Extended HAYASHI's Quantification Method Type III and its Application in Corpus Linguistics[1]

Junsaku NAKAMURA

Abstract

The present study focuses on one of several methods useful in describing text types based upon the vocabulary frequency list, Extended HAYASHI's Quantification Method Type III, a statistical method used in multivariate analysis. The main objective of this study is to introduce notions underlying this statistical method and to demonstrate its use in describing text typology in large corpora.

Extended HAYASHI's Quantification Method Type III is useful in the processing of the frequency table of the pronominal forms across 728 textual samples taken from the Book Corpus, the Spoken Corpus and the Times Corpus accumulated in the Bank of English at COBUILD. On the basis of the distribution of pronouns, the structure of the Bank of English is determined. Relationships among pronominal forms, relationships among textual samples and, more importantly, relationships between pronominal forms and textual samples are made explicit through the use of this statistical method. The "personal involvement vs. personal detachment" factor plays an important role in determining the structure of the corpus.

This study concludes with suggestions for possible applications in corpus linguistics.

[1] The original version of this paper was presented at the 14th ICAME conference held on May 19-23, 1993, at Panorama Hotel, Feusisberg, Switzerland. The study was supported by a Japanese Ministry of Education grant for studying abroad for the fiscal year 1991. The author hereby wishes to express his gratitude to CARE (Center for Advanced Research in English) at the University of Birmingham and COBUILD, which provided the data in the Bank of English as well as hardware and software. He also thanks the computer officers of COBUILD, Zoe James and Tim Lane, for their patient assistance, without which this work could have never been completed.

0. Introduction

As Fig. 1 shows, various methods have been proposed for analysing and classifying English texts on the basis of frequency count of such items as vocabulary and grammatical tags. Basically, there are two major techniques for classifying English texts: 1) techniques based upon the similarity or distance matrix calculated from the frequency table and 2) techniques based on direct processing of the frequency table.

