***):

> C. It was on this pleasant basis of costly disorder, consequently, that they eventually seated themselves, on either side of a small table, at a window adjusted to the busy quay and the shining ***barge-burdened*** Seine; where, for an hour, in the matter of letting himself go, of **diving** deep, Strether was to feel he had touched bottom. He was to feel many things on this occasion, and one of the first of them was that he had travelled far since that evening in London, before the theatre, when his dinner with Maria Gostrey, between the ***pink-shaded*** candles, had struck him as requiring so many explanations. He had at that time gathered them in, the explanations — he had **stored** them up; but it was at present as if he had either **soared** above or sunk below them — he couldn't tell which; he could somehow think of none that didn't seem to leave the appearance of collapse and cynicism easier for him than lucidity. How could he wish it to be lucid for others, for any one, that he, for the hour, saw reasons enough in the mere way the bright clean ordered ***water-side*** life came in at the open window? — the mere way Madame de Vionnet, opposite him over their intensely white ***table-linen***, their **omelette** ***aux tomates***, their bottle of straw-coloured ***Chablis***, thanked him for everything almost with the smile of a child, while her grey eyes moved in and out of their talk, back to the quarter of the warm spring air, in which early summer had already begun to throb, and then back again to his face and their human questions.

The self-interrupting style of the first sentence, so characteristic of the later James, might alone allow a reader to identify this as a late novel, but the frequent words tell a powerful confirmatory story. Of the 164 different words, thirteen appear only in late novels, including three proper names, *Strether, Gostrey, Vionnet*, and six other words, *Chablis, barge-burdened, pink-shaded, table-linen, tomates, water-side,* that appear only in *The Ambassadors,* the last five only once in the twenty novels. Fifty-five of the 164 steadily increase from the early to the late novels, making this pattern even more prevalent than the decreasing pattern in Passage A. Another forty-seven are more frequent in early than in late novels, so that more than two-thirds of the words of Passage C, representing more than two-thirds of the different words it contains, are more frequent in late than early novels.

One final expansion of the lists of characteristically early and late words shows how pervasive the changes in lexis are in James's style. Ignoring the intermediate novels, I created a list of all words that appear at least twice in either the early or the late novels. Although the kinds of words that appear only once in the early novels or once in the late novels seem stylistically different, I wanted to concentrate on words that could potentially increase or decrease in frequency. I also removed frequent proper nouns and other obvious content words, but not variant spellings for this final demonstration (see Appendix 5 for some comments on "content words"). The lists are so large that removing variant spellings is a tedious and error-prone process, and a British spelling may represent a word adopted by James in his late period that is either rare or non-existent in the early novels rather than a variant spelling of a word also used earlier.

After calculating the average frequencies of the words in the seven early and eight late novels, I calculated the Distinctiveness Ratio between the two and sorted the word list into two groups: words more frequent in the early novels and words more frequent in the late novels. Sorting each of these groups on the Distinctiveness Ratio and collecting all the words in each group that are at least three times as frequent in the early as in the late novels or vice versa identifies about 4400 characteristically early words and about 3900 characteristically late words. Although there is some overlap with Zeta words, these words tend to be much more frequent: the most frequent Zeta word occurs only about forty times in the twenty novels, while at least fifty of the words in each of these lists occurs more than 200 times in the twenty novels. Including words that are at least twice as frequent in one style as the other identifies about 1200 additional characteristically early and 1100 characteristically late words. Marking these four groups of words with distinctive typefaces paints a dramatic picture of the lexical differences between the early and late novels, a picture easily visible on nearly every page. First, consider Passages D and E, from *The Europeans* (1878) and *The Golden Bowl* (1909 [1904]), which are particularly rich in early and late words (the marking is reversed in the two passages to make the early and late words equally salient visually):

> D. And after a while they went out. The air had grown warm as well as **brilliant**; the **sunshine** had **dried** the **pavements**. They **walked** about the **streets** at **hazard**, **looking** at the people and the houses, the ***shops*** and the **vehicles**, the **blazing blue sky** and the **muddy** crossings, the **hurrying men** and the **slow-strolling maidens**, the fresh red **bricks** and the bright green trees, the ***extraordinary*** mixture of smartness and **shabbiness**. From one hour to another the day had grown **vernal**; ***even*** in the bustling **streets** there was an **odor** of earth and **blossom**. Felix was immensely **entertained**. He had called it a **comical country**, and he went about **laughing** at everything he saw. You would have **said** that American **civilization** expressed itself to his sense in a **tissue** of **capital** jokes. The jokes were certainly **excellent**, and the **young** man's **merriment** was **joyous** and genial. He possessed what **is** called the **pictorial** sense; and this first glimpse of **democratic manners** stirred the same sort of attention that he would have given to the movements of a **lively young** person with a bright **complexion**.

