Narrative Style and the Frequencies of Very Common Words: A Corpus-Based Approach to Dickens's First Person and Third Person Narratives*

Tomoji Tabata

Abstract

The present article is devoted to statistical analysis of the language of Dickens's novels. The particular problem is to examine structural and stylistic features of the first person and third person narratives. In the following analysis, I apply Principal Component Analysis (PCA) to the examination of the frequency-patterns of very common word-types of the text-samples. What emerges from this approach is a remarkable contrast between the two narrative modes. The differentiae between Dickens's first person and third person narratives suggest a broad opposition between a more oral, subjective, verbal style and a more literate, descriptive, nominal style.

0. Introduction

The choice of a particular point of view is arguably one of the most crucial decisions in beginning a fictional discourse. Decisions may include whether to employ a first person narrator or a third person narrator, whether to narrate in the present tense or the past tense, and so forth. While a first person narrative and a third person narrative differ obviously from each other in terms of the presence, or the total absence, of the first person pronouns, *I*, *me*, and *my*, the two narrative modes are likely to differ in less obvious ways. They share, no doubt, many characteristics of the language of narrative distinct from other genres of English prose.

This study examines, from a quantitative viewpoint, linguistic and stylistic attributes of the two narrative modes to demonstrate how Dickens differentiates one mode of narrative from another. The approach I adopt in this study has three characteristics. First of all, it is corpus-based: the word frequency profiles that will appear a little later and other frequency counts are derived from a computerised text-corpus. Second, it focuses on very common word-types, most of

English Corpus Studies, No. 2, 1995, pp. 91-109.

which are function words, rather than rare words or so-called “key-words” that are usually focused on in studies of literary texts. Third, it is based on multivariate statistics to illustrate relationships among very common word-types; relationships among text-samples; relationships between the very common word-types and the text-samples. This method has produced justifiable results in studies of disputed authorship (Burrows: 1989, Craig: 1992); literary idiolects (Burrows: 1987a, Tabata: 1991); stylistic changes that occur over an author’s career (Tabata 1993 & 1994 forthcoming); and in other areas as well.

Table 1. The Set of Eleven Narratives

Label	**Narrator [*TEXT*] & Date**		**Word-tokens** [Pure-Narrative]	**Segments**
	First Person Narratives			
David#1-5	David [*David Copperfield*]	(1849-50)	20145	5
Esther#1-4	Esther [*Bleak House*]	(1852-3)	18399	4
Pip#1-4	Pip [*Great Expectations*]	(1860-1)	18359	4
	Group Total		56903	13
	Third Person Narratives			
SB#1-3	Sketches by Boz	(1836)	12569	3
PP#1-3	The Pickwick Papers	(1836-7)	11081	3
OT#1-4	Oliver Twist	(1837-8)	16677	4
NN#1-3	Nicholas Nickleby	(1838-9)	12863	3
BH#1-2	Bleak House	(1852-3)	7389	2
TTC#1-3	A Tale of Two Cities	(1859)	12798	3
OMF#1-3	Our Mutual Friend	(1864-5)	13117	3
ED#1-3	The Mystery of Edwin Drood	(1870)	11973	3
	Group Total		98467	24

1. Data

1.1 *Corpus*

The corpus draws on ten novels from Dickens’s *oeuvre* (See Table 1).[1] 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 of comparison.[2] The current corpus consists of three first person narratives—*David Copperfield*, Esther’s narrative, and *Great Expectations*—and eight third person narratives. *Bleak House* provides two narratives: one is the first person narrative by the character narrator Esther Summerson, the other is the anonymous third person narrative. The two narratives of *Bleak House* are also contrasted in the use of verb tenses. While Esther uses the past tense, the anonymous narrator employs the “dramatic present.” The dramatic present is also used in *The Mystery of Edwin Drood* and in some parts of *Sketches by Boz* and *David Copperfield*.

Each text is then divided into successive 4000-word segments. Segmentation of text has two objectives. First, to give each variable (i.e., word) as appropriate a number of samples as possible in order to reduce the possibility of chance effect. Second, to help observe internal variation (or consistency) in each text. In all, the present study analyses 37 segments (or text-samples), of which 13 are first person narratives and 24 are third person narratives.[3]

1.2 *Some preliminary treatments of data*

In the present case, the discrepancy between the first person and the third person narratives in an incidence of first person pronouns is too obvious to require a statistical analysis. It is desirable, therefore, to exclude those pronouns from the following statistical analysis so as to diminish the overshadowing effect of what is already evident. Otherwise the difference due to the incidence of first person pronouns will become so inflated through statistical treatments that other subtler differences may be submerged. This exclusion of first person pronouns deprives my data of some interesting subjects for computational stylistics, but in return it makes them sensitive to evidence of subtler stylistic differences.

Another problem is concerned with verb forms. My earlier studies have shown that the top 100 words include only a small number of verbs—mostly preterite forms of common verbs, such as *was*, *had*, and *said*.[4] The size of the present corpus, in addition, is not large enough to process verbs of lower frequency. If words of low frequency are subjected to a statistical analysis, the dearth of numbers may cause an aberrant result. The recognised solution is lemmatisation. For example, *take*, *takes*, *took*, *taken*, and *taking* are lemmatised as *take*. Lemmatisation enables a number of verbs to rank higher than in my

Table 2. Eleven Narrators in Dickens's Novels:
Standardised (text-percentage) frequencies for the 100 most common word-types in the "pure-narrative."