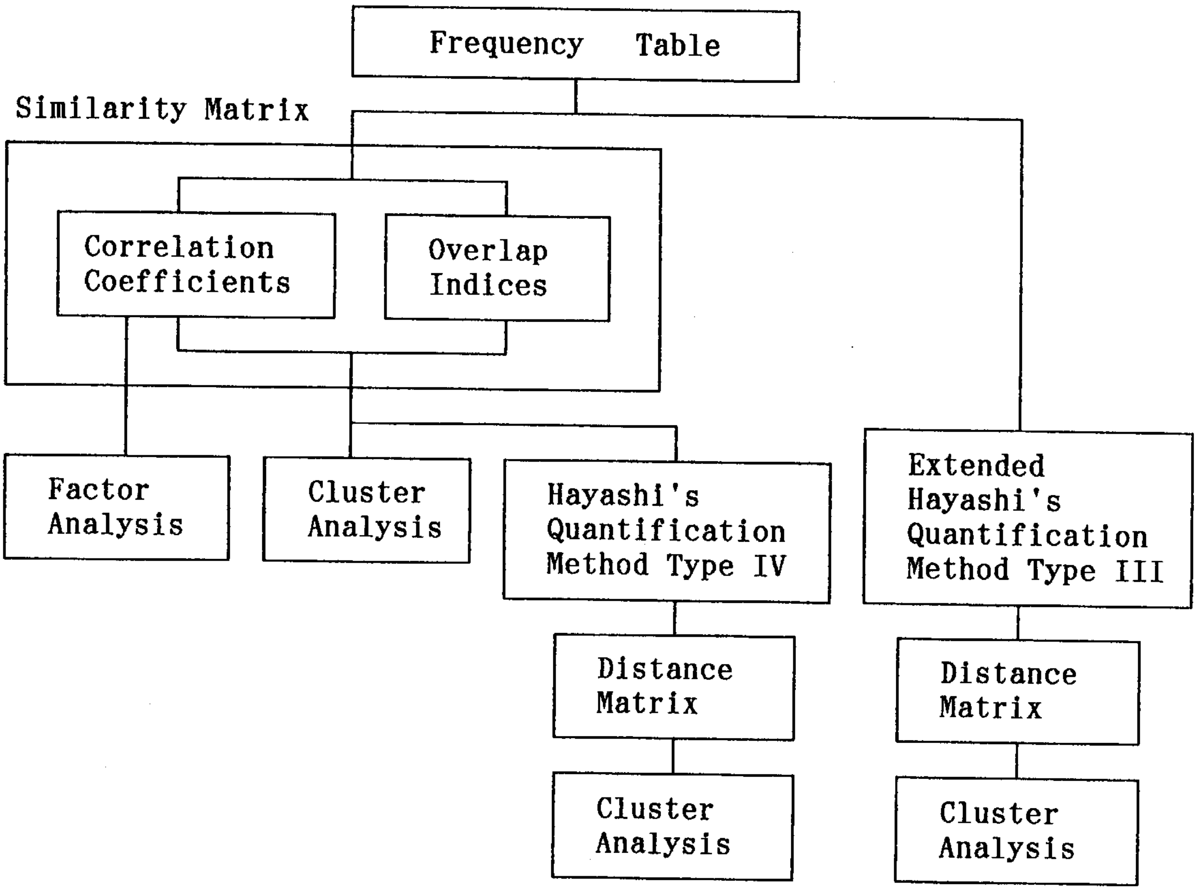

Fig. 1. Various Methods for Analysing Frequency Table.

The techniques based on the similarity or distance matrix are again subdivided into two methodologies depending on the kind of matrix involved. One of these methodologies uses a matrix of correlation coefficients calculated from the frequency table to indicate the degree of similarity among various text categories. The other methodology uses a matrix showing the degree of overlap of the vocabulary of one text category with the vocabulary of other text categories.

HAYASHI's Quantification Method Type IV (a statistical technique essentially similar to multidimensional scaling) is useful in processing the similarity matrix obtained either through correlation coefficients or through overlap indices, locating the categories in a multi-dimensional space and showing the relationships among categories. Alternately, the similarity matrices may be fed into cluster analysis of one kind or another. The matrix consisting exclusively of correlation coefficients can further be processed by factor analysis, a powerful statistical technique used in BIBER (1988 and 1989).

The statistical technique employed in this study differs from the techniques mentioned above because it directly deals with the frequency table without intervening steps such as similarity matrices or correlation coefficients. This method is not so well-known in the West, probably because it was developed by a Japanese statistician, and the main objective of this paper is to introduce the notions underlying HAYASHI's Quantification Method Type III and its extended version and to demonstrate its application in describing text typology in large corpora.

To describe the structure of the Bank of English from the viewpoint of the use of pronominal forms, an enormous amount of textual data created and accumulated in COBUILD (Collins Birmingham University International Language Database) has been processed, using the extended version of HAYASHI's Quantification Method Type III. Focusing on the distribution of the personal pronouns, this method was used to examine how 728 texts relate to one another. In so doing, this method also proved its worth in determining text typology.

1. Data

1.1. Corpora

Three extensive corpora summarised in Table 1, the Book Corpus, the Spoken Corpus and the Times Corpus installed and accessed in the Bank of English as of April 1992, were chosen to show how this methodology works.

The Book Corpus consists of 319 texts, averaging about 60,000 words, taken mainly from books published by HarperCollins and Thorsons from

Table 1. Corpora Analyzed in the Bank of English.

Name of Corpus	Number of Tokens	Number of Texts	Average Text Size
Book Corpus	19,333,737	319	60,607.3
Spoken Corpus	2,950,502	288	10,244.8
Times Corpus	11,089,548	121	91,649.2
Total	33,373,787	728	45,843.1

1983 to 1992. Those texts represent various genres ranging from mystery and crime fiction, to health and fitness, and to astrology. The Spoken Corpus comprises 288 texts, representing various examples of spontaneous spoken English of various modes and events, such as face-to-face conversations, telephone conversations, radio phone-in programmes, seminars, lectures, and so on. Topics range from general (mainly current affairs) to highly specialised, such as "Coleridge as a Critic." The Times Corpus, representing newspaper English, includes 121 texts taken from *The Times* from April through July, 1991. The texts, averaging around 90,000 words, were collected daily. The BBC Corpus was then a part of the Bank of English, but the present analysis does not include it because, unlike other corpora, it is not divided into subtexts. In all, the present study analyses 728 texts.[2]

1.2. Personal Pronouns Selected and their Groupings

This study of the structure of the Bank of English focuses on personal pronouns since personal pronouns are quite context-dependent. This high context-dependency of personal pronouns is discussed in KUČERA and FRANCIS (1967), MOSTELLER and ROURKE (1982) and FRANCIS and KUČERA (1982) for the Brown Corpus and summarised in NAKAMURA (1989: 53-54).