> Such attention would have been **demonstrative** and **complimentary**; and in the present ***case*** Felix might have passed for an undispirited **young** exile revisiting the **haunts** of his **childhood**. He kept **looking** at the **violent blue** of the **sky**, at the **scintillating** air, at the scattered and ***multiplied*** patches of **color**. (**early 3X late**, **early 2X late**, ***late 3X early***, ***late 2X early***)

> E. Fanny herself limited **indeed**, she **minimised**, her office; you **didn't** need a jailor, she **contended**, for a **domesticated lamb** tied up with pink ribbon. This **wasn't** an animal to be **controlled** — it was an animal to be, at the most, educated. She admitted **accordingly** that she was **educative** — which Maggie was so **aware** that she herself **inevitably wasn't**; so it came round to being true that what she was most in charge of was his **mere intelligence**. This left, **goodness** knew, plenty of different calls for Maggie to **meet** — in a **case** in which so much pink ribbon, as it might be symbolically **named**, was lavished on the creature. What it all **amounted** to at any **rate** was that Mrs. Assingham would be **keeping** him quiet now, while his wife and his **father-in-law** carried out their own little ***frugal picnic***; **quite moreover**, **doubtless**, not much less neededly in respect to the members of the circle that were with them there than in respect to the pair they were **missing** almost for the first time. It was present to Maggie that the Prince **could** bear, when he was with his wife, almost any **queerness** on the part of people, strange English **types**, who bored him, **beyond convenience**, by being so little as he himself was; for this was one of the **ways** in which a wife was **practically sustaining**. But she was as **positively aware** that she **hadn't** yet ***learned*** to see him as meeting such **exposure** in her absence. How did he move and walk, how **above** all did he, or how WOULD he, look — he who with his so **nobly** handsome face **could** look such **wonderful** things — in **case** of being left alone with some of the subjects of his **wonder**? (**late 3X early**, **late 2X early**, ***early 3X late***, ***early 2X late***)

In Passage D, Felix, a young European, describes Boston and simultaneously characterizes himself, and in Passage E, in one of the James's extended metaphors, Fanny Assingham, a virtuoso of social perception, occupies the Prince while his wife Maggie and his father-in-law have some private time together.

The density of these characteristically early and late words is obviously much higher than that of the Zeta and Iota words, so that there are hundreds of densely characteristic passages. A thorough examination of these passages would shed a great deal of light on James's styles. Simply paging through the marked early and late novels, however, leaves some strong impressions. For example, many of the passages that are very dense in early words are, like Passage D, largely descriptive. And it might seem that description, by its very nature, requires a substantial proportion of concrete nouns and adjectives. Perhaps, then, one significant cause of the difference between the early and late style is that James became more interested in the psychological landscape and the social scene of his novels than in their physical settings. This reasonable hypothesis would require extensive testing, and Passage C shows that the later James can set a Seine-side scene in Paris in the springtime without much recourse to concrete physical detail. Note, too, how in Passage C Strether *dives deep, touches bottom, soars, sinks,* and *travels far* without moving. The late James typically focuses on intricate and delicate social relationships, just as he does in Passage E, where even *jailor* and *domesticated lamb tied up with pink ribbon* lose their concreteness and elaborately and metaphorically particularize Fanny’s activities and her complex relationship with the Prince.

If description figures as a sign of early James, and contrasts with the subtle and complex depiction of the social scene in the late James, it is also striking how many heavily late passages include dialogue. I have noted that the frequency of contractions increases generally in the later James (note the four in Passage E), but they are especially frequent in dialogue. The slang and colloquialism often noted in the later James are also naturally more frequent in dialogue. Indeed, the unusual combination of increased slang and colloquialism and increasingly convoluted and difficult syntax seem to me a hallmark of the later James. This is especially true when they appear together: “He’s ***prodigious***; but what **is** there — as ***you’ve*** fixed it — TO ***dodge***? ***Unless***,” she ***pursued***, “it’s her getting near him; it’s — if ***you’ll*** pardon my vulgarity — her getting AT him” (*The Golden Bowl*; marked as in Passage E). Here both of the medial interruptions set off with dashes contain characteristically late contractions and postpone the arrival of the *dodge* and *getting at him.*

On a smaller scale, the effects are often even more concentrated, as in the brief passages below, six from early novels followed by six from late novels (arranged in order of increasing prevalence of early or late words):