Rank	Word-types	SB	PP	OT	NN	David	Esther	BH	TTC	Pip	OMF	ED	Total (raw)	(%)
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	be	3.477	2.969	3.352	2.946	3.783	3.565	3.816	3.110	3.470	2.851	2.923	5163	3.323
4	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
5	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
6	in(p)	2.164	1.660	1.847	1.788	1.812	1.853	2.463	2.016	1.672	2.173	2.096	2983	1.920
7	his	1.090	2.265	2.027	1.998	0.521	1.005	0.947	2.188	1.117	2.295	2.038	2373	1.527
8	have	1.567	1.516	1.619	1.174	1.762	1.631	1.719	1.469	1.759	1.243	0.969	2358	1.518
9	to(i)	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
10	he	1.034	1.354	1.961	1.454	0.789	1.223	1.177	1.453	1.073	1.479	1.178	1983	1.276
11	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
12	to(p)	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
13	say	0.151	1.724	1.133	1.508	1.142	1.614	0.650	0.445	0.980	0.991	0.793	1630	1.049
14	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
15	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
16	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
17	that(c)	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
18	on(p)	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
19	by(p)	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
20	her(a)	0.271	0.108	0.168	0.288	0.988	0.598	0.839	0.656	0.376	0.793	0.685	821	0.528
21	which(r)	0.812	0.641	0.762	0.669	0.417	0.424	0.420	0.398	0.381	0.435	0.309	794	0.511
22	him	0.342	0.298	0.899	0.474	0.392	0.478	0.338	0.641	0.507	0.496	0.443	772	0.497
23	for(p)	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
24	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
25	she	0.127	0.009	0.126	0.155	0.963	0.902	0.568	0.445	0.616	0.267	0.292	700	0.451
26	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
27	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
28	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
29	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
30	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
31	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
32	they	0.509	0.325	0.444	0.389	0.268	0.217	0.284	0.539	0.207	0.175	0.409	518	0.333
33	look	0.127	0.244	0.216	0.334	0.432	0.321	0.298	0.453	0.468	0.358	0.292	516	0.332

*(a) = adjective, (adv) = adverbials, (a.d.) = adverb of degree, (c) = conjunction, (d) = demonstrative, (i) = infinitive, (r) = relative, (p) = preposition, (pron) = pronoun

Table 2. (continued)

Rank	Word-types	SB	PP	OT	NN	David	Esther	BH	TTC	Pip	OMF	ED	Total (raw)	(%)
34	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
35	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
36	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
37	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
38	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
38	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
40	that(d)	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
41	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
42	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
43	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
44	up(adv)	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
45	go	0.080	0.099	0.096	0.210	0.442	0.408	0.217	0.219	0.370	0.320	0.217	408	0.263
46	so(a.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
47	do	0.095	0.162	0.198	0.233	0.501	0.261	0.176	0.242	0.289	0.168	0.184	383	0.247
48	upon(p)	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
49	take	0.183	0.208	0.246	0.264	0.253	0.174	0.135	0.250	0.338	0.252	0.284	375	0.241
50	their	0.549	0.316	0.216	0.381	0.104	0.082	0.338	0.273	0.180	0.191	0.242	372	0.239
51	make	0.088	0.171	0.228	0.194	0.357	0.250	0.284	0.219	0.332	0.183	0.192	368	0.237
52	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
53	come	0.111	0.072	0.144	0.117	0.377	0.315	0.325	0.234	0.289	0.236	0.184	355	0.228
54	them	0.278	0.099	0.138	0.288	0.194	0.207	0.244	0.367	0.267	0.130	0.242	343	0.221
55	would	0.334	0.135	0.330	0.179	0.199	0.234	0.203	0.164	0.245	0.099	0.109	325	0.209
56	see	0.183	0.027	0.180	0.124	0.278	0.288	0.189	0.148	0.430	0.061	0.150	319	0.205
57	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
58	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
59	could	0.127	0.126	0.168	0.171	0.298	0.326	0.054	0.164	0.256	0.137	0.042	295	0.190
60	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
61	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
62	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
63	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
64	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
65	her(pron)	0.095	0.018	0.012	0.054	0.412	0.402	0.257	0.195	0.114	0.076	0.159	274	0.176
66	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
67	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
68	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

*(a) = adjective, (adv) = adverbials, (a.d.) = adverb of degree, (c) = conjunction, (d) = demonstrative, (i) = infinitive, (r) = relative, (p) = prepos
(pron) = pronoun

Table 2. (continued)