[2] The number of texts in the Spoken Corpus was considerably increased and some of the texts were taken out of the Book Corpus after the data for the present study was collected. These changes were not dealt with here. The present study reflects only the Bank of English as of April, 1992.

Later NAKAMURA (1989 and 1990) adds evidence by studying the structure of the Brown Corpus from the viewpoint of distribution of personal pronouns and from the viewpoint of distribution of grammatical tags. NAKAMURA (1991) also studies the corpus structure based upon the distribution of grammatical tags and obtains exactly the same kind of evidence for the LOB Corpus.

All these studies suggest that the structure of the corpus can be determined by the way personal pronouns are used. When a writer writes a text, he has to provide an appropriate, definite perspective for the reader, the people and the things described and the writer himself. Depending upon the

Table 2. Groupings of Personal Pronouns and the Word Types Included in Each Category.

Category	Word Types
First Person Singular Pronouns (I)	I, I'd, I'll, I'm, I've, me, my, mine, mine's, myself, gimme
First Person Plural Pronouns (WE)	we, we'd, we'll, we're, we've, us, our, ours, ourselves
Second Person Pronouns (YOU)	you, you'd, you'll, you're, you've, your, yours, yourself, yourselv, yourselves, y'all, y'd, y'know, y're, y'see, y've, y'from, ye, ye're, you's, yuh, d'you, d'ye, whaddya, willya
Third Person Sigular Masculine Pronouns (HE)	he, he'd, he'll, he's, his, him, himself, 'im, himselfe, hisself, himsijlf
Third Person Singular Feminine Pronouns (SHE)	she, she'd, she'll, she's, her, hers, herself, hir
Third Person Singular Neuter Pronouns (IT)	it, it'd, it'll, it's, its, itself
Third Person Plural Pronouns (THEY)	they, they'd, they'll, they're, they've, their, theirs, them, 'em, their's, themselves, themselv, 'emselves, thay, ther

kind of text he is writing, the writer will choose pronouns to reflect his perspective. And this is also true of the spoken texts.

The types of personal pronouns used in the present analysis include a base form, its inflectional or spelling variants, and merged or contracted forms. Reflexive pronouns with spelling variants are included since they also reflect the writer/speaker's perspective in exactly the same way as the personal pronouns do. The pronouns chosen then fall into seven categories according to person, number and gender (as they are distinguished in one way or another by these grammatical categories). Table 2 shows the groupings of personal pronouns and word types included in each category.[3] Hereafter, the generic notations using uppercase letters (such as "I" and "WE") placed in the parentheses identify these seven categories.

1.3. Original Frequency Data

The number of occurrences of 7 categories of pronouns was counted across the 728 texts. Table 3 is a part of the results of this count, listing the first, the middle and the last fifteen texts.[4] The figures at the bottom of the table indicate a total of 33,373,787 tokens, of which 2,151,555 tokens are pronouns. The most frequent type of pronoun, "HE," occurs 422,631 times. The least frequent type, "WE," occurs only 175,486 times, much less than half of the most frequent types.

Genre sizes in the column at far right range from 222 tokens to 196,207 tokens. The smallest text comes from the Spoken Corpus, i.e., S12, and likewise, most of the texts of small sizes (usually telephone conversations) are from the Spoken Corpus. The largest text is T13791, *The Times* of July 13th, 1991. Most big texts are either from the Times Corpus or from the Book Corpus. The average text size is 45,843.1 tokens.

Since the raw frequency data in Table 3 are from texts of different size,

[3] Disambiguation of homographs such as "mine" (meaning "a place where minerals are dug" or "dig up minerals from such a place" or "a kind of bomb") is done partly by "grepping out" collocations of "mine" used as a noun or as a verb and partly by looking through the concordance lines on the screen.

[4] For complete data, contact the author.

Table 3. Raw Frequency Data.