Early 3X late, early 2X late, *late 3X early*, *late 2X early*

1. They were all **standing** round his **sister**, as if they were **expecting** her to **acquit** herself of the ***exhibition*** of some **peculiar faculty**, some **brilliant talent**. Their attitude seemed to **imply** that she was a kind of **conversational mountebank**, attired, intellectually, in **gauze** and **spangles**.
   (*The Europeans*)
2. His **collegiate peccadilloes** had **aroused** a domestic **murmur** as **disagreeable** to the **young** man as the **creaking** of his **boots** would have been to a **house-breaker**. (*The Europeans*)
3. Florence in **midsummer** was perfectly void of **travelers**, and the dense little **city** gave forth its **aesthetic aroma** with a larger **frankness**, as the **nightingale sings** when the listeners have **departed**. (*Roderick Hudson*)
4. The long rain had **freshened** the air, and **twelve hours’** brilliant **sunshine** had **dried** the **roads**; . . . . (*The Europeans*)
5. She was **gentle**, accessible, tenderly **gracious**, **expressive**, **demonstrative**,

almost **flattering**. (*Confidence*)

6. He was **shamefully idle**, **spiritless**, **sensual**, snobbish. (*The American*)

Late 3X early, late 2X early, *early 3X late*, *early 2X late*

7. These things **shimmered** in the silver air of the **wondrous perspective** ahead, the ***region*** off there that **awaited** her present **approach** and where Gussy **hovered** like a bustling ***goddess*** in the **enveloping** cloud of her court. (*The Ivory Tower*)

8. The fine old **presence** on the ***pillow*** had **faltered** before expression; then it appeared rather **sighingly** and **finally** to give the **question** up. (*The Ivory Tower*)

9. Cissy, from below, her **charmingly** cool cove, had **watchfully signalled** up, and they **met afresh**, on the firm clear **sand** where the drowsy waves **scarce even** lapsed, with forms of intimacy that the **sequestered spot happily favoured**. (*The Ivory Tower*)

10. The **shopman**, who **hadn't** stirred, ***stood*** there in his patience — which, his **mute intensity helping**, had almost the **effect** of an **ironic comment**. (*The Golden Bowl*)

11. Her **straightness**, **visibly**, was all his own **loyalty could** ask. (*The Wings of the Dove*)

12. Strether, **consciously gaping** a little, had **fairly hung** ***upon*** her lips. (*The Ambassadors*)

It is little wonder that passages like these, with as many as four words in succession that are at least three times as frequent in the early novels as in the late novels, or vice versa, seem quintessentially early or late, and similar passages can be found in any early or late novel. In Passage two, "His collegiate peccadilloes had aroused a domestic murmur" seems almost stuffily formal, while the concrete and forceful description of the Passages three and four and the heavy use of adjectives in Passages five and six would be unusual in the later James. The late passages are more difficult to characterize, but verbs like *shimmered* and *hovered* (Passage seven), *faltered* (Passage eight), and *gaping* (Passage twelve) strike me as especially characteristic of the late James, as does the adverb *scarce* (Passage nine) and the adjective *ironic* (Passage ten). The *-ly* adverbs, especially rarities like *sighingly*, have already been mentioned, and the presence of as many as three of them in a single sentence, as in Passage nine, is quite telling.

5. Conclusion

Henry James certainly has a style, a style that resides not only in the tendency of his late novels toward convoluted syntax, but also in the frequencies of words of various kinds. It pervades his entire vocabulary, from the most frequent words in English to the rarest and most peculiarly Jamesian adverbs. But that distinctive style is not monolithic. Rather, it develops so gradually and consistently throughout his career that quantitative evidence from his use of words places his novels in almost perfectly chronological order.

Whether other authors display the almost astonishingly unidirectional stylistic development seen in James remains an open question, though preliminary research toward a more comprehensive study suggests that other authors do not generally show the same kind of easily interpretable trajectory. The early novels of Charles Dickens, for example, seem to develop chronologically, but those following *Dombey and Son* show no clear pattern. Willa Cather's early and late novels fall into very consistent patterns of similarity and difference, but those patterns are not chronological. However the styles of other authors develop, or even oscillate, stylometric techniques can profitably be applied to ensure that stylistic conclusions have a sound basis. Identifying words of exceptionally variable frequency, especially moderately frequent words, can direct our attention back toward the text. Finally, distinctively highlighting a large number of words that are distributed very differently in two or more authors or sub-styles can provide a vivid picture of the statistical and distributional nature of lexical style—a picture that suggests why and how readers can intuitively identify different styles.