Rank	Word-types	SB	PP	OT	NN	David	Esther	BH	TTC	Pip	OMF	ED	Total (raw)	(%)
69	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
69	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
71	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
72	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
73	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
74	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
74	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
76	himself	0.127	0.208	0.198	0.233	0.050	0.076	0.162	0.102	0.109	0.274	0.217	233	0.150
77	gentleman	0.064	0.334	0.438	0.272	0.069	0.163	0.068	0.078	0.000	0.122	0.033	232	0.149
78	know	0.159	0.063	0.090	0.086	0.233	0.239	0.284	0.039	0.174	0.084	0.109	226	0.145
78	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
80	reply	0.024	0.433	0.378	0.459	0.050	0.027	0.041	0.031	0.027	0.076	0.117	224	0.144
81	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
81	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
83	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
84	face	0.072	0.081	0.150	0.132	0.134	0.136	0.068	0.211	0.076	0.183	0.284	216	0.139
85	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
86	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
87	like(p)	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
88	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
88	turn	0.072	0.144	0.114	0.148	0.134	0.125	0.068	0.133	0.158	0.168	0.134	202	0.130
90	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
91	get	0.080	0.036	0.102	0.086	0.139	0.120	0.108	0.172	0.245	0.099	0.159	199	0.128
92	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
93	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
93	seem	0.072	0.027	0.066	0.078	0.154	0.223	0.095	0.063	0.185	0.183	0.084	188	0.121
95	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
95	sit	0.024	0.081	0.096	0.093	0.204	0.152	0.135	0.141	0.109	0.122	0.109	186	0.120
97	think	0.056	0.063	0.072	0.078	0.243	0.196	0.027	0.031	0.240	0.030	0.067	183	0.118
97	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
97	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
100	never	0.143	0.000	0.066	0.054	0.179	0.228	0.217	0.023	0.136	0.061	0.125	181	0.116
Sum		**52.28**	**52.10**	**54.70**	**53.43**	**54.71**	**53.70**	**53.59**	**56.24**	**54.00**	**53.34**	**52.66**	**83613**	**53.82**

*(a) = adjective, (adv) = adverbials, (a.d.) = adverb of degree, (c) = conjunction, (d) = demonstrative, (i) = infinitive, (r) = relative, (p) = prepos
(pron) = pronoun

earlier studies and brings the number to a comparatively safe (though not all-sufficient) working-level.

On the other hand, common homographic forms are tagged to specify each usage: for example, “that” is separated into “that(c)” (conjunctive), “that(r)” (relative), “that(d)” (demonstrative); “her” is tagged to separate possessive adjective, “her(a)”, from pronouns, “her(pron)”; verbs like *look*, *reply*, etc. are also tagged to distinguish from nouns. For the purpose of counting, furthermore, 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”).[5]

1.3 *The 100 most common word-types*

Table 2 lists the 100 most common word-types whose occurrence in segments of text is frequent enough to allow multivariate analysis.[6] While most of the words with higher frequency are function words, several common adjectives and nouns find themselves towards the bottom of the list. What deserves attention is that lemmatisation has enabled seventeen verbs to rank within the top 100 words: *be* [3], *have* [8], *say* [13], *look* [33], *go* [45], *do* [47], *take* [49], *make* [51], *come* [53], *see* [56], *know* [78], *reply* [80], *turn* [88], *get* [91], *seem* [93], *sit* [95], and *think* [97] (numbers in square brackets show respective frequency rankings).[7] Since the 100 hundred words account for 53.8 % of all the word-tokens in the pure-narrative, it may not be inappropriate to assume that some major determinants of style are reflected in the data.

2. Analysis and Interpretation of the results

2.1 *Principal Component Analysis of the 100 words*

Figure 1 shows the results of a principal component analysis (PCA) of the hundred words in the texts divided into thirty-seven segments. The first step of PCA is to measure the correlations of each of the hundred words with each of the other ninety-nine across the thirty-seven segments of text, using Pearson’s Product Moment Formula.[8] This procedure generates a matrix of 4450 correlation

Table 3. Pearson's Product-Moment Correlation Matrix

		the	and	be	of	a	in(p)	his	have	to(i)	he	. . .
1	**the**	—										
2	**and**	-0.076	—									
3	**be**	-0.306	0.002	—								
4	**of**	0.702	-0.225	-0.279	—							
5	**a**	0.035	-0.357	-0.264	0.309	—						
6	**in(p)**	0.180	-0.352	0.201	0.552	0.324	—					
7	**his**	0.549	-0.127	-0.640	0.357	0.210	-0.003	—				
8	**have**	-0.235	-0.028	0.376	-0.281	-0.190	-0.124	-0.474	—			
9	**to(i)**	-0.646	-0.073	0.309	-0.575	-0.159	-0.360	-0.441	0.284	—		
10	**he**	0.180	-0.316	-0.274	0.032	0.225	-0.160	0.643	-0.104	-0.081	—	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
45	**go**	-0.667	0.383	0.166	-0.717	-0.212	-0.275	-0.475	0.140	0.335	-0.264	. . .
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
100	**never**	-0.517	-0.066	0.539	-0.298	-0.280	0.003	-0.613	0.335	0.454	-0.346	. . .

coefficients. Table 3 gives a part of the matrix. Correlation-coefficients range in value from +1.000, a perfect positive correlation, to -1.000, a perfect negative correlation. A coefficient of 0.000 indicates no correlation whatsoever. The matrix of coefficients reflects similarity or contrast among the hundred words in their "behaviour," or concomitant frequency variation over the thirty-seven text segments. A coefficient of 0.702 between *the* and *of*, a comparatively high score allowing for the number of text-samples examined, indicates that where the relative frequency of *the* runs high, that of *of* tends to be concomitantly high. A score of -0.717 obtained between *of* and *go* provides a contrary example: a text that has pronounced recourse to *of* tends to show comparatively sparing use of *go*.

