# Dickens's Narrative Style: A Statistical Approach to Chronological Variation*

Tomoji TABATA

**Résumé.** Cette communication est consacrée aux analyses statistiques de la langue romanesque de Dickens. Le problème spécifique est de trouver des marqueurs linguistiques de la chronologie dans les narrations de l'auteur. Dans les analyses qui suivent, je me sers de l'analyse du composant principal et du « test *t* de Student » pour examiner la distribution des fréquences de vocables les plus communs dans certains extraits de son œuvre. Cette approche fait ressortir le contraste entre les premières œuvres et les plus récentes. Les facteurs différentiateurs entre les premières narrations et les suivantes suggèrent une mutation du complexe style formel en style oral.

**Keywords:** Language of Dickens, chronology, style, narrative, principal component analysis, Student's *t*-test, computational stylistics, literary statistics, very common words.

**Mots-clés :** Langue de Dickens, chronologie, style, narration, analyse des principaux composants, test *t* de Student, stylistique computationnelle, statistiques littéraires, vocables communs.

## 1. Introduction

This essay is a sequel to the approach that I adopted in my previous paper (Tabata: 1993), in which I applied statistical methods to the stylistic study of four volumes of literary texts by Dickens. Two of the four were published in his twenties: *Sketches by Boz* (1836) and *Oliver Twist* (1837–8), while the *Christmas Books* were written in his thirties (1843–8). The last of the set was his last and unfinished novel, *The Mystery of Edwin Drood* (1870). The results of this

* This study came out of a group project being pursued at the Centre for Literary and Linguistic Computing, the University of Newcastle, NSW, Australia. I would like to thank John Burrows, Alexis Antonia, Hugh Craig, and Wayne McKenna for technical advice given at all times, and to Nicole Cox for her service in the preparation of texts. I am also grateful to John Lambert for software support.

✉ Ariake National College of Technology; 150 Higashihagio-machi; Omuta; Fukuoka 836 (Japan).
Fax: + 81 944 53-1361 E-mail: tabata@ariake-nct.ac.jp

analysis suggested gradual changes of Dickens's style in some dimensions, such as elaboration, pronominal reference, nominalisation, and so on, though not in exact chronological terms. While the set I studied was not large enough to allow any chronological generalisation, the length of his career (from his early twenties to his death at the age of fifty-eight) encourages a longitudinal study of style. The purpose of this paper therefore is to present a test for chronological variation of Dickens's style by enlarging the size and the scope of text corpus, with due modification of the methods.

Key to Tables and Graphs

| Label | Narrator [*Text*] & Date | | Word-tokens [Pure-Narrative] | Segments |
|---|---|---|---|---|
| Works Written in 1830s | | | | |
| SB#1–3 | Sketches by Boz | (1836) | 12569 | 3 |
| PP#1–2 | The Pickwick Papers | (1836–7) | 11081 | 2 |
| OT#1–4 | Oliver Twist | (1837–8) | 16677 | 4 |
| NN#1–3 | Nicholas Nickleby | (1838–9) | 12863 | 3 |
| Works Written after 1849 | | | | |
| David#1–5 | David [*David Copperfield*] | (1849–50) | 20145 | 5 |
| Esther#1–4 | Esther [*Bleak House*] | (1852–3) | 18399 | 4 |
| BH#1–2 | Bleak House | (1852–3) | 7389 | 2 |
| TTC#1–3 | A Tale of Two Cities | (1859) | 12798 | 3 |
| Pip#1–4 | Pip [*Great Expectations*] | (1860–1) | 18359 | 4 |
| OMF#1–3 | Our Mutual Friend | (1864–5) | 13117 | 3 |
| ED#1–3 | The Mystery of Edwin Drood | (1870) | 11973 | 3 |

| Group | | Word-tokens [Pure-Narrative] | Segments |
|---|---|---|---|
| 1 | Early Dickens: Third-person Narrative | 53190 | 12 |
| 2 | Late Dickens: First-person Narrative | 56903 | 13 |
| 3 | Late Dickens: Third-person Narrative | 45277 | 11 |

The corpus draws on four novels written in the 1830s as an early subset, and six novels written after 1849, when he started *David Copperfield*, as a later subset (see Key to Tables and Graphs)[1]. An interval of ten years was allowed between the two sets in an effort to make interpretation of the results

[1] The Oxford Illustrated Dickens, 21 vols (London: OUP) is the source of the copytexts for:
- *Sketches by Boz*: 1989 [1st p.: 1836]
- *David Copperfield*: 1987 [1st p.: 1850]

as straightforward as possible. As a result, the *Christmas Books* were forced out of the set. Each text is represented by approximately twenty thousand words from the beginning of the novel, and the language of "pure-narrative" is extracted as a basis for comparison[2]. *Bleak House*, since it is recounted by two irregularly alternating narrators, provides two narratives: one is the first-person retrospective narrative by Esther Summerson, who uses the past tense, while the other is recounted by an anonymous third-person narrator, who constantly uses the present tense.

*David Copperfield, Great Expectations*, and Esther's narrative are chosen with an eye to first-person narrative. The current corpus, seen in terms of person, consists of three first-person narratives and eight third-person narratives. Three shorter first-person narratives, "The Stroller's Tale", "The Convict's Return", which are interpolated in *The Pickwick Papers*, and Miss Wade's narrative, "The History of a Self-Tormentor", which is embedded in *Little Dorrit* (1855–7), will also be added to analysis in due course[3]. Of the eleven main narratives, three are narrated in the present tense (*Sketches by Boz* and *Edwin Drood* in addition to the anonymous narrative of *Bleak House*).

Since differences in point of view are obvious between first-person narratives and third-person narratives, between narratives told in the past tense and those in the present tense, it is desirable to exclude a number of words most

– *Great Expectations*: 1992 [1st p.: 1860]
– *Our Mutual Friend*: 1989 [1st p.: 1865]

The Penguin English Library (Harmondsworth: Penguin Books) is used as copytexts for:
– *The Pickwick Papers*, ed. Robert L. Patten: 1986 [1st p.: 1837]
– *Nicholas Nickleby*, ed. Michael Slater: 1978 [1st p.: 1839]
– *Oliver Twist*, ed. Peter Fairclough: 1985 [1st p.: 1838]
– *Bleak House*, ed. Norman Page: 1987 [1st p.: 1853]
– *A Tale of Two Cities*, ed. George Woodcock: 1970 [1st p.: 1859]
– *The Mystery of Edwin Drood*, ed. Arthur J. Cox: 1985 [1st p.: 1870].

[2] In principle, a chapter ending nearest to twenty-thousand words was taken as the end of sample in this study. For *Bleak House*, however, the end of sample was prolonged to the end of chapter seven to generate two narrative samples. The size of narrative text differs considerably from text to text according to the portion dialogue and free indirect discourse occupy in each sample.

The distinction between narrative and dialogue adopted in this paper is based upon the presence of quotation marks, a crude but tolerably objective distinction. Within the domain of narrative there is a further need to distinguish "pure-narrative" and the special category of fictional discourse, which is often referred to as "free indirect discourse" (FID), and in which the "voice" of the narrator and that of the character that s/he reports are merged into a hybrid style. How and where to set the boundary between pure-narrative and FID is a touchy question: but I base my distinction between pure-narrative and free indirect discourse on my interpretation of the texts.