I conclude with two extraordinary graphs that display the combined relative frequencies of the approximately 8300 words that occur three or more times as often in early as in late James or vice versa. Remember that these words have been selected on the basis of their average frequencies in all seven early and all eight late novels. Yet, as Figure 4 shows, the characteristically early words dwindle steadily, except for a surge in early words in *The Europeans,* until they reach an apparent base-line frequency in the late novels. The characteristically late words steadily increase from a similar base-line frequency in the early novels to a peak in the posthumous novel *The Ivory Tower.* They, too, follow a remarkably regular pattern, disrupted only by a large surge in late words in *What Maisie Knew*, and a smaller one in *The Sacred Fount.* These trends persist even through the four novels of the brief intermediate period that played no part in the creation of the word lists. That the slight anomalies displayed by *The Reverberator* and *What Maisie Knew* result from James's 1908 revisions is nicely confirmed by Figure 5, in which the same behavior can be seen for the original and New York editions of *The American, Daisy Miller,* and *The Portrait of a Lady.*

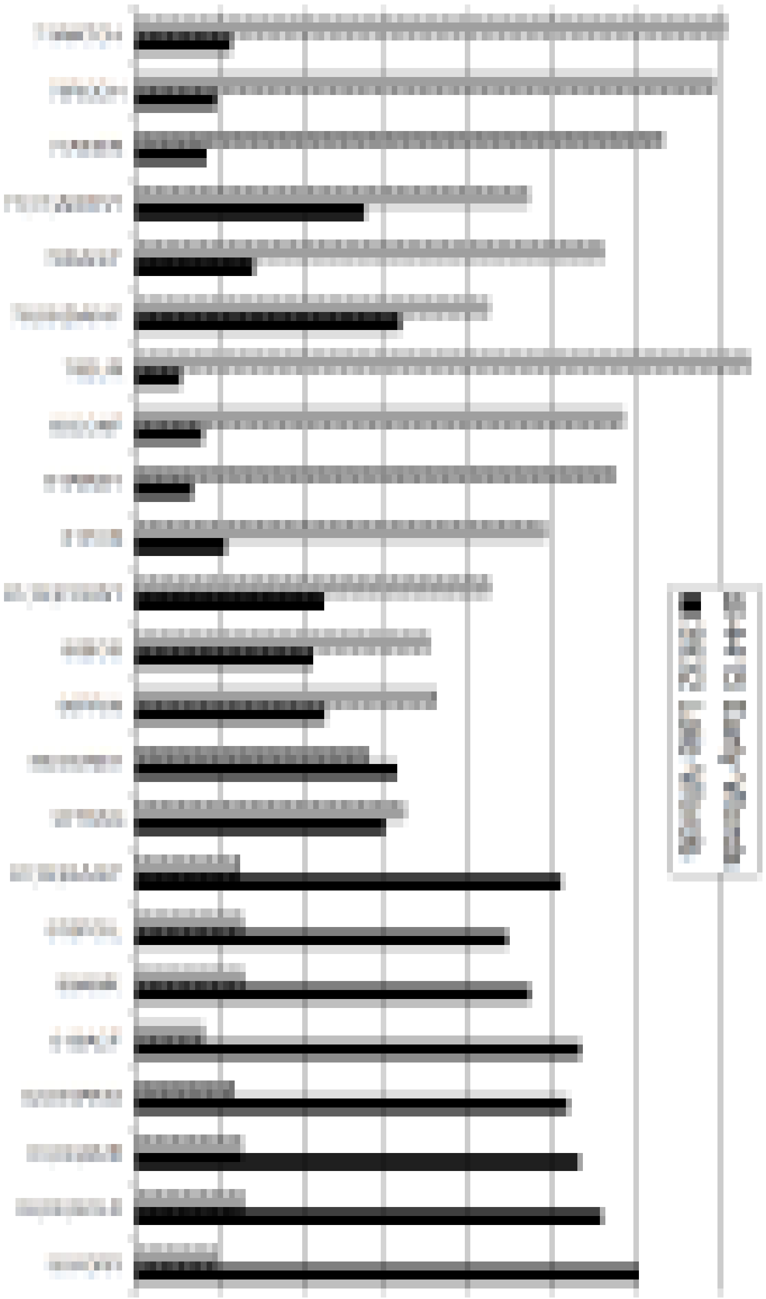

Fig. 4. Early and Late Words in Early, Intermediate, and Late Novels

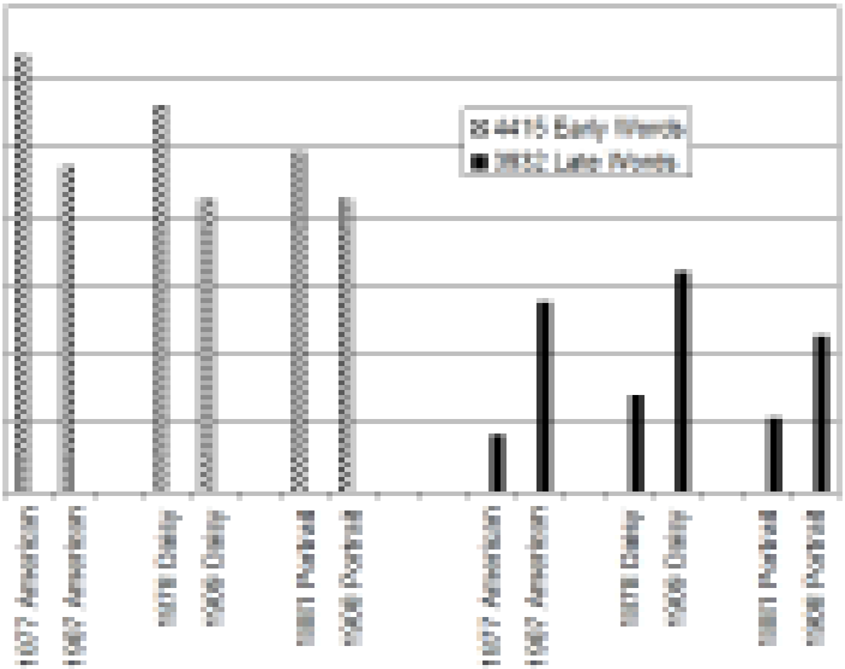

Fig. 5. Early and Late Words in Revised Early Novels

Much work remains to be done, both in adequately characterizing the styles of Henry James and in the study of style development more generally. We need better methods of visualizing and clarifying large-scale stylistic characteristics and changes. We need simpler ways of discovering passages of particular stylistic interest. We need, always, more direct maps leading revealingly and insightfully back to the text. Corpus-based and computer-assisted methods will surely be at the center of much of this work.

Notes