In a small matrix like Table 3, it may be easy to see certain patterns like those I sketched above. However the entire matrix, in which 4450 coefficients are given, makes it practically impossible to grasp complex interrelationships among the hundred words. The next step therefore is to subject the correlation matrix to eigen-analysis. By eigen-analysis, the complex patterns of a correlation coefficient are reduced to a succession of eigen-vectors, and an appropriate weighting, "eigen-value," is assigned to each eigen-vector. The first principal vector, or the first principal component, arrays the coefficients in a sequence

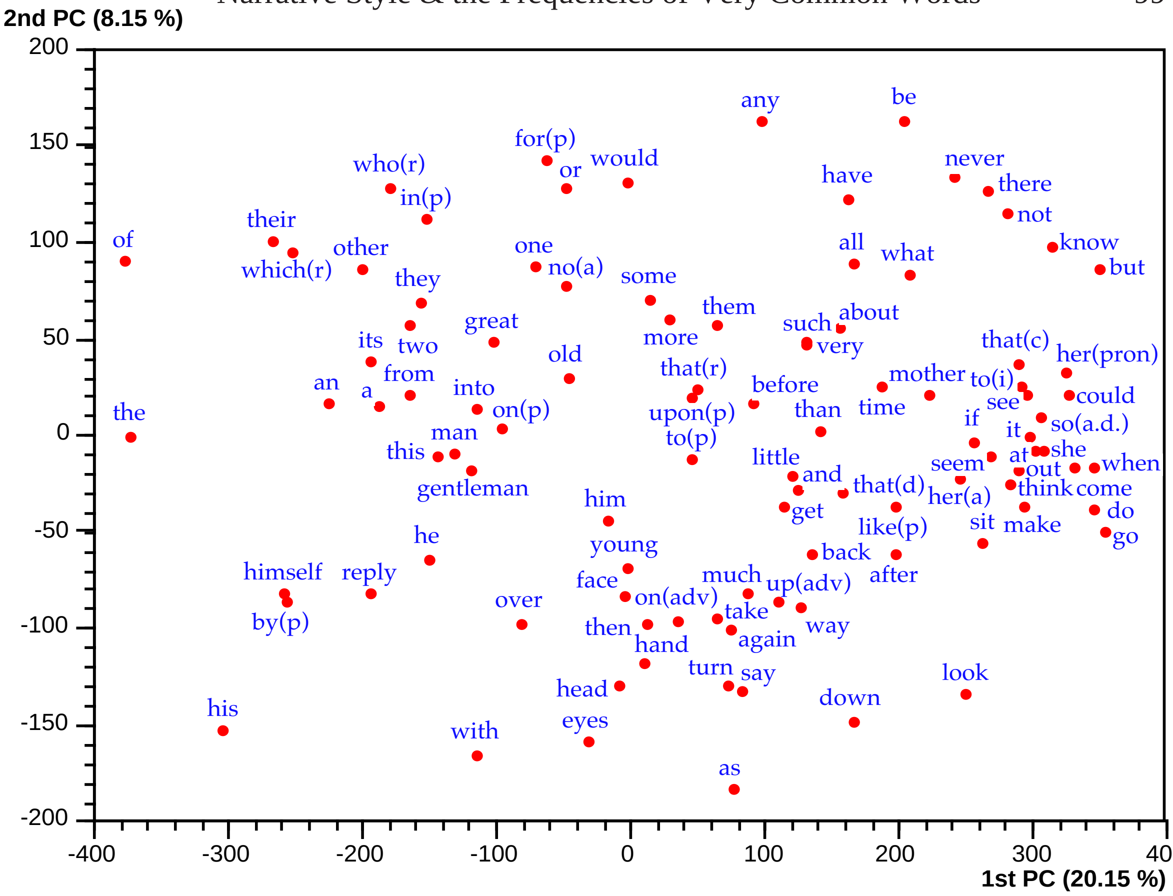

Figure 1. First person narratives versus Third person narratives: Word-plot (for the 100 most common words of the narrative corpus).