[3] John Burrows was kind enough to offer the machine-readable set of these three narratives, which were used as Dickens's text-base in his analysis of first person retrospective narratives. See BURROWS (1992 and 1994).

Table 1

Eleven Narrators in Dickens's Novels: Standardised (text-percentage) frequencies for the seventy-four very common word-types in the "pure-narrative."

(For the column labelled "Total", frequencies are given in raw score.)

| Rank | Word-types* | SB | PP | OT | NN | David | Esther | BH | TTC | Pip | OMF | ED | Total (raw) | Mean (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | the | 7.606 | 9.097 | 7.327 | 6.320 | 4.433 | 4.723 | 6.834 | 7.462 | 5.817 | 6.602 | 6.515 | 9935 | 6.394 |
| 2 | and | 3.914 | 4.088 | 3.598 | 4.019 | 3.927 | 4.310 | 3.424 | 4.329 | 4.210 | 3.690 | 3.792 | 6164 | 3.967 |
| 3 | of | 4.225 | 3.592 | 3.190 | 3.281 | 2.636 | 2.462 | 3.424 | 3.469 | 2.511 | 3.255 | 3.566 | 4879 | 3.140 |
| 4 | a | 2.912 | 2.346 | 2.908 | 3.001 | 2.442 | 2.571 | 2.774 | 2.508 | 2.495 | 3.171 | 2.773 | 4194 | 2.699 |
| 5 | in | 2.164 | 1.661 | 1.847 | 1.788 | 1.812 | 1.853 | 2.463 | 2.016 | 1.672 | 2.173 | 2.096 | 2983 | 1.920 |
| 6 | to(inf) | 1.201 | 1.101 | 1.325 | 1.314 | 1.524 | 1.549 | 1.340 | 1.188 | 1.416 | 1.189 | 1.111 | 2055 | 1.323 |
| 7 | with | 1.154 | 1.263 | 1.091 | 1.104 | 1.052 | 1.277 | 0.920 | 1.274 | 1.149 | 1.395 | 1.336 | 1841 | 1.185 |
| 8 | to(prep) | 1.034 | 1.119 | 1.091 | 1.026 | 1.176 | 1.163 | 0.988 | 1.203 | 1.024 | 1.235 | 1.169 | 1736 | 1.117 |
| 9 | it | 0.676 | 0.605 | 0.768 | 0.700 | 1.365 | 1.076 | 1.272 | 1.399 | 1.285 | 1.113 | 1.052 | 1624 | 1.045 |
| 10 | as | 0.692 | 0.957 | 0.851 | 1.011 | 1.082 | 0.848 | 0.826 | 0.938 | 1.008 | 1.037 | 1.128 | 1476 | 0.950 |
| 11 | at | 0.756 | 0.496 | 0.738 | 0.910 | 1.023 | 0.962 | 0.758 | 0.836 | 1.100 | 0.953 | 0.618 | 1337 | 0.861 |
| 12 | that(conj) | 0.549 | 0.388 | 0.660 | 0.599 | 0.933 | 0.989 | 0.595 | 0.484 | 1.002 | 0.602 | 0.501 | 1098 | 0.707 |
| 13 | on(prep) | 0.835 | 0.713 | 0.522 | 0.575 | 0.660 | 0.554 | 0.555 | 0.766 | 0.757 | 0.724 | 0.727 | 1040 | 0.669 |
| 14 | by | 0.525 | 0.578 | 0.672 | 0.536 | 0.457 | 0.435 | 0.528 | 0.524 | 0.452 | 0.640 | 0.585 | 826 | 0.532 |
| 15 | which(r) | 0.812 | 0.641 | 0.762 | 0.669 | 0.417 | 0.424 | 0.420 | 0.399 | 0.381 | 0.435 | 0.309 | 794 | 0.511 |
| 16 | for(prep) | 0.732 | 0.415 | 0.570 | 0.459 | 0.491 | 0.484 | 0.568 | 0.328 | 0.479 | 0.343 | 0.543 | 762 | 0.490 |
| 17 | but | 0.422 | 0.289 | 0.414 | 0.342 | 0.660 | 0.582 | 0.555 | 0.445 | 0.523 | 0.450 | 0.317 | 729 | 0.469 |
| 18 | not | 0.501 | 0.262 | 0.336 | 0.327 | 0.551 | 0.554 | 0.447 | 0.391 | 0.523 | 0.381 | 0.267 | 664 | 0.427 |
| 19 | be | 0.461 | 0.162 | 0.444 | 0.435 | 0.382 | 0.419 | 0.663 | 0.305 | 0.392 | 0.335 | 0.326 | 603 | 0.388 |
| 20 | from | 0.485 | 0.478 | 0.402 | 0.443 | 0.308 | 0.326 | 0.406 | 0.484 | 0.376 | 0.267 | 0.326 | 595 | 0.383 |
| 21 | when | 0.159 | 0.343 | 0.342 | 0.350 | 0.536 | 0.462 | 0.392 | 0.344 | 0.485 | 0.252 | 0.309 | 585 | 0.377 |
| 22 | this | 0.294 | 0.307 | 0.546 | 0.498 | 0.382 | 0.217 | 0.298 | 0.313 | 0.338 | 0.442 | 0.451 | 579 | 0.373 |
| 23 | all | 0.326 | 0.171 | 0.300 | 0.420 | 0.367 | 0.451 | 0.352 | 0.336 | 0.479 | 0.328 | 0.334 | 561 | 0.361 |
| 24 | an | 0.493 | 0.433 | 0.348 | 0.233 | 0.308 | 0.342 | 0.392 | 0.336 | 0.289 | 0.450 | 0.443 | 560 | 0.360 |

* (a) = adjective, (adv) = adverbials, (adv.d.) = adverb of degree, (conj) = conjunction, (dem) = demonstrative, (inf) = infinitive, (r) = relative, (prep) = preposition.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.

Table 1 (continued)