No	Text ID	HE	SHE	IT	THEY	YOU	WE	I	Total	Size
1	B 4	2329	996	982	410	810	161	1155	6843	62343
2	B 5	2553	710	1353	739	1041	370	1346	8112	72031
3	B 6	1223	1317	913	953	857	472	1151	6886	56191
4	B 7	1995	390	936	548	836	252	1293	6250	57587
5	B 8	2009	1206	799	575	1065	271	1527	7452	64833
6	B 9	2252	1155	826	561	778	312	1247	7131	58304
7	B 10	996	862	1006	495	968	557	3447	8331	71042
8	B 11	1880	974	868	569	809	277	1208	6585	56674
9	B 12	1916	2102	873	717	681	240	873	7402	64453
10	B 13	2277	1438	1026	606	711	168	968	7194	60242
11	B 14	1501	2147	1077	326	1238	168	1618	8075	52177
12	B 15	1768	2235	1011	570	1123	363	1587	8657	57674
13	B 16	2007	691	885	749	820	311	924	6387	53316
14	B 17	1037	1282	953	731	672	338	787	5800	49883
15	B 18	1159	828	923	550	805	384	2921	7570	61272
364	S 53	25	13	261	126	359	87	352	1223	9830
365	S 54	46	0	301	158	361	131	425	1422	10817
366	S 55	105	23	233	170	350	99	465	1445	10558
367	S 56	76	50	295	122	339	104	562	1548	10860
368	S 57	49	17	237	143	348	72	398	1264	9835
369	S 58	39	65	195	252	272	65	423	1311	10231
370	S 59	50	7	198	114	264	124	420	1177	9265
371	S 60	87	3	250	122	281	79	493	1315	10962
372	S 61	43	58	195	174	349	107	523	1449	10064
373	S 62	56	16	320	172	263	119	459	1405	11020
374	S 63	47	79	295	125	366	104	419	1435	11215
375	S 64	64	11	216	123	422	114	470	1420	10576
376	S 65	88	44	175	179	256	165	407	1314	10107
377	S 66	38	64	236	158	242	109	402	1249	10664
378	S 67	139	44	212	147	236	92	422	1292	10895
714	T 27791	1382	366	1016	716	272	379	480	4611	106307
715	T 28491	2271	647	1604	1162	423	540	892	7539	143021
716	T 28591	933	171	590	432	65	159	226	2576	75749
717	T 28691	913	107	668	496	104	190	216	2694	73323
718	T 28791	2217	593	1645	1249	360	504	863	7431	152859
719	T 29491	964	150	603	714	72	263	237	3003	75656
720	T 29591	789	138	665	518	69	161	220	2560	75858
721	T 29691	1503	582	1036	773	312	411	719	5336	105402
722	T 29791	953	162	701	583	108	248	230	2985	81770
723	T 30491	795	122	651	406	70	191	181	2416	71851
724	T 30591	1097	142	697	552	102	268	288	3146	82407
725	T 30691	1989	698	1796	1284	357	658	773	7555	158413
726	T 30791	729	169	603	482	107	180	215	2485	72938
727	T 31591	805	286	589	454	102	126	265	2627	68026
728	T 31791	793	225	656	493	111	218	249	2745	73430
Total		422631	180267	399514	279859	321789	175486	372009	2151555	33373787

the figures in the table cannot be compared as given. The figures also cannot be processed directly by Extended HAYASHI's Quantification Method Type III because the ways in which the writer/speaker of the texts uses pronouns are apparently independent of text size. This extended version of HAYASHI's Quantification Method Type III is sensitive to the text sizes. Therefore, the next step is to adjust the frequency figures of Table 3 to equal text sizes so that comparison of texts and pronouns is uninfluenced by the text size. In the present case, the figure 45,000 (close to the average text size and easy to process) is used as a text size for adjusting frequencies, making the size of the entire corpus 32,760,000 tokens.

1.4. Adjusted Frequency Data

Table 4 shows some results of adjusting the original frequencies to the text size of 45,000 tokens. The frequencies are calculated on an equal basis, yielding interesting observations. For example, the column sums in Table 4 show the distribution of pronouns in the entire corpus as in Table 5.

The total number of pronouns used in the entire corpus of 32,760,000 tokens is 2,619,475.9, representing 8.00% of the total tokens. About one in 12.5 words in the texts is actually a pronoun of one kind or another.

The most frequent type of pronouns is "I," occurring 592,135.7 times out of 2,619,475.9 tokens of pronouns, 22.61%. Its ratio of occurrence among the total number of tokens is 1.81%, meaning that about two out of ten pronouns are "I" and there is a possibility of encountering about two occurrences of "I" in the text of 100 tokens. The least frequent type of pronoun is "SHE," occurring 139,120.4 times and representing only 5.31% of the pronouns. The frequency of "SHE" is approximately one fourth of the frequent types like "I," "YOU" and "IT."