[1] For novels with two dates, the first is that of the New York edition, from which the e-text in my corpus is taken; the original publication date is in brackets. The e-texts for the other novels come from the original version. Both early and New York edition versions of *Daisy Miller, The American,* and *The Portrait of a Lady,* three of the most heavily revised early works (McWhirter 7), are included, so that the effects of James's extensive revisions can be examined briefly. An electronic text of the New York edition of *The American* was kindly supplied by John Burrows, who created it for a comparison of James and Austen ("Not Unless" 91-99). *The Outcry* has recently become available as an electronic text and will be included in future work on James's style. For more details, and the sources of the e-texts, see

Hoover, et al., chapter four.

Works Cited

Beach, Joseph Warren, ed. *The American*. New York: Holt, 1963.

Burrows, John F. “All the Way Through: Testing for Authorship in Different Frequency Strata.” *Literary and Linguistic Computing* 22 (2007): 27–47.

——. *Computation into Criticism.* Oxford: Clarendon Press, 1987.

——. “‘Delta’: A Measure of Stylistic Difference and a Guide to Likely Authorship.” *Literary and Linguistic Computing* 17 (2002): 267–87.

——. “The Englishing of Juvenal: Computational Stylistics and Translated Texts.” *Style* 36 (2002): 677–99.

——. “Not Unless You Ask Nicely: The Interpretive Nexus Between Analysis and Information.” *Literary and Linguistic Computing* 7 (1992): 91–109.

——. “Questions of Authorship: Attribution and Beyond.” *Computers and the Humanities* 37 (2003): 5–32.

——. “Who Wrote *Shamela*? Verifying the Authorship of a Parodic Text.” *Literary and Linguistic Computing* 20 (2005): 437–450.

Chatman, Seymour B. *The Later Style of Henry James*. Oxford: Blackwell. 1972.

Gettmann, Royal A. “Henry James’s Revision of *The American”*, *American Literature*, 16.4 (1945) 279–95.

Hoover, David L. “Altered Texts, Altered Worlds, Altered Styles.” *Language and Literature.* 13.2 (2004): 99–118.

——. “Delta Prime?” *Literary and Linguistic Computing* 19 (2004): 477–95.

——. “Multivariate Analysis and the Study of Style Variation.” *Literary and Linguistic Computing* 18 (2003): 341–60.

——. “Quantitative Analysis and Literary Studies.” *Blackwell Companion to Digital Literary Studies*. Ed. Ray Siemans and Susan Schreibman. Forthcoming, London: Blackwell, 2007.