| Rank | Word-types* | SB | PP | OT | NN | David | Esther | BH | TTC | Pip | OMF | ED | Total (raw) | Mean (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 25 | or | 0.398 | 0.153 | 0.372 | 0.365 | 0.357 | 0.255 | 0.379 | 0.305 | 0.283 | 0.381 | 0.342 | 505 | 0.325 |
| 26 | out | 0.080 | 0.199 | 0.222 | 0.194 | 0.506 | 0.364 | 0.487 | 0.391 | 0.414 | 0.282 | 0.342 | 503 | 0.324 |
| 27 | there | 0.294 | 0.217 | 0.288 | 0.319 | 0.387 | 0.375 | 0.365 | 0.328 | 0.327 | 0.198 | 0.234 | 480 | 0.309 |
| 28 | into | 0.382 | 0.208 | 0.408 | 0.365 | 0.268 | 0.217 | 0.392 | 0.352 | 0.283 | 0.160 | 0.393 | 474 | 0.305 |
| 29 | who(r) | 0.358 | 0.253 | 0.384 | 0.404 | 0.218 | 0.223 | 0.514 | 0.211 | 0.245 | 0.435 | 0.134 | 457 | 0.294 |
| 29 | one | 0.446 | 0.235 | 0.282 | 0.350 | 0.323 | 0.239 | 0.217 | 0.367 | 0.234 | 0.236 | 0.309 | 457 | 0.294 |
| 31 | that(dem) | 0.239 | 0.280 | 0.222 | 0.334 | 0.472 | 0.239 | 0.203 | 0.273 | 0.272 | 0.274 | 0.259 | 447 | 0.288 |
| 32 | very | 0.223 | 0.244 | 0.462 | 0.365 | 0.338 | 0.413 | 0.271 | 0.211 | 0.185 | 0.175 | 0.150 | 445 | 0.286 |
| 33 | if | 0.207 | 0.208 | 0.186 | 0.272 | 0.377 | 0.288 | 0.352 | 0.211 | 0.468 | 0.198 | 0.284 | 443 | 0.285 |
| 34 | little | 0.199 | 0.190 | 0.222 | 0.404 | 0.338 | 0.413 | 0.298 | 0.266 | 0.169 | 0.229 | 0.384 | 442 | 0.284 |
| 35 | up | 0.151 | 0.280 | 0.288 | 0.327 | 0.333 | 0.228 | 0.257 | 0.289 | 0.376 | 0.282 | 0.200 | 435 | 0.280 |
| 36 | so(adv.d.) | 0.151 | 0.126 | 0.174 | 0.187 | 0.283 | 0.554 | 0.230 | 0.227 | 0.278 | 0.175 | 0.242 | 394 | 0.254 |
| 37 | upon | 0.191 | 0.190 | 0.228 | 0.334 | 0.268 | 0.212 | 0.284 | 0.336 | 0.240 | 0.206 | 0.234 | 382 | 0.246 |
| 38 | no(a) | 0.326 | 0.153 | 0.258 | 0.210 | 0.223 | 0.163 | 0.244 | 0.242 | 0.267 | 0.244 | 0.192 | 356 | 0.229 |
| 39 | down | 0.088 | 0.117 | 0.198 | 0.109 | 0.228 | 0.207 | 0.203 | 0.313 | 0.272 | 0.252 | 0.209 | 318 | 0.205 |
| 40 | some | 0.263 | 0.135 | 0.210 | 0.179 | 0.243 | 0.158 | 0.203 | 0.211 | 0.218 | 0.168 | 0.234 | 316 | 0.203 |
| 41 | more | 0.239 | 0.180 | 0.210 | 0.187 | 0.194 | 0.207 | 0.135 | 0.180 | 0.196 | 0.107 | 0.192 | 292 | 0.188 |
| 42 | old | 0.255 | 0.162 | 0.294 | 0.117 | 0.114 | 0.304 | 0.284 | 0.094 | 0.065 | 0.114 | 0.284 | 287 | 0.185 |
| 43 | man | 0.294 | 0.343 | 0.126 | 0.155 | 0.055 | 0.147 | 0.176 | 0.133 | 0.283 | 0.259 | 0.125 | 285 | 0.183 |
| 44 | then | 0.175 | 0.262 | 0.126 | 0.124 | 0.164 | 0.136 | 0.041 | 0.164 | 0.278 | 0.198 | 0.284 | 281 | 0.181 |
| 45 | before | 0.223 | 0.108 | 0.168 | 0.163 | 0.179 | 0.207 | 0.149 | 0.211 | 0.174 | 0.229 | 0.117 | 277 | 0.178 |
| 46 | other | 0.271 | 0.190 | 0.174 | 0.264 | 0.129 | 0.114 | 0.108 | 0.211 | 0.153 | 0.206 | 0.117 | 269 | 0.173 |
| 47 | over | 0.167 | 0.153 | 0.192 | 0.093 | 0.129 | 0.125 | 0.108 | 0.250 | 0.174 | 0.267 | 0.200 | 262 | 0.169 |
| 48 | again | 0.127 | 0.171 | 0.108 | 0.086 | 0.238 | 0.168 | 0.149 | 0.203 | 0.153 | 0.221 | 0.192 | 260 | 0.167 |
| 49 | its | 0.247 | 0.171 | 0.060 | 0.155 | 0.079 | 0.092 | 0.325 | 0.328 | 0.093 | 0.145 | 0.359 | 258 | 0.166 |
| 50 | that(r) | 0.215 | 0.072 | 0.090 | 0.124 | 0.194 | 0.130 | 0.217 | 0.344 | 0.153 | 0.160 | 0.167 | 258 | 0.166 |
| 51 | time | 0.151 | 0.126 | 0.180 | 0.225 | 0.208 | 0.158 | 0.135 | 0.133 | 0.212 | 0.084 | 0.125 | 255 | 0.164 |

* (a) = adjective, (adv) = adverbials, (adv.d.) = adverb of degree, (conj) = conjunction, (dem) = demonstrative, (inf) = infinitive, (r) = relative, (prep) = preposition.

Extrait de la Revue Informatique et Statistique dans les Sciences humaines
XXX, 1 à 4, 1994. C.I.P.L. - Université de Liège - Tous droits réservés.

Table 1 (continued)