The "Total" column in Table 4 enables the text-by-text comparison of the use of pronouns. Table 6 lists 15 texts characterised either by the most frequent or the least frequent pronoun use. Texts B47, S265, B19, B27, and B40 use pronouns most frequently, and their percentages of pronouns among total tokens range from 16.9% to 15.9%, meaning that in these texts, every sixth word is a pronoun. In contrast to this high ratio of pronouns, B695 and B646 contain only 0.3% and 0.4% of pronouns respectively, and it is

Table 4. Adjusted Frequency Data.

No	Text ID	HE	SHE	IT	THEY	YOU	WE	I	Total
1	B 4	1681.1	718.9	708.8	295.9	584.7	116.2	833.7	4939.4
2	B 5	1594.9	443.6	845.3	461.7	650.3	231.2	840.9	5067.8
3	B 6	979.4	1054.7	731.2	763.2	686.3	378.0	921.8	5514.6
4	B 7	1559.0	304.8	731.4	428.2	653.3	196.9	1010.4	4883.9
5	B 8	1394.4	837.1	554.6	399.1	739.2	188.1	1059.9	5172.4
6	B 9	1738.1	891.4	637.5	433.0	600.5	240.8	962.5	5503.8
7	B 10	630.9	546.0	637.2	313.5	613.2	352.8	2183.4	5277.1
8	B 11	1492.8	773.4	689.2	451.8	642.4	219.9	959.2	5228.6
9	B 12	1337.7	1467.6	609.5	500.6	475.5	167.6	609.5	5168.0
10	B 13	1700.9	1074.2	766.4	452.7	531.1	125.5	723.1	5373.8
11	B 14	1294.5	1851.7	928.9	281.2	1067.7	144.9	1395.4	6964.3
12	B 15	1379.5	1743.9	788.8	444.7	876.2	283.2	1238.3	6754.6
13	B 16	1694.0	583.2	747.0	632.2	692.1	262.5	779.9	5390.8
14	B 17	935.5	1156.5	859.7	659.4	606.2	304.9	710.0	5232.2
15	B 18	851.2	608.1	677.9	403.9	591.2	282.0	2145.3	5559.6
364	S 53	114.4	59.5	1194.8	576.8	1643.4	398.3	1611.4	5598.7
365	S 54	191.4	0.0	1252.2	657.3	1501.8	545.0	1768.1	5915.7
366	S 55	447.5	98.0	993.1	724.6	1491.8	422.0	1981.9	6158.8
367	S 56	314.9	207.2	1222.4	505.5	1404.7	430.9	2328.7	6414.4
368	S 57	224.2	77.8	1084.4	654.3	1592.3	329.4	1821.1	5783.4
369	S 58	171.5	285.9	857.7	1108.4	1196.4	285.9	1860.5	5766.3
370	S 59	242.8	34.0	961.7	553.7	1282.2	602.3	2039.9	5716.7
371	S 60	357.1	12.3	1026.3	500.8	1153.5	324.3	2023.8	5398.2
372	S 61	192.3	259.3	871.9	778.0	1560.5	478.4	2338.5	6479.0
373	S 62	228.7	65.3	1306.7	702.4	1074.0	485.9	1874.3	5737.3
374	S 63	188.6	317.0	1183.7	501.6	1468.6	417.3	1681.2	5757.9
375	S 64	272.3	46.8	919.1	523.4	1795.6	485.1	1999.8	6042.0
376	S 65	391.8	195.9	779.2	797.0	1139.8	734.6	1812.1	5850.4
377	S 66	160.4	270.1	995.9	666.7	1021.2	460.0	1696.4	5270.5
378	S 67	574.1	181.7	875.6	607.2	974.8	380.0	1743.0	5336.4
714	T 27791	585.0	154.9	430.1	303.1	115.1	160.4	203.2	1951.8
715	T 28491	714.5	203.6	504.7	365.6	133.1	169.9	280.7	2372.1
716	T 28591	554.3	101.6	350.5	256.6	38.6	94.5	134.3	1530.3
717	T 28691	560.3	65.7	410.0	304.4	63.8	116.6	132.6	1653.4
718	T 28791	652.7	174.6	484.3	367.7	106.0	148.4	254.1	2187.6
719	T 29491	573.4	89.2	358.7	424.7	42.8	156.4	141.0	1786.2
720	T 29591	468.0	81.9	394.5	307.3	40.9	95.5	130.5	1518.6
721	T 29691	641.7	248.5	442.3	330.0	133.2	175.5	307.0	2278.1
722	T 29791	524.5	89.2	385.8	320.8	59.4	136.5	126.6	1642.7
723	T 30491	497.9	76.4	407.7	254.3	43.8	119.6	113.4	1513.1
724	T 30591	599.0	77.5	380.6	301.4	55.7	146.3	157.3	1717.9
725	T 30691	565.0	198.3	510.2	364.7	101.4	186.9	219.6	2146.1
726	T 30791	449.8	104.3	372.0	297.4	66.0	111.1	132.6	1533.2
727	T 31591	532.5	189.2	389.6	300.3	67.5	83.4	175.3	1737.8
728	T 31791	486.0	137.9	402.0	302.1	68.0	133.6	152.6	1682.2