——. “Testing Burrows’s Delta.” *Literary and Linguistic Computing* 19 (2004): 453–75.

——. “Word Frequency, Statistical Stylistics, and Authorship Attribution.” *Advanced ICT Methods Guide to Linguistics*. Ed. Tony McEnery. Forthcoming, 2007.

Hoover, David L., Jonathan Culpeper, Bill Louw, and Martin Wynne. *Approaches* *to Corpus Stylistics*. Forthcoming, London: Routledge, 2007.

Kenny, Anthony. *The Computation of Style*. Oxford: Pergamon, 1982.

Krause, Sydney J. 'James's Revisions of the Style of *The Portrait of a Lady.*' *American Literature* 30.1 (1958): 67–88.

Krook-Gilead, Dorothea. *The Ordeal of Consciousness in Henry James*. Cambridge: Cambridge UP, 1962.

Lee, Vernon. *The Handling of Words*. Lincoln: U of Nebraska P.1968.

Leech, Geoffrey N., and Short, Mick H. *Style in Fiction: A Linguistic Introduction to English Fictional Prose.* London: Longman, 1981.

Lodge, David. *Language of Fiction*. London: Routledge & K. Paul, 1966.

McWhirter, David, ed. *Henry James's New York Edition: The Construction of Authorship*. Stanford: Stanford UP, 1995.

Ohmann, Richard. "Generative Grammars and the Concept of Literary Style." *Linguistics and Literary Style.* Ed. D. C. Freeman. New York: Holt, Rinehart and Winston, 1970.

Poirier, Richard. *The Comic Sense of Henry James: A Study of the Early Novels.* New York: Oxford UP, 1960.

Rybicki, Jan. "Burrowing into Translation: Character Idiolects in Henryk Sienkiewicz's Trilogy and its Two English Translations." *Literary and Linguistic Computing* 21 (2006): 91–103.

Short, R.W. "The Sentence Structure of Henry James," *American Literature* 18.2 (1946): 71–88.

Stewart, Larry. "Charles Brockden Brown: Quantitative Analysis and Literary Interpretation." *LLC* 18 (2003): 129–38.

Van Doren, Carl. *The American Novel*. New York: Macmillan, 1921. 4 July 2005. <http://www.bartleby.com/187/11.html>.

Watt, Ian. "The First Paragraph of *The Ambassadors*." *The Ambassadors: An Authoritative Text, the Author on the Novel, Criticism* 2nd Ed. Ed. S. P. Rosenbaum. New York: Norton. 1994.

Appendix 1

The results reported above are intended to be comprehensible as they stand, but some readers may want further explanation and clarification of the stylometric methods used. I have avoided the terms *type* and *token* above to prevent confusion, but will use them below for precision and brevity. A word **Type** is defined here as a unique spelling (ignoring capitalization). This unfortunately treats unrelated homographs like the verb/noun *bear* and the proper name/common noun *Mark* as equivalent, but the corpus is too large for manual disambiguation of such words, and, fortunately, the huge numbers of words being analyzed tend to minimize their effects. Hyphenated words and contractions are treated as single types. A **Token** is an occurrence of a type. Thus there are forty-five words (tokens) in the fourth sentence of this paragraph, but only thirty-eight different words (types) because there are two tokens each of the types *and, noun, of,* and *words,* and four of the type *the.*

1. Delta

Burrows's Delta is designed for situations in which an anonymous text has many possible authors. He initially tests Delta on the 150 most frequent words of samples of English Restoration poetry ("Delta" 269–73), beginning with a primary set of twenty-five samples by twenty-five possible authors. My American novel corpus is organized similarly into fifteen primary samples by fifteen authors and fifty test novels by twenty authors, forty by ten of the primary authors and ten by ten other authors. These were downloaded from on-line collections, usually Project Gutenberg (http://www.gutenberg.org/), and lightly edited to remove Gutenberg information, prefaces, introductions, tables of contents, and so forth. To succeed, Delta must correctly attribute texts by primary authors but avoid incorrect attributions. My chronological focus here dictated the choice of primary samples and test novels for authors represented by more than two novels (with just two novels the choice is irrelevant). For authors represented by three novels, I selected the chronologically middle novel; for those represented by four or more novels, I combined an early and a late novel; for James I combined four novels from throughout his career. All words absent from all primary samples were removed from the word lists because they would produce a division by zero in the calculation of Delta and are irrelevant in any case. All personal pronouns were removed because they are closely related to the numbers and genders of characters, the proportion of dialogue to narration, and point of view and so can skew the results. All words for which a single text supplies more than seventy per cent of its occurrences were removed to eliminate most character and place names (for further information about this corpus, see Hoover, et al., chapter four).