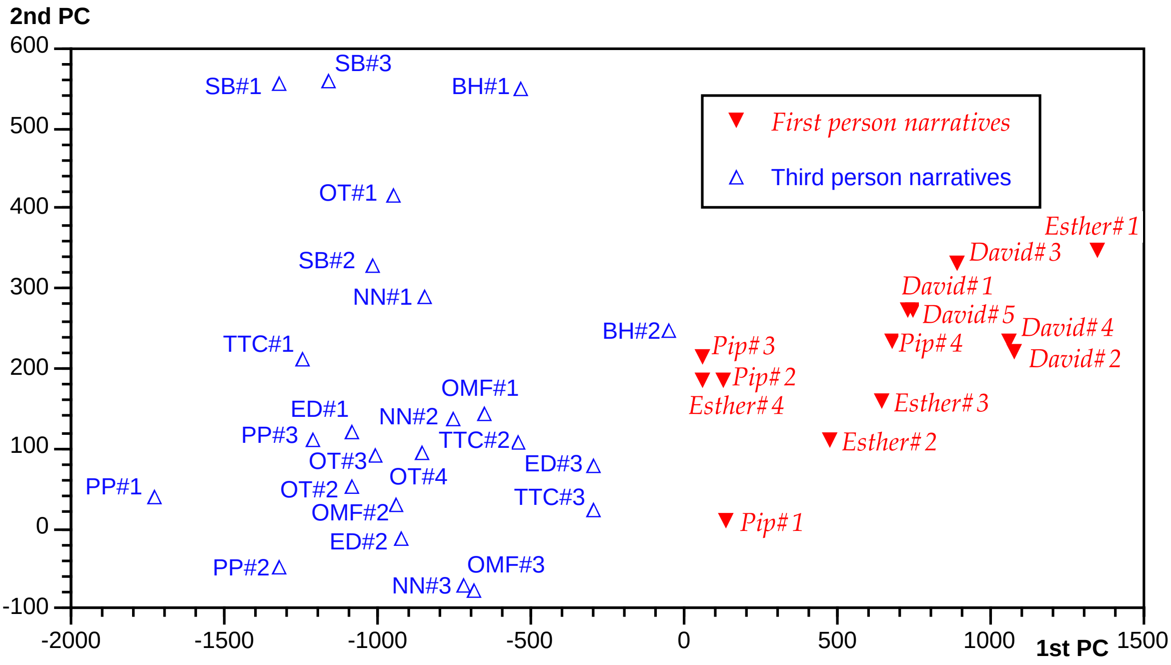

Figure 2. First person narratives versus Third person narratives: Texts in 4000-word segments (based on the 100 most common words of the narrative corpus).

which stands for the most consistent pattern of coefficients. The second principal vector accounts for the most consistent residual pattern of coefficients, and so on. To put it in another way, each principal component (PC) shows an important and independent dimension in the data.

It is also possible to project the most powerful components in a scatter diagram like Figure 1. The percentage-values given beside “1st PC” and “2nd PC” in Figure 1 indicate the extent to which these first two principal components account for the complex interrelationships of the correlation coefficients. The first PC accounts for 20.15 % of the original matrix, and the second PC 8.15 %. Figure 1 shows a picture of the reciprocal relationships among the hundred words. Relative distance between the entries reflects similarity or contrast among these words in their comparative frequency patterns over the text segments. Words located towards the east and those located towards the west of the figure 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, the concomitant pair of *the* and *of*, which I noted above, are at the western extremity, while *go*, a negative correlate with *the* and *of*, goes to the opposite extreme.

The merit of Figure 1 may be best explained in conjunction with Figure 2, which gives a picture of relationships among the text-samples. To arrive at this figure, the eigen-matrix that produced Figure 1 has been multiplied back through the original frequency table for the 37 text-segments. Through this procedure, the text samples are distributed on the word-pattern, and it becomes possible to discern relationships between the hundred words and the text samples. Easterly entries of words in Figure 1 occur more frequently in text-samples lying towards the east of Figure 2 than in those lying towards the west (and vice versa), while northern entries of words in Figure 1 predominate in the text-samples situated in the north of Figure 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 figure.

2.2 *Interpretation of the results*

Since the two figures correspond to each other, the configuration of one figure can be used to interpret the other. The most salient feature of Figure 2 is the overall differentiation between the first person narratives and the third person narratives along the horizontal axis. This result testifies that a difference between the two sets of narrative is the most powerful differentia in the data, even without the first person pronouns. The entries for first person narrative have positive scores for the first PC, while the entries for third person narrative have negative scores. It follows that words leaning towards the eastern extremity of Figure 1, or words with a higher *positive* score for the first PC, run higher in the first person narratives and are comparatively avoided in the third person narratives. On the contrary, words located towards the western extremity of Figure 1, or words that have a strongly *negative* score, are more conspicuous in third person narratives and run lower in the first person narratives.

Let us examine Figure 1 to see how reciprocities among the 100 words are associated with the separation of the first person narratives from the third person narratives. To begin with, words that characterise the first person narratives include all verbs but *reply*. Verbs like *go, do, come, make, think, sit, seem,* and *see,* especially, cluster at the eastern edge of Figure 1, indicating that they are words that discriminate strongly in favour of first person narratives. The conjunctions, *when, that(c), but,* and *and,* also run higher in the first person texts. My earlier studies (Tabata 1993 and 1994) have shown that these conjunctions are prominent in a narrative of a comparatively colloquial kind. Words that mark another “oral” tendency in the first person narratives are those which betoken intensification and comparison: *so, very, much, never, than,* and *like* are all in the “first person” side of the spectrum. Among the first person narrators, the first segment of Esther’s narrative (*Esther#1*) is an extreme example. The following passage from Esther’s narrative typically illustrates this strain:

> [My godmother] was *so very* good herself, I thought, that the badness of other people made her frown all her life. I felt *so* different from her, even making every allowance for the differences between a child and a woman; I felt *so* poor, *so* trifling, and *so* far off; that I *never* could be unrestrained with her—no, could *never* even love her as I wished. (*Bleak House*, p. 63: my italics)

Interestingly, the north-easterly entries for the negatives *not* and *never* seem to reflect the comparative weighting on negation in the first person set. Watt (1960: 275), taking note of the use of negation in Henry James's *The Ambassadors*, states that "there are no negatives in nature, but only in human consciousness." Applying Watt's observations in the present case, prominent recourse to negation can be considered a measure of subjectivity.