Table 5. Frequency Distribution of Pronouns.

Category	Absolute Frequency	Proportion among Pronouns (%)	Proportion in the Entire Corpus (%)
HE	337,094.1	12.85	1.03
SHE	139,120.4	5.31	0.42
IT	505,148.8	19.28	1.54
THEY	315,306.4	12.04	0.96
YOU	512,407.0	19.56	1.56
WE	218,263.4	8.33	0.67
I	592,135.7	22.61	1.81
Total	2,619,475.9	100.00	8.00
Total Number of Tokens in the Entire Corpus			32,760,000

Table 6. Comparison of the Total Number of Pronouns Used in Each Text.

Maximum 15 Texts			Minimum 15 Texts		
Text ID	Frequency	Ratio	Text ID	Frequency	Ratio
B 47	7623.5	16.9	B 695	130.1	0.3
S 265	7293.4	16.2	B 646	170.6	0.4
B 19	7238.2	16.1	B 595	312.2	0.7
B 27	7199.5	16.0	B 697	344.4	0.8
B 40	7162.8	15.9	B 749	352.0	0.8
S 138	7032.9	15.6	B 632	367.6	0.8
S 199	6970.1	15.5	B 615	376.0	0.8
B 14	6964.3	15.5	B 820	398.1	0.9
B 25	6920.9	15.4	B 592	402.0	0.9
B 37	6916.9	15.4	B 747	408.6	0.9
S 317	6916.1	15.4	B 678	413.2	0.9
B 48	6817.7	15.2	B 621	449.9	1.0
S 137	6809.7	15.1	B 746	451.4	1.0
S 119	6802.6	15.1	B 804	481.2	1.1
S 13	6793.0	15.1	B 779	492.0	1.1

expected that only three or four pronouns appear in the 1,000 tokens text. B47, B19 and B27 are forms of fiction categorised as "crime, mystery and thriller" written by the same author, while B695 is a book about solving crossword puzzles which lists the words matching clues. B646 is a book concerning acupuncture.

Table 4 also enables comparison of texts from the viewpoint of the use of individual pronouns. Table 7 lists another 15 texts characterised either by the most frequent or least frequent use of "HE." Most of texts that use "HE" frequently are from the Book Corpus. Five books, B101, B76, B609, B96, and B41, use "HE" more than 2,250 times, i.e., 5% of the text size. B101, B96, and B41 are works of fiction. B76 is a biography of Charlie Chaplin and B609 is a book titled *Men! A Collector's Guide*. Among 15 texts with the highest ratio of "HE," one text from the Spoken Corpus, S156, is a university lecture, "Coleridge as Thinker." In 19 texts, 5 in the Book Corpus and 14 in the Spoken Corpus, "HE" is not used although only 15

Table 7. Comparison of the Number of HE Used in Each Text.

Maximum 15 Texts			Minimum 15 Texts		
Text ID	Frequency	Ratio	Text ID	Frequency	Ratio
B 101	2519.8	5.6	B 592	0.0	0.0
B 76	2453.1	5.5	B 593	0.0	0.0
B 609	2407.4	5.3	B 595	0.0	0.0
B 96	2331.4	5.2	B 615	0.0	0.0
B 41	2268.5	5.0	B 618	0.0	0.0
B 100	2249.3	5.0	S 7	0.0	0.0
B 22	2103.3	4.7	S 9	0.0	0.0
B 34	2062.2	4.6	S 11	0.0	0.0
B 49	2058.7	4.6	S 12	0.0	0.0
S 156	1993.9	4.4	S 23	0.0	0.0
B 55	1992.9	4.4	S 24	0.0	0.0
B 693	1968.6	4.4	S 77	0.0	0.0
B 650	1905.7	4.2	S 78	0.0	0.0
B 38	1846.3	4.1	S 79	0.0	0.0
B 48	1838.2	4.1	S 87	0.0	0.0

texts are listed in the table. All five books are about cooking in relation to health and fitness. *Cooking for Diabetes* is an example. Six of the 14 texts from the Spoken Corpus are the lecture series "Development Finance."