Delta is a simple measure of textual difference, but its derivation can be confusing, so an example seems appropriate. Delta begins with lists of the frequencies of all types found in each text, arranged in descending frequency order. Consider *the,* the most frequent type in this corpus, with a mean frequency of 5.13% of all tokens in the primary samples. Its frequency is 7.12% in the primary sample *White Fang* and 3.46% in the test text *Alice Adams*. The calculation of Delta begins with the difference between each text and the mean of the primary samples: 1.99 for *White Fang* (7.12 - 5.13) and -1.67 for *Alice Adams* (3.46 - 5.13). Thus *the* is much more frequent than the mean in the former and much less frequent in the latter. The frequencies of the most frequent types fall rapidly: *the* constitutes about five per cent of all the tokens, but *had,* the tenth most frequent type, only about one per cent. So that differences in the frequencies of all words can have an effect, they are converted to z-scores by dividing the difference between the test text and the mean by the standard deviation of the word in the primary samples. This transforms the difference between the raw word-frequencies into a measure of the distance (in standard deviations) of each from the mean. The z-score for *the* is 2.20 in *White Fang* and -1.85 in *Alice Adams:* its frequency is 2.20 standard deviations above the mean in *White Fang* and 1.85 standard deviations below it in *Alice Adams*. Subtracting the z-score for *White Fang* from the z-score for *Alice Adams* gives a difference between the differences from the mean of -4.05 standard deviations (-1.85 - 2.20), showing that *the* is used extremely differently in *Alice Adams* and *White Fang*. This procedure is followed for all words included in the analysis, comparing each text sequentially with each primary sample and yielding a list of differences between the differences from the mean for each pair of texts. Because Delta is a measure of pure difference, the signs of the differences are eliminated before the final calculation, and Burrows defines his new measure as “the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text” (“Delta” 271). Once Delta is calculated for each test text, the primary author whose sample shows the smallest mean difference from the test text, the smallest Delta, is identified as its likeliest author.

2. Delta-Lz

Delta-Lz, my proposed modification of Delta, limits its calculation to words with relatively large (> 0.7) z-scores, words with frequencies nearly a standard deviation from the mean (for a detailed explanation, see my “Delta Prime?”; “Word Frequency”). It seems intuitively reasonable that such words may be more distinctive than those with frequencies nearly the same as the mean. Note, however, that Delta-Lz is based on a different word list for each test text, while all Delta calculations are based on the same word list.

3. Cluster Analysis

Cluster Analysis is an inductive statistical method of classification that is usually used when the numbers and sizes of the groups are not known in advance. Its results are normally presented as dendograms like Figure 1 and Figure 3. The technique has many different forms with varied results, and there is no space here to discuss this topic fully. The crucial point is that the closer to the bottom or left of the graph that items join together, the more statistically similar they are to each other. Conversely, the higher or further right two items join, the more different they are. Figure 1 shows that James's novels are very different from those of his contemporaries, and Figure 3 shows that the two versions of *The Portrait of a Lady* are, unsurprisingly, the most similar. Figure 1 and Figure 3 both show hierarchical cluster analyses with Ward linkage and squared Euclidean distance. Because the true authors of the texts in these analyses are known, the clusters are much easier to interpret then they would otherwise be. Cluster Analysis, because it requires no set of primary authors, can include all 71 novels in the American novel corpus or all of James's novels.

4. Principal Components Analysis

Principal Components Analysis, a statistical method for simplifying the description of a set of related variables, seems well suited to investigating the complexly interrelated frequencies of large numbers of words. It groups words with similar distributions together into new variables (principal components), each of which accounts for a larger proportion of the differences among the texts than do the frequencies of any one word. The results are typically presented as a scatter graph in which the distances between items reflect their differences. In Figure 2 the first principal component, which accounts for 24.8% of the total variance, is related closely to the dates of the novels: the closer to the top of the graph a novel falls, the later its composition. The relative positions of the early and revised editions of *The American, Daisy Miller,* and *The Portrait of a Lady* show that they are quite different from each other and that the revised editions are more like the later novels.