To turn to Figure 2, the locations of the two first segments (*Esther#1* and BH#1) of *Bleak House* are also revealing. The entry for *Esther#1* at the eastern end of the spectrum is in marked contrast with that for BH#1 at the top centre. The remarkable distance between the pair betokens 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.

On the other hand, markers of third person narratives include the definite and indefinite articles, *the*, *a*, and *an*. The definite article, *the*, in particular, lying in the western extremity, finds itself one of the most "strongly discriminating" words. The westerly locations of the articles reflect the relative preference for nominal phrases in the third person narratives and its comparative avoidance in the first person narratives. The idea is supported by the westerly entries for possessive pronouns and adjectives, such as *his*, *their*, *its*, *other*, *great*, *no*(a), and so on. The entries for major prepositions in the same direction add to this interpretation (*of*, *by*, *in*, *from*, *into*, *with*, *on*, and *for*). The preference for nominal structures stands in sharp contrast to the verbal tendency in the first person text-samples. The relatives, *which* and *who*, markers of elaboration (or syntactic complexity), are observed in the top left of the chart. Figure 2 demonstrates the two segments from *Sketches by Boz* (SB#1 and SB#3) as the most given to this set of words. In contrast to *which* and *who*, often used to introduce embedded structures, another relative *that* stands in the "first person" region, though not very far from the middle. 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 associate with spoken narratives and WH-relatives with written narratives.

Personal pronouns show interesting patterns in their dispersion along the

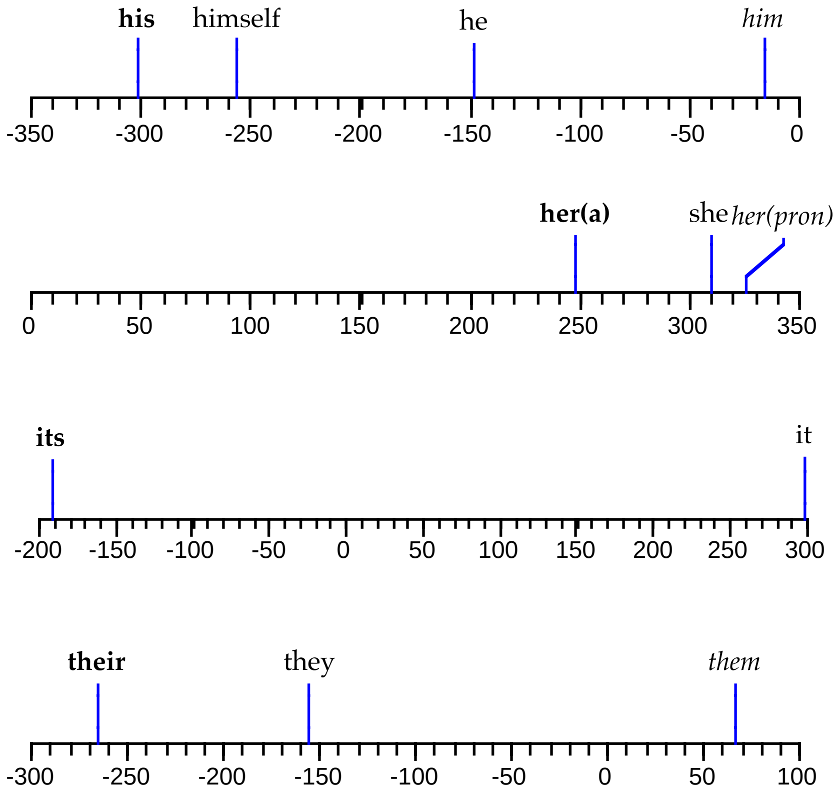

Figure 3. Distribution of personal pronouns in each category (based on the respective scores for the 1st PC).

horizontal axis. The third person singular feminine pronouns, *she*, *her*(a), and *her*(pron), are displayed in the "first person" domain. The third person singular masculine pronouns, *he*, *his*, and *himself*, by contrast, find themselves in the "third person" side of Figure 1, with *him* near the middle. There is something unexpected about the behaviours of the third person singular neuter pronouns and the third person plural pronouns. Whereas *it* is in the cluster of words that predominate in the first person narratives (*so*(a.d.), *she*, *at*, *out*, *think*, *when*, and so on), *its* is in the opposite side of the figure, being a close neighbour to the indefinite articles, *a* and *an*. *Them* also behaves differently from the other two of its grammatical trio, *they* and *their*: while *them* is in the positive range of the horizontal axis (around 70), *they* and *their* are close to *he* and *him* in their respective behaviours. On a closer inspection, a further interesting pattern can be observed in the locations of personal pronouns: in each category, possessive adjective forms lie in the westernmost, and objectives are the easternmost. A clearer illustration of this pattern is given in Figure 3.