The adjusted frequency table yields interesting information, but the ordinary descriptive statistics method cannot show the relationships 1) among pronominal forms themselves, 2) among texts themselves and 3) between pronominal forms and texts in sufficient detail. A more sophisticated statistical technique of multivariate analysis such as HAYASHI's Quantification Method Type III is necessary.

2. Method

2.1. Gist of HAYASHI's Quantification Method Type III

HAYASHI's Quantification Method Type III quantifies qualitative or attributive categories and samples simultaneously. The distinctive feature of this method is that it can classify or quantify, without requiring any external criterion, both categories and samples by processing qualitative dichotomic response patterns. Table 8 is a schematic example. Each row

Table 8. Data Matrix for Dichotomic Response Patterns.

	1	2	3	4	5	6	7
1		x			x		
2	x			x			x
3			x				x
4	x				x		
5			x	x			
6		x				x	
7			x				x
8						x	
9	x					x	x
10		x		x			

representing a sample reacts positively (indicated by "x") or negatively to several columns representing categories.

Although this method was originally proposed to deal with dichotomic response patterns such as in Table 8, it can be easily extended to deal with ordinary frequency cross tables like Table 4. The present article discusses this extended version. To introduce the statistical method in question, a simple illustration using the adjusted frequency will obviate the necessity of over complicated, highly technical explanation.[5]

At first sight, Table 4 might seem to have no definite pattern or clear tendency for categories and samples. It would be impossible to process so many items visually and manually. The fact that there is no definite pattern or clear tendency is shown when quantities are given to categories and samples (as in the parentheses of Table 9) and the correlation coefficient is calculated. Here, the quantities given to categories are called "category weights" and the quantities given to samples are called "sample scores." This set of category weights and sample scores yields a correlation coefficient of about 0.103, indicating very low correlation of categories and samples.[6]

Generally speaking, when the data are like the data in Table 4, we can

[5] Those interested in the algorithm of the original version of this statistical technique can refer to a standard textbook on statistics, but as this technique was developed by a Japanese statistician, it may be difficult to find in the textbooks written in English or other European languages. Most Japanese statistics textbooks (including the one in the references) contain HAYASHI's Quantification Method Type III. Statistical packages of the original version both for main-frame computers and personal computers are also available in Japan.

[6] Actually, the seemingly low correlation coefficient obtained here is relatively high, considering the number of cases involved in its calculation. This relatively high correlation coefficient is probably due to the fact that all the texts pertaining either to the Book Corpus, to the Spoken Corpus, or to the Times Corpus are listed together consecutively. If the order of 728 texts and the order of pronominal categories are completely randomised, the correlation coefficient close to nil will certainly be obtained, indicating no correlation whatsoever. However, this statistical method still works for the randomised frequency table.

Table 9. Data Matrix with Category Weights and Sample Scores.