| Rank | Word-types* | SB | PP | OT | NN | David | Esther | BH | TTC | Pip | OMF | ED | Total (raw) | Mean (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 52 | being | 0.088 | 0.090 | 0.222 | 0.225 | 0.189 | 0.130 | 0.149 | 0.109 | 0.158 | 0.168 | 0.251 | 255 | 0.164 |
| 53 | two | 0.239 | 0.153 | 0.120 | 0.272 | 0.114 | 0.125 | 0.108 | 0.250 | 0.120 | 0.145 | 0.184 | 251 | 0.162 |
| 54 | than | 0.095 | 0.099 | 0.168 | 0.109 | 0.174 | 0.158 | 0.257 | 0.219 | 0.191 | 0.114 | 0.184 | 248 | 0.160 |
| 55 | about | 0.167 | 0.090 | 0.114 | 0.179 | 0.169 | 0.201 | 0.149 | 0.148 | 0.207 | 0.091 | 0.175 | 245 | 0.158 |
| 56 | head | 0.064 | 0.126 | 0.204 | 0.117 | 0.169 | 0.163 | 0.068 | 0.219 | 0.125 | 0.206 | 0.226 | 245 | 0.158 |
| 57 | gentleman | 0.064 | 0.334 | 0.438 | 0.272 | 0.070 | 0.163 | 0.068 | 0.078 | 0.000 | 0.122 | 0.033 | 232 | 0.149 |
| 58 | what | 0.095 | 0.054 | 0.114 | 0.187 | 0.169 | 0.217 | 0.122 | 0.094 | 0.202 | 0.114 | 0.150 | 226 | 0.145 |
| 59 | after | 0.048 | 0.144 | 0.186 | 0.140 | 0.204 | 0.141 | 0.081 | 0.109 | 0.169 | 0.099 | 0.150 | 220 | 0.142 |
| 60 | much | 0.064 | 0.126 | 0.138 | 0.194 | 0.134 | 0.136 | 0.135 | 0.078 | 0.196 | 0.198 | 0.134 | 220 | 0.142 |
| 61 | any | 0.199 | 0.081 | 0.120 | 0.086 | 0.223 | 0.130 | 0.217 | 0.156 | 0.131 | 0.084 | 0.117 | 219 | 0.141 |
| 62 | face | 0.072 | 0.081 | 0.150 | 0.132 | 0.134 | 0.136 | 0.068 | 0.211 | 0.076 | 0.183 | 0.292 | 217 | 0.140 |
| 63 | great | 0.103 | 0.180 | 0.216 | 0.233 | 0.089 | 0.174 | 0.149 | 0.125 | 0.093 | 0.099 | 0.058 | 213 | 0.137 |
| 64 | hand | 0.040 | 0.153 | 0.120 | 0.086 | 0.124 | 0.082 | 0.054 | 0.352 | 0.142 | 0.160 | 0.150 | 207 | 0.133 |
| 65 | like(prep) | 0.080 | 0.072 | 0.048 | 0.047 | 0.159 | 0.136 | 0.122 | 0.188 | 0.261 | 0.168 | 0.100 | 204 | 0.131 |
| 66 | eyes | 0.048 | 0.144 | 0.156 | 0.140 | 0.104 | 0.109 | 0.027 | 0.227 | 0.136 | 0.137 | 0.175 | 202 | 0.130 |
| 67 | mother | 0.056 | 0.000 | 0.024 | 0.016 | 0.874 | 0.005 | 0.027 | 0.000 | 0.011 | 0.008 | 0.050 | 201 | 0.129 |
| 68 | such | 0.151 | 0.117 | 0.060 | 0.155 | 0.169 | 0.212 | 0.135 | 0.125 | 0.087 | 0.076 | 0.075 | 196 | 0.126 |
| 69 | on(adv) | 0.103 | 0.117 | 0.096 | 0.187 | 0.099 | 0.114 | 0.068 | 0.086 | 0.163 | 0.114 | 0.167 | 188 | 0.121 |
| 70 | back | 0.088 | 0.063 | 0.114 | 0.086 | 0.134 | 0.168 | 0.054 | 0.117 | 0.142 | 0.122 | 0.159 | 186 | 0.120 |
| 71 | way | 0.056 | 0.099 | 0.150 | 0.070 | 0.144 | 0.158 | 0.054 | 0.156 | 0.131 | 0.084 | 0.117 | 183 | 0.118 |
| 72 | young | 0.048 | 0.090 | 0.150 | 0.132 | 0.025 | 0.136 | 0.176 | 0.141 | 0.114 | 0.099 | 0.251 | 183 | 0.118 |
| 73 | never | 0.143 | 0.000 | 0.066 | 0.054 | 0.179 | 0.228 | 0.217 | 0.023 | 0.136 | 0.061 | 0.125 | 181 | 0.117 |
| 74 | now | 0.151 | 0.045 | 0.078 | 0.062 | 0.114 | 0.141 | 0.095 | 0.117 | 0.163 | 0.160 | 0.092 | 178 | 0.115 |

* (a) = adjective, (adv) = adverbials, (advd.) = adverb of degree, (conj) = conjunction, (dem) = demonstrative, (inf) = infinitive, (r) = relative, (prep) = preposition.

obviously influenced by the choice of point of view, such as personal pronouns and finite verbs, from statistical analysis so as to diminish the overshadowing effect of what is already evident. Otherwise vectors that differentiate first-person narratives from third-person narratives and texts couched in the past tense from those in the present tense will be so strong that other determinants of style may be submerged. This exclusion of personal pronouns and finite verbs deprives my data of some of the interesting subjects for computational stylistics[4]. In return, however, it makes them sensitive to evidence of subtler stylistic differences.

## 2. Analysis and Interpretation

### 2.1. Principal component analysis of the 74 very common word-types

The following statistical work is based on the seventy-four very common word-types (as listed in Table 1) whose occurrence in segments of text is frequent enough to allow multivariate analysis[5]. Initially I selected one hundred commonest words, which included thirteen personal pronouns and as many finite verbs[6]. Even after these twenty-six words are excluded, the remaining seventy-four words still account for more than 40 % of all the word-tokens in the pure-narrative.

Fig. 1 shows the results of a principal component analysis (PCA) of the seventy-four words in the texts divided into thirty-six 4,000-word segments. Segmentation of text, one of modifications adopted in this approach, has two objectives. First, to give each variable (word) as appropriate a number of

[4] Works by BRAINERD (1979), BURROWS (1987a), and CRAIG (1991) show stimulating results of statistical analyses of a set of personal pronouns.

[5] "Word-type" will be referred to as "word" henceforward unless it is necessary to distinguish. It may be necessary to add that considerable amounts of time and efforts are needed for post-editing of texts apart from proof-reading: for the purpose of counting, contractions are expanded (*e.g.* "can't" counts as "can" and "not"), proper names like "Mr Copperfield", on the other hand, are united with asterisk ("Mr*Copperfield") so as to count as one word. Words that are usually hyphenated or treated as one word but do not appear as one word in my texts are also joined ("for ever" is joined as "for*ever"). Furthermore, common homographic forms are tagged to specify each usage: for example, "that" is separated into "that(c)" (conjunctive), "that(r)" (relative), "that(d)" (demonstrative), and so on.

[6] Personal pronouns and verbs that ranked within the top 100 word-types are: *I, me, my, we, he, him, his, himself, she, her, they, them, their, is, was, were, been, have, had, came, looked, made, said, replied, could*, and *would*.

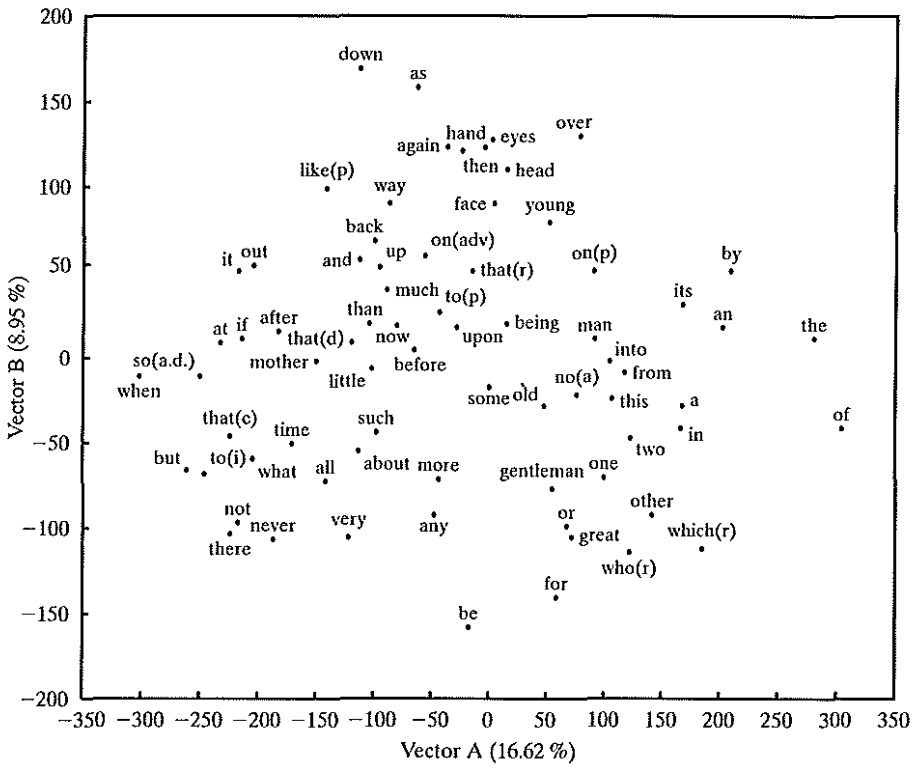


Fig. 1.– Eleven Narrators in Dicken's Novels: Word-plot
(for the 74 very common word-types in the corpus)