To look at Figure 2 along the vertical axis is to see that compared with the wide range of dispersion of the third person narratives, the first person narratives are more concentrated around the middle. The vertical axis of Figure 2, not differentiating the text samples according to the narrative modes, does not allow a straightforward interpretation at a glance. A close examination of Figure 1, however, leads to a realisation that the vertical axis of Figure 2 distributes the more stative towards the north and the more dynamic towards the south. The north-easterly positions of stative verbs *be, have*, and *know* stand opposite to the southerly locations of verbs that describe “action,” such as *turn, say, look, take*, etc. It is not difficult to find illuminating evidence for this: the adverbials, especially those expressing “direction,” *down, up, on*(adv), and *over* are shown towards the bottom, together with the words denoting parts of body, such as *eyes, head, hand*, and *face*, situated around the bottom centre. In fact, the segments situated towards the bottom—mostly second and third segments—are richer in “stage-directions” that accompany dialogues, such as “..., *said* Mr Bumble, waving his right *hand* ...,” and in depictions that bear the action forward. On the other hand, those situated at the top of Figure 2—the first and the third segments of *Sketches by Boz* and the opening segment of *Bleak House*—are made up of descriptions comparatively static and elaborately picturesque. The segments of *Sketches by Boz* are, just as the title tells, “sketches” of London. The opening segment of *Bleak House* describes the High Court of Chancery shrouded in the fog. In the first segment of *Oliver Twist*, the narrator tells of the circumstances attending the birth of Oliver Twist and elaborates on the notoriety of the workhouse with mock seriousness. Such a static and picturesque strain, which can be found, more or less, in other first segments, seems to account for their relatively high scores for the second PC. Notable departures from this pattern are the opening segments of the *Pickwick Papers*, which reports the meeting of the Pickwick Club, and *Great Expectations*, which begins with Pip’s encounter with Magwitch and in which dialogues are often brought into the foreground.

3. Conclusion

The results of the present analysis revealed that the difference between the first person and the third person narratives is statistically a most powerful differentia in the data. The evidence given above may allow the following generalisations.

First, the most striking contrast is the preference for verbal structures in the first person set versus the tendency to prefer nominal structures in the third person set, as is illustrated by the locations of the verbs, the definite and indefinite articles, and the major pronouns. The distribution of adverbials of direction towards one end of the spectrum and words functioning as a determiner towards the other is in keeping with this pattern.

Second, the weighting on determiners in the third person narratives bears relation to another point: habits of reference. While the third person narratives are characterised by more elaborate and explicit reference, which is manifest in the incidence of the determiners and the relatives, *which* and *who*, the first person narrator turn more freely to looser reference. One can derive further evidence from the presence of the pronoun *it*, the demonstrative *that*(d), and the place adverbial *there*.

Third, the frequent recourse to the relatives *which* and *who* betokens a tendency towards hypotactic structures in the third person set of texts. On the other hand, the coordinate conjunctions *and* and *but* are more commonly associated with the first person set of texts.

Fourth, the language of the first person narratives seems to be tinged with more emotional colouring and subjectivity compared with that of the third person narratives. The predominance of a set of intensifiers betokens the first person narrators' distinctive habits of emphasis. The abundance of negatives, as I noted earlier, can be considered a measure of subjectivity. It may not be inappropriate to assume that these tendencies illustrate one feature of the language of narrative in which a narrator's sense of value tends to be more freely reflected. The third person narrators, on the other hand, assume a stance closer to "what is called omniscient point of view" and seem to narrate with some degree of emotional detachment.

Turning finally to the distribution of personal pronouns, in this study, third person singular masculine pronouns and third person plural pronouns except *them* are highly associated with the third person narratives. Third person singular feminine pronouns, on the other hand, are predominant in the first person narratives. Such a difference in distributions between the masculine and the feminine can be also observed in the Bank of English, as Nakamura (1994) points out.[9] However given the size and the number of texts examined in this study, it may be safe to assume that the different distributions of those pronouns are to some extent thematically determined. Contextual examination of each instance of the pronouns, with the help of a concordance, also adds to this view.

These features, taken together, seem to indicate that the differentiae between the first person and the third person narratives point towards a broad opposition between a more "oral" style and a more "literate" style: a language tinged with more emotional colouring versus a language of more elaborate and generalising cast; more subjective narratives versus narratives with greater descriptive emphasis; a style more given to verbal structures versus a style that has a comparative preference for nominal structures.

In this analysis, a larger number of verbs ranked within the top 100 words than in my earlier studies as a result of lemmatisation. As a future step, it would be worthwhile to examine the distribution of those verbs in greater detail, by casting light on their collocation, to look deeper into the structure of Dickens's language of narrative.

Notes

* Part of 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 am grateful for the support and assistance of my colleagues: John Burrows, Hugh Craig, Wayne McKenna, Nicole Cox, Alexis Antonia, and John Lambert. The study was also supported by a Japanese Ministry of Education grant for studying abroad for the academic year 1993.