5. Distinctiveness Ratio

This is a simple measure of difference, but the creation of the list of 100 very distinctive words discussed above may need clarification. Beginning with a complete word list for the twenty novels (omitting the New York editions of the early novels) and removing personal pronouns and all words for which a single text supplies more than sixty per cent of the occurrences, I also removed words found in only one period and words that neither increase or decrease in frequency. From the remaining words that occur at least twenty times I selected the 500 with the most variable frequencies, as measured by the total Distinctiveness Ratio (the ratio of the maximum to the minimum frequency plus the ratio of the maximum to the median frequency). More than half the remaining words decrease or increase steadily in frequency from the early to the late style. I limited the list only to these words, and then eliminated differences resulting from changes in spelling conventions, function words, proper names, and words clearly related to content. “Words clearly related to content” is a somewhat subjective category, but *dressmaker* and *sculptor,* for example, are typically Jamesian ways to avoid repeating a character’s name, and *studio* is more likely to appear frequently in novels involving artists. Removing such words seems unproblematic because it works against the point I am making by weakening the separation of the three periods. Finally, I reduced the remaining list of 221 words to the 100 with the most variable frequencies, thirty-six of which decrease steadily in frequency and sixty-four of which increase steadily. The full list of 221 was used to search for additional word families.

6. Zeta and Iota

Both measures begin with a word frequency list for a primary text sample. This sample is divided into five equal sections, and a record is kept of how many of them contain each word. Calculating a Zeta score begins with words that occur in at least three of the five sections. In two-author tests, words are then removed that exceed a specific frequency in the works of the second author; in many-author tests, words are removed that appear in the samples of most of the other authors. Calculating an Iota score begins with all words that appear in just one or two of the five sections. In two-author tests, words that appear in the second author’s sample are removed; in many-author tests, words are removed that appear in more than about half the other authors. The Zeta or Iota score is the total frequency (in tokens per 1000 words) of all words that remain after the stipulations are made, and the higher the Zeta or Iota score of a text, the more likely the primary author wrote it (Burrows, “All the Way Through”).

Further study of these two measures will be needed, but Burrows shows that Zeta and Iota are effective for poems by Waller and Marvell (“All the Way Through”), and preliminary tests on 20th century American poetry show strong results (Hoover, “Quantitative Analysis”). Zeta and Iota seem to be capturing genuine authorial idiosyncrasies, and Zeta, based on words of moderate frequency that are used at very different rates by different authors, seems especially promising for investigating the lexical aspects of style. Zeta is such a new measure of similarity that no precise methodological principles have been established, but its general applicability to James’s stylistic chronology seems clear. Zeta analysis requires that some texts be set aside as a primary authorial set against which other texts are tested, and I limit myself here to tests with early and late James as the primary "author." (The intermediate novels, a transitional group of only four novels, are a less attractive focus.) For Zeta analysis, I divided longer novels into sections and omitted *Daisy Miller* to reduce the differences in text sizes, and omitted the revised early novels. The resulting thirty-four sections and novels range in size from *The Reverberator* at about 53,000 words to *What Maisie Knew* at about 95,000 words. The primary texts for early James consist of four novels of moderate size from throughout the early period (*Watch and Ward, The Europeans, Confidence, Washington Square*) and one of the two sections of *Roderick Hudson*. The counter set consists of twenty-three sections of intermediate and late novels, and the test set consists of the other six sections of the early novels. Beginning with words found in at least three of the five early sections, I deleted those found in more than six of the counter samples, after which only 597 words remained. All six early test sections have higher Zeta scores than any of the intermediate or late sections; that is, they have higher relative frequencies of these words than do any of the counter samples. As words are deleted which appear in a minimum of five, four, three, two, and one of the counter samples, the results continue to be correct, and the difference between the lowest score for any early sample and the highest score for any counter sample increases to a maximum for the 161 words that appear in at least three of the primary samples but in no more than two of the counter samples. These results are impressive, especially for a technique designed to distinguish different authors rather than the different styles of a single author.

An examination of these 161 words reveals a few proper nouns that seem inappropriate as markers of the early and later styles. There are also many words like *honor, civilization,* and *self-defence*, for which James uses different spellings in early and late novels. Though good markers of early texts, these are not very interesting stylistically, and they seem especially vulnerable to editorial alteration. When proper nouns and spelling variations are removed, one intermediate text ties with one early text. Limiting the list to words that occur at least six times in the primary set before applying the same stipulations yields the list of 122 words discussed above after proper nouns and spelling variants are removed. With *The Ivory Tower, What Maisie Knew, The Sacred Fount,* and the first sections of *The Awkward Age* and *The Ambassadors as the primary set, similar restrictions yield the group of 137 characteristic late Zeta words discussed above.*

The expanded lists of 446 early and 673 late Zeta words and 4952 early and 4696 late Iota words discussed above were created using the seven early and eight late *novels as primary sets. Stipulations similar to those described above give excellent Zeta and Iota results using these expanded word lists.*