samples as possible in order to reduce the possibility of chance effect. Second, to help observe internal variation (or consistency) in each text. The first step of PCA is to measure correlation of each of the seventy-four words with each of the other seventy-three across the thirty-six segments of text[7]. In the process, a matrix of 2,701 correlation coefficients is generated. The next step is to subject the resulting correlation matrix to eigen-analysis. By eigen analysis, the principal components of the matrix are extracted, and it is possible to project the most powerful components in a scatter diagram like Fig. 1.

The graph gives a picture of the reciprocal relationships among the seventy-four words. Relative distance between the entries reflects similarity or contrast among these words in their concomitant variation over the thirty-six

[7] All the calculation in this study was undertaken using *MINITAB, Release 8.2 Accelerated*, The *Apple Macintosh* Version. (Minitab Inc, State College, PA, 1991).

text segments. Words located towards the east and those located towards the west of the graph tend to be mutually opposed: when the frequency of one set of words rises high in a given text-sample, the frequency of the other tends to fall low. The same applies to the vertical axis. To give a crude example, where the relative frequency of "of" runs high "the" also tends to occur frequently, while the relative incidence of words like "when" and "so" (adverb of degree) tend to fall away. The final step is to multiply the eigen-matrix that produced Fig. 1 back through the original frequency table for the 36 segments of text. The purpose is to distribute the text segments on the word-pattern. This procedure produces Fig. 2.

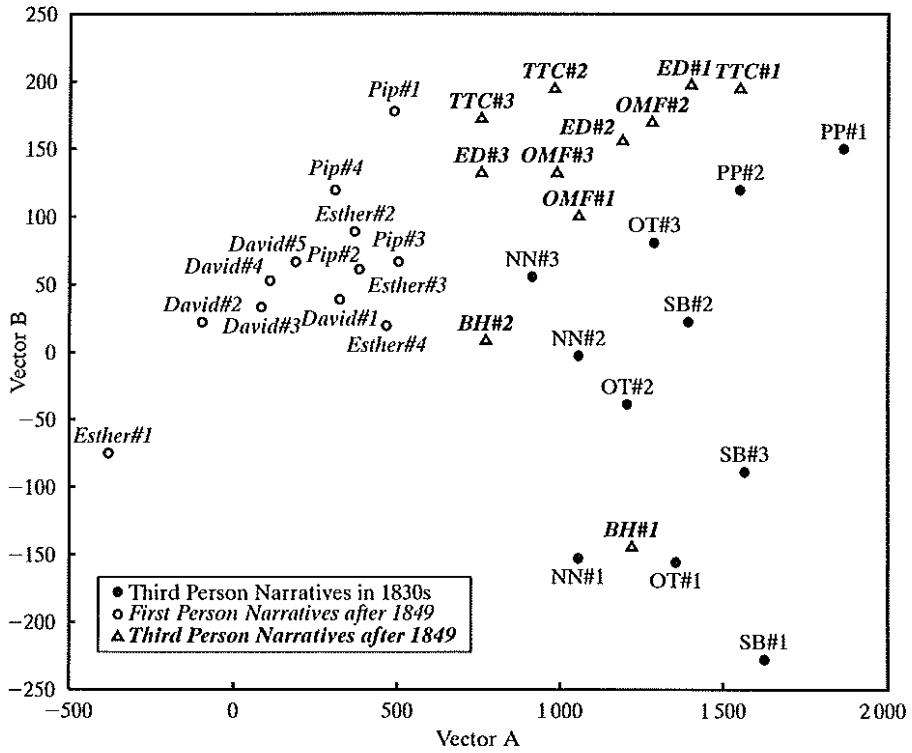


Fig. 2.– Eleven Narrators in Dickens's Novels: Text-plot
(based on the 74 very common word-types in the corpus)

Since this graph is a product of the previous one, they correspond to each other. When one compares the two graphs, one can see that easterly entries of words in Fig. 1 occur more frequently in text-entries lying towards the east

of Fig. 2 than in those lying towards the west (and vice versa), while northern entries of words in Fig. 1 predominate in the text-entries situated in the north of Fig. 2 and are outnumbered in texts that find their place in the south (and vice versa). Additionally, words that contribute little to the horizontal and vertical differentiations of texts lie around the middle of the graph.

The most obvious feature of Fig. 2 is the separation between the first-person and the third-person narratives along the horizontal axis. This axis represents the strongest component of the matrix, which accounts for 16.74 % of the complex interrelationships among the 74 words. That the first and third-person distinction predominates along the horizontal axis, with no interpenetration between the two groups, testifies that stylistic difference between the two narrative modes is still remarkably consistent even after the removal of the personal pronouns. Another feature of interest, which is more relevant to the present study, is that the overall contrast can be drawn diagonally between the entries for earlier novels and those for the later. With an imaginary border running from the north-east to the south-west, the segments of earlier texts find themselves in the lower right triangle, while those of the later sit in the upper left. Such a diagonal separation takes place when the two vectors function in conjunction with each other. A notable departure from this pattern is the first segment for the anonymous narrator of *Bleak House* (BH#1), which goes deep into the middle of the earlier narrators' cluster. The second segment lies just across the border.

Let us scrutinise Fig. 1 to see how reciprocities among the 74 words are related to the overall configuration of Fig. 2. The definite and the indefinite articles, *the, a*, and *an* all incline towards the east. The first segments of *The Pickwick Papers* (PP#1) and *Sketches by Boz* (SB#1) are among the most given to the use of these words. The easterly locations of the articles seem to reflect the comparative weighting on nominal phrases in the third-person and its comparative avoidance in the first-person narratives. The major prepositions, such as *of, by*, and *in*, also leaning towards the east end, add to this interpretation. The relatives *which* and *who*, markers of elaboration or identification, go to the far south-east, the territory where the earlier third-person narratives dominate. Among others, SB#1 has frequent recourse to these devices for elaboration.

On the other hand, one can notice the westerly entries of words that betoken intensification and comparison (the adverb of degree *so, never, very, much*, the prepositional *like*, and *than*). Such a group of words represent habits of emphasis shared by the first-person narrators, among whom the first segment of Esther's narrative is an extreme example. Habits of reference provide another revealing contrast. The first-person narrators appear to be more given

to reference by pronouns as is illustrated by westerly positions of the pronoun *it* and the demonstrative *that*, while the third-person narrators seem to favour more exact and explicit reference, which is implied by a set of determiners and the two relatives lying towards the eastern extremity of the chart.