[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)
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 (1986, 1st p.: 1837) ed. Robert L. Patten
Nicholas Nickleby (1978, 1st p.: 1839) ed. Michael Slater
Oliver Twist (1985, 1st p.: 1838) ed. Peter Fairclough
Bleak House (1987, 1st p.: 1853) ed. Norman Page
A Tale of Two Cities (1970, 1st p.: 1859) ed. George Woodcock
The Mystery of Edwin Drood (1985, 1st p.: 1870) ed. Arthur J. Cox

[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] As to residual segments, those of more than 3000 words are standardised to 4000: smaller residues are incorporated in the preceding segment.

[4] Cf. Tabata (1993: 119), and (1994: 168-71).

[5] It may be necessary to mention here that considerable amounts of time and effort are needed to carry out these preparations of texts, such as tagging, lemmatisation, hyphenation, and so on.

[6] "Word-type" will be referred to as "word" henceforward unless it is necessary to make a clear-cut distinction.

[7] In Tabata (1994), which is based on the same corpus, only seven types of verb are ranked within the top 100: *came*, *looked*, *made*, *said*, *replied*, *HAVE* (*have* & *had*), *BE* (*be*, *is*, *was*, *were*, and *been*).

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

[9] In his analysis of personal pronouns in the Bank of English, Nakamura demonstrates the dissimilar distribution of *she* and *he* along Axes 2 and 3 (Nakamura, 1994: 164).

References

Beaman, K. (1984) “Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse.” In Tannen (ed.), pp. 45-80.

Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge UP.

——— (1993) “The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings.” *Computers and the Humanities*, 26: 331-345.

Biber, D. and E. Finegan (1989) “Drift and the Evolution of English Style: A History of Three Genres.” *Language*, Volume 65, 3: 487-517.

——— eds. (1994) *Sociolinguistic Perspectives on Register*. New York & Oxford: Oxford UP.

Brainerd, B. (1979) “Pronouns and Genre in Shakespeare’s Drama.” *Computers and the Humanities*, 13: 3-16.

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

——— (1987b) “Word-Patterns and Story-Shapes: The Statistical Analysis of Narrative Style.” *Literary and Linguistic Computing*, 2: 61-70.

——— (1989) “‘A Vision’ as a revision?” *Eighteenth-Century Studies*, 22: 551-65.

——— (1992) “Computers and the Study of Literature.” In Butler, pp. 167-204.

——— (1994) “Tiptoeing into the Infinite: Testing for National Differences in the Language of English Narrative.” In Hockey and Ide.

Burrows, J. F. and D. H. Craig (1994) “Lyrical Drama and the ‘Turbid Mountebanks’: Styles of Dialogue in Romantic and Renaissance Tragedy.” *Computers and the Humanities*, 28: 63-86.

Butler, C. ed. (1992) *Computers and Written Texts*. Oxford: Blackwell.

Computers and the Humanities. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Cluett, R. (1976) *Prose Style and Critical Reading*. New York & London: Teacher’s College Press.

Craig, D. H. (1991) “Plural Pronouns in Roman Plays by Shakespeare and Johnson”, *Literary and Linguistic Computing*, 6: 180-186.

——— (1992) “Authorial Styles and the Frequencies of Very Common Words: Jonson, Shakespeare and the Additions to *The Spanish Tragedy*”, *Style*, Volume 26, 2: 199-220.

Hockey, S. and N. Ide eds. (1996) *Research in Humanities Computing 4*. London: Oxford UP.

Literary & Linguistic Computing: Journal of the Association for Literary and Linguistic Computing. Oxford: Oxford UP.

Nakamura, J. (1994) "Extended HAYASHI's Quantification Method Type III and its Application in Corpus Linguistics." *Journal of Language and Literature, Faculty of Integrated Arts and Sciences, University of Tokushima*, 1: 141-192.

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

Potter, R. G. ed. (1989) *Literary Computing and Literary Criticism*. Philadelphia: University of Pennsylvania Press.

Pugh, C. S. (1992) "Steinbeck no shousetsu—Computer ni yoru buntaibunseki no kanousei—." In Saito ed.

Revue Informatique et Statistique dans les Sciences Humaines (*RISSH*), University of Liege, Belgium.

Saito, T. ed. (1992) *Eigoeibungakukenkyu to Computer*. Tokyo: Eichosha.

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

Tabata, T. (1991) "Characterization in Dickens's *Christmas Books*: A Computer-Assisted Approach to Idiolects." *Kumamoto Studies in English Language and Literature*, 34: 98-126.

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

——— (1994) "Dickens's Narrative Style: A Statistical Approach to Chronological Variation." *Revue Informatique et Statistique dans les Sciences Humaines* (*RISSH*), 30: 165-182.

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

Toolan, M. J. (1988) *Narrative: A Critical Linguistic Introduction*. London & New York: Routledge.

Watt, I. (1960) "The first Paragraph of *The Ambassadors*: an explication." *Essays in Criticism*, 10: 250-74.

(Osaka University E-mail: tabata@lisa.lang.osaka-u.ac.jp)