The word-pattern along the horizontal axis points to contrast between the more formal, generalising style and the more subjective, emphatic style. It may not be inappropriate to associate the former with the overall tendency to describe things and events externally (from a so-called "omniscient" point of view) in the third-person narrative. The latter may illustrate one feature of the language of narrative in which a narrator's sense of value tends to be more freely reflected.

From a chronological viewpoint, the relatives *which* and *who*, lying at the south-eastern end of the chart, are words that discriminate strongly in favour of the texts written in 1830s. What is interesting here is the distance that separates another relative, *that*, from this set: while the early style is characterised by pronounced recourse to WH-relatives, the later Dickens is inclined in comparative terms to *that*-relatives. Such different distributions of *that*-relatives and WH-relatives are in keeping with Beaman's (1984) treatment of the two types of relatives as separate classes. Beaman furthermore finds *that*-relatives highly associated with spoken narratives and WH-relatives with written narratives. Major prepositions like *of* and *in*, which often form postmodifying phrases, are also found in the early Dickensian quarter, together with the set of articles.

Words situated at the north-western, or late Dickensian, end of the chart include the adverbial particles *out* and *down*, the pronoun *it*, the prepositional *like* (often used as in comparison), words often associated with oral style. Burrows (1992), in treating the chronological shift of English first-person narrative over the past three centuries, notes the increasing incidence of the adverbial particles and the pronoun *it* towards the present. The mildly increasing incidence of the connectives *and* and *then* is also recognised in the later works. Such an overall oral and less formal tendency in the later narrative stands in contrast with the comparatively more disquisitory and formal cast of the early narratives, which is reflected in the predominance of WH-relatives, the determiners, and some of major prepositions. The words denoting parts of body, such as *hand, eyes, head*, and *face*, also characterise the texts written after 1849.

On a closer study of Fig. 2, another interesting feature is brought into relief. The first segments tend to lie about the outskirts in the chart, whereas other segments are positioned in relative terms nearer to the centre. Nine of the eleven first segments lie towards the various extremities. Three of the four first segments from the 1830s' texts (SB#1, OT#1 and NN#1) are in the south-east,

typifying the early Dickens's preference for nominal and prepositional phrases (as is reflected in the comparative weighting on articles and prepositions) and tendency to elaboration. The anachronistically placed BH#1 (as noted earlier) accompanies the trio. Its counterpart in Esther's narrative (Esther#1) is at the westmost end, demonstrating predominance of words related to intensification and negatives[8]. The remarkable contrast between the pair reflects how Dickens, as Page (1990) points out as the result of a manuscript study, took pains to differentiate sharply the two narrators' voices (especially at the outset of each narrative) at different stylistic levels, to say nothing of tense and person. ED#1 and TTC#1, standing furthest from both the early and the first-person narratives, show the most sparing use of "old-fashioned" elaboration.

The seemingly aberrant position of BH#1 in Fig. 2 needs some explanation here. A close inspection of the word frequency table and a concordance for each segment of texts reveals that, whereas the anonymous narrator shares most of the late Dickensian characteristics, he finds more need to use *who* than any other late narrator at the outset of the novel where he mentions the people involved in the notorious law suits[9]. This narrator, on the other hand, makes comparatively limited reference to the bodily words *eyes, face, head*, and *hand*, to which the narrators of other late novels tend to refer more frequently. Such, rather thematic, factors seem to function both positively and negatively in placing BH#1 apart from the other later narrators. The idea will be given credence in the analysis in the next section.

## 2.2. Distribution test and the strongly discriminating words

Though the Fig. 2 shows a general contrast between the texts written in the 1830s and those written after 1849, the contrast is somewhat blurred by the fact that the principal vector is dominated by the difference between the first-person and third-person narratives. To diminish the effect of the first-person subset and gain a more rigorous distinction, it seems necessary to establish a basis of comparison excluding the first-person narratives.

Student's *t*-test is a tool often used for comparison between two groups. It gives a measure of judgement on whether discrepancy between two sets of

---

[8] On negation, WATT (1960) states that "there are no negatives in nature, but only in human consciousness." If we apply Watt's philosophical remarks in the present case, recourse to negation can be considered a measure of subjectivity. Watt, 274.

[9] *Conc* v. 1.71 beta (Summer Institute of Linguistics, Dallas TX, 1992) was used for building a computerised concordance to the texts I analysed.

Table 2

Dickens's Novels in 1830s versus Novels after 1849:
The 21 most common strongly discriminating words.

(Sorted in the ascending order of "probability" level.)

| No. | Rank | Word-types | Freqs. Means (%) 1830s | 1849– | *t* | Probability | DF |
|---|---|---|---|---|---|---|---|
| 1 | 15 | which(r) | 0.732 | 0.387 | 7.7 | 0.0000 | 20 |
| 2 | 9 | it | 0.696 | 1.201 | −6.94 | 0.0000 | 20 |
| 3 | 26 | out | 0.1761 | 0.3662 | −5.12 | 0.0001 | 20 |
| 4 | 65 | like(p) | 0.0601 | 0.1465 | −4.31 | 0.0003 | 20 |
| 5 | 48 | again | 0.1173 | 0.1966 | −3.21 | 0.0044 | 20 |
| 6 | 32 | very | 0.341 | 0.1935 | 3.12 | 0.0057 | 19 |
| 7 | 39 | down | 0.1346 | 0.2466 | −3.03 | 0.0066 | 20 |
| 8 | 57 | gentleman | 0.279 | 0.0726 | 2.97 | 0.011 | 13 |
| 9 | 41 | more | 0.2054 | 0.1561 | 2.66 | 0.018 | 15 |
| 10 | 63 | great | 0.194 | 0.1023 | 2.59 | 0.021 | 15 |
| 11 | 5 | in | 1.893 | 2.166 | −2.4 | 0.026 | 20 |
| 12 | 54 | than | 0.1230 | 0.1908 | −2.36 | 0.030 | 18 |
| 13 | 51 | time | 0.1717 | 0.1162 | 2.35 | 0.030 | 19 |
| 14 | 62 | face | 0.1130 | 0.199 | −2.36 | 0.032 | 15 |
| 15 | 36 | so(adv.d.) | 0.1614 | 0.2185 | −1.99 | #0.060 | 20 |
| 16 | 46 | other | 0.2254 | 0.1648 | 1.98 | #0.062 | 20 |
| 17 | 49 | that(r) | 0.128 | 0.224 | −1.98 | #0.062 | 20 |
| 18 | 49 | its | 0.1482 | 0.286 | −2.03 | #0.065 | 12 |
| 19 | 47 | over | 0.1547 | 0.2133 | −1.87 | #0.077 | 20 |
| 20 | 64 | hand | 0.0944 | 0.187 | −1.86 | #0.087 | 12 |
| 21 | 16 | for(p) | 0.559 | 0.430 | 1.74 | #0.098 | 18 |

Words with # are those with higher probability score than "significant" level ($p > 0.05$). "Rank" indicates the original word frequency hierarchy. "DF" stands for degree of freedom. (adv.d.) = adverb of degree, (p) = preposition, (r) = relative. For probability level, the score shown in MINITAB as "0.0000" is too small to be round up to 0.0001.

data is significant enough to be pursued, by calculating the divergences from means. With the help of *t*-test, I compared the twelve segments of the early third-person narratives with the eleven segments of the later. The summary of the result is shown in Table 2.

The *t*-score is given in the fifth column. The positive value indicates that the word discriminates in favour of the early subset. The negative value means the converse. The sixth column indicates a probability that chance has caused discrepancy in word-frequency between the two sets. Though a probability of 0.05, one in twenty, is often used as a threshold at which the result is considered statistically "significant," I included in the list the words with probability level more than 0.05 and less than 0.1. A score of 0.01, which is

regarded as "highly significant," means that difference between the two groups is 99 % reliable. The corollary, however, also must not be forgotten: provided the calculation is accurate, such a difference *does* happen by chance once in a hundred trials. In such a case the result is dismissed as not significant. However, when such a (or smaller) probability score is gained several times in a hundred trials, it becomes possible to take the result as significant. In the present case, in which seventy-four words are subjected to the test, seven words show a score of less than 0.01, and another seven one of less than 0.05. The discrepancy between the texts in 1830s and those after 1849 in the incidence of *which, it, out*, and *like*, which I noted earlier, is demonstrated again as "very highly significant" ($p < 0.001$).

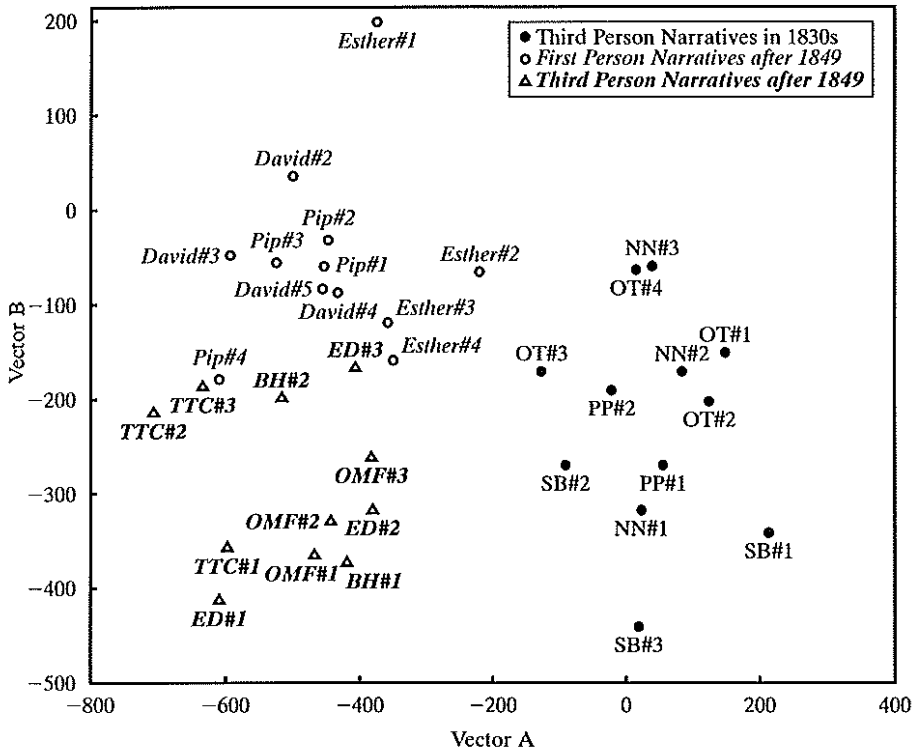


Fig. 3.– Early Style versus Late Style: Text-plot
(based on the 21 most common "strongly discriminating" words of the main corpus)

Figs. 3 and 4 are derived from principle component analysis of the twenty-one commonest "strongly discriminating" words listed in Table 2. The salient

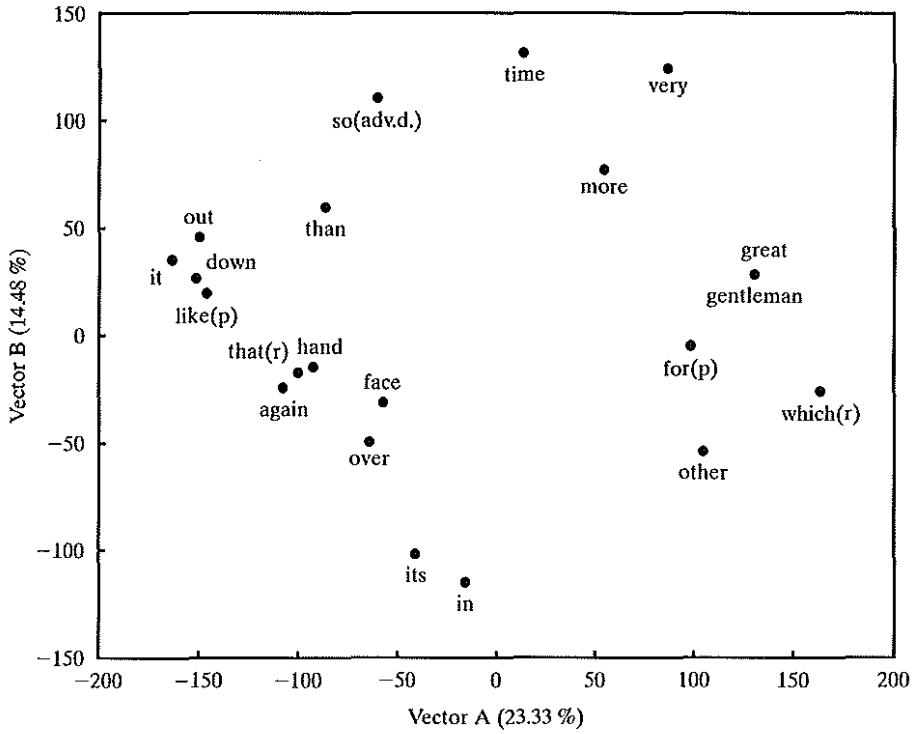


Fig. 4.– Early Style versus Late Style: Word-plot
(based on the 21 most common "strongly discriminating" words of the main corpus)

feature of Fig. 3 is sharper differentiation between the texts written in 1830s and those written after 1849, now displayed along the most powerful vector. There is no interpenetration between the two subsets, with BH#1 and BH#2 sitting within the large cluster of later texts. Esther's (#2–4) positions in the "early Dickens" side of the "late Dickens" cluster might seem inappropriate at first thought, but this phenomenon can be accounted for by her marked preference for intensification. She stands highest among the later narrators in the incidence of *very, more,* and *great.* In this respect, she is sharply differentiated from other first-person narrators, David and Pip.

Another noteworthy feature is a contrast between the later first-person and the later third-person narratives along the second strongest vector. It is remarkable that the set of twenty-one "chronologically-oriented" words still has the power to separate the two narrative modes, though less sharply than in Fig. 2. Taking into account the fact that I excluded the first-person narratives

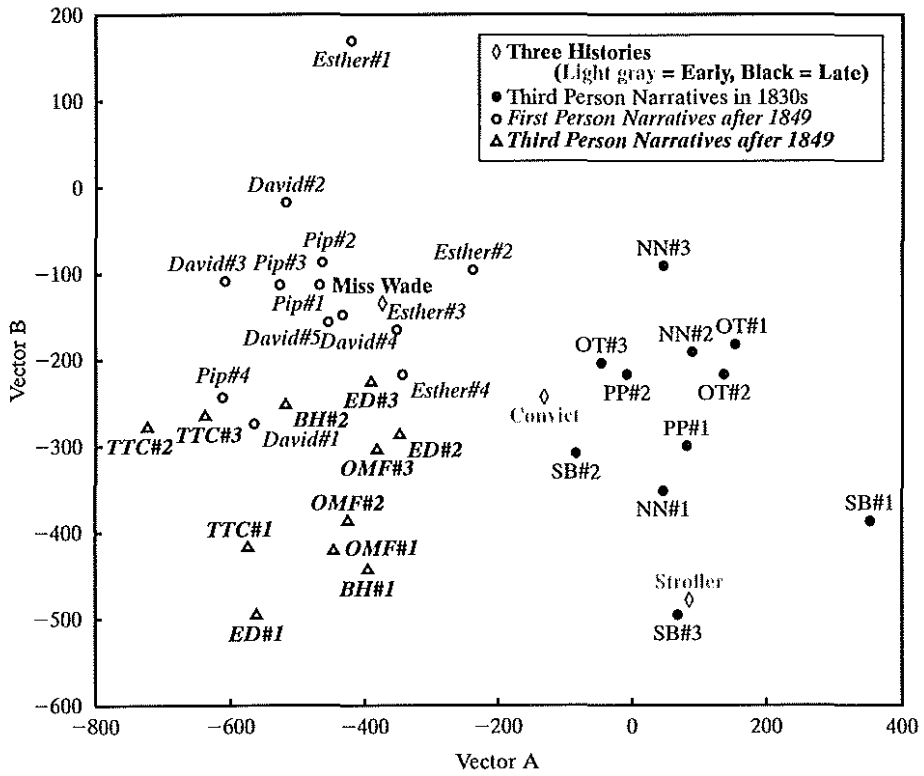


Fig. 5.– Early Narratives, Late Narratives and Three Histories: Text-plot
(based on the 21 most common "strongly discriminating" words of the main corpus)

when I extracted the set of differentiae, it would not be unreasonable to regard the clusterings as formed by a chronological determinant.

The force of the distinction based on the twenty-one words (or to say sceptically the possibility that this result came out of mere freak of chance) can be tested by introducing a new set of texts. Here three shorter first-person narratives will be employed as a control set. Fig. 5 summarises the result in a convincing way, by placing the new set in appropriate clusters. Two histories, "The Stroller's Tale" and "The Convict's Return" (both from *The Pickwick Papers*) lie in the 1830s' cluster, while Miss Wade's narrative (from *Little Dorrit*) lies in the "after-1849" cluster. Interestingly, while Miss Wade finds herself in the "first-person domain", neither the stroller nor the convict is distinguished from the third-person narrators. The essential configuration of Fig. 5 is not dissimilar to Fig. 3. Since the three new texts did not participate in the compilation of

the original word-list or in the selection of the discriminating words, I take this evidence as confirming that Dickens's early and later styles can be distinguished from each other in quantitative terms.

## 3. Final Remarks

The evidence this paper has shown is not only consistently in keeping with my earlier findings from a smaller set of texts, but also capable of distinguishing Dickens's early and later styles with a sharpness that could not be achieved before. Whether to regard such a change in one's style as a development or not is a matter of aesthetic judgements, and the approach I pursued in this study does not lead to any such value judgements. But the evidence from very common vocabulary of Dickens's narrative at least betokens a movement from a style that characterises *Sketches by Boz* and *The Pickwick Papers* to a style typically demonstrated in *David Copperfield*, from a more formal and elaborate style to a style of an increasing oral tendency.

In this approach, I excluded personal pronouns and finite verbs due to the difference in narrative mode and tense. By limiting comparison to a particular narrative mode (either first-person or third-person), frequency distribution of personal pronouns can be examined in much the same way. For the treatment of finite verbs, lemmatisation is another option, by means of which it is possible to process a number of verbs statistically. My next step, then, is to complement this approach by focusing attention on what I have excluded from this study.

## Bibliography

### 1. Books

BIBER (Douglas): 1988, *Variation across Speech and Writing* (Cambridge: Cambridge UP).

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

CLUETT (Robert): 1976, *Prose Style and Critical Reading* (New York–London: Teacher's College Press, Columbia University, "New Humanistic Research Series, 3").

GORDON (Ian A.): 1966, *The Movement of English Prose* (London: Longman, "English Language Series, 2").

PAGE (Norman): 1990, *Bleak House: A Novel of Connections* (Boston: Twayne Publishers).

STEVENS (James): 1986, *Applied Multivariate Statistics for the Social Sciences* (Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers).

TOOLAN (Michael J.): 1988, *Narrative: A Critical Linguistic Introduction* (London–New York: Routledge, "The Interface Series, 1").

## 2. Papers

BEAMAN (Karen): 1984, "Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse" in TANNEN, pp. 45–80.

BIBER (Douglas) and FINEGAN (Edward): 1989, "Drift and the Evolution of English Style: A History of Three Genres", *Language*, Volume 65, 3, pp. 487–517.

BIBER (Douglas): 1993, "The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings", *Computers and the Humanities*, 26, pp. 331–345.

BRAINERD (Baron): 1979, "Pronouns and Genre in Shakespeare's Drama", *Computers and the Humanities*, 13, pp. 3–16.

BURROWS (J.F.): 1987b, "Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style", *Literary and Linguistic Computing*, 2, pp. 61–70.

BURROWS (J.F.): 1992, "Computers and the Study of Literature" in BUTLER, pp. 167–204.

BURROWS (J.F.): forthcoming 1994, "Tiptoeing into the Infinite: Testing for National Differences in the Language of English Narrative" in HOCKEY and IDE.

BURROWS (J.F.) and CRAIG (D. H.): forthcoming 1994, "Lyrical Drama and the "Turbid Mountebanks": Styles of Dialogue in Romantic and Renaissance Tragedy", *Computers and the Humanities*, 28.

BUTLER (Christopher): 1992, ed. *Computers and Written Texts* (Oxford: Blackwell).

CRAIG (D.H.): 1991, "Plural Pronouns in Roman Plays by Shakespeare and Johnson", *Literary and Linguistic Computing*, 6, pp. 180–186.

CRAIG (D.H.): 1992, "Authorial Styles and the Frequencies of Very Common Words: Jonson, Shakespeare and the Additions to *The Spanish Tragedy*", *Style*, Volume 26, 2, pp. 199–220.

HOCKEY (Susan) and IDE (Nancy): forthcoming 1994, *Research in Humanities Computing '92* (London: Oxford UP).

TABATA (Tomoji): 1993, "The Language of Dickens and Its Computer-Based Evidence: A Step towards a Chronological Study", *Kumamoto Studies in English Language and Literature*, 36, pp. 116–134.

TANNEN (Deborah): 1984, ed. *Coherence in Spoken and Written Discourse* (Norwood, N.J.: Ablex).

WATT (Ian): 1960, "The first Paragraph of *The Ambassadors*: an explication", *Essays in Criticism*, 10, pp. 250–74.