No	Text ID		HE (1.0)	SHE (2.0)	IT (3.0)	THEY (4.0)	YOU (5.0)	WE (6.0)	I (7.0)
1	B 4	(1.0)	1681.1	718.9	708.8	295.9	584.7	116.2	833.7
2	B 5	(2.0)	1594.9	443.6	845.3	461.7	650.3	231.2	840.9
3	B 6	(3.0)	979.4	1054.7	731.2	763.2	686.3	378.0	921.8
4	B 7	(4.0)	1559.0	304.8	731.4	428.2	653.3	196.9	1010.4
5	B 8	(5.0)	1394.4	837.1	554.6	399.1	739.2	188.1	1059.9
6	B 9	(6.0)	1738.1	891.4	637.5	433.0	600.5	240.8	962.5
7	B 10	(7.0)	630.9	546.0	637.2	313.5	613.2	352.8	2183.4
8	B 11	(8.0)	1492.8	773.4	689.2	451.8	642.4	219.9	959.2
9	B 12	(9.0)	1337.7	1467.6	609.5	500.6	475.5	167.6	609.5
10	B 13	(10.0)	1700.9	1074.2	766.4	452.7	531.1	125.5	723.1
11	B 14	(11.0)	1294.5	1851.7	928.9	281.2	1067.7	144.9	1395.4
12	B 15	(12.0)	1379.5	1743.9	788.8	444.7	876.2	283.2	1238.3
13	B 16	(13.0)	1694.0	583.2	747.0	632.2	692.1	262.5	779.9
14	B 17	(14.0)	935.5	1156.5	859.7	659.4	606.2	304.9	710.0
15	B 18	(15.0)	851.2	608.1	677.9	403.9	591.2	282.0	2145.3
364	S 53	(364.0)	114.4	59.5	1194.8	576.8	1643.4	398.3	1611.4
365	S 54	(365.0)	191.4	0.0	1252.2	657.3	1501.8	545.0	1768.1
366	S 55	(366.0)	447.5	98.0	993.1	724.6	1491.8	422.0	1981.9
367	S 56	(367.0)	314.9	207.2	1222.4	505.5	1404.7	430.9	2328.7
368	S 57	(368.0)	224.2	77.8	1084.4	654.3	1592.3	329.4	1821.1
369	S 58	(369.0)	171.5	285.9	857.7	1108.4	1196.4	285.9	1860.5
370	S 59	(370.0)	242.8	34.0	961.7	553.7	1282.2	602.3	2039.9
371	S 60	(371.0)	357.1	12.3	1026.3	500.8	1153.5	324.3	2023.8
372	S 61	(372.0)	192.3	259.3	871.9	778.0	1560.5	478.4	2338.5
373	S 62	(373.0)	228.7	65.3	1306.7	702.4	1074.0	485.9	1874.3
374	S 63	(374.0)	188.6	317.0	1183.7	501.6	1468.6	417.3	1681.2
375	S 64	(375.0)	272.3	46.8	919.1	523.4	1795.6	485.1	1999.8
376	S 65	(376.0)	391.8	195.9	779.2	797.0	1139.8	734.6	1812.1
377	S 66	(377.0)	160.4	270.1	995.9	666.7	1021.2	460.0	1696.4
378	S 67	(378.0)	574.1	181.7	875.6	607.2	974.8	380.0	1743.0
714	T 27791	(714.0)	585.0	154.9	430.1	303.1	115.1	160.4	203.2
715	T 28491	(715.0)	714.5	203.6	504.7	365.6	133.1	169.9	280.7
716	T 28591	(716.0)	554.3	101.6	350.5	256.6	38.6	94.5	134.3
717	T 28691	(717.0)	560.3	65.7	410.0	304.4	63.8	116.6	132.6
718	T 28791	(718.0)	652.7	174.6	484.3	367.7	106.0	148.4	254.1
719	T 29491	(719.0)	573.4	89.2	358.7	424.7	42.8	156.4	141.0
720	T 29591	(720.0)	468.0	81.9	394.5	307.3	40.9	95.5	130.5
721	T 29691	(721.0)	641.7	248.5	442.3	330.0	133.2	175.5	307.0
722	T 29791	(722.0)	524.5	89.2	385.8	320.8	59.4	136.5	126.6
723	T 30491	(723.0)	497.9	76.4	407.7	254.3	43.8	119.6	113.4
724	T 30591	(724.0)	599.0	77.5	380.6	301.4	55.7	146.3	157.3
725	T 30691	(725.0)	565.0	198.3	510.2	364.7	101.4	186.9	219.6
726	T 30791	(726.0)	449.8	104.3	372.0	297.4	66.0	111.1	132.6
727	T 31591	(727.0)	532.5	189.2	389.6	300.3	67.5	83.4	175.3
728	T 31791	(728.0)	486.0	137.9	402.0	302.1	68.0	133.6	152.6

R = 0.1029827D+00

imagine that certain types of samples or texts may show preference for certain types of categories or pronouns. In other words, there might be a hidden tendency or pattern in the original data matrix. To find this hidden tendency, we change the positions of rows and columns of the original data matrix (see Table 10) so that the proportionally large frequency figures can converge around the diagonal. Rows and columns are so completely changed that the table does not look like the original data matrix, but the response of each text to each pronoun is retained. The category weights and sample scores (as in Table 10) yield the correlation coefficient of about 0.392, considerably high in comparison with the correlation obtained for the original data matrix, showing a positive correlationship between categories and samples. Consequently, the categories placed close together and the samples placed close together are considered to be qualitatively similar. Those located distant from one another are qualitatively different.

In Table 10, there is a unit distance between categories and between samples, but there will be a still higher correlation coefficient if distances between categories and distances between samples are properly adjusted as in Table 11. Here the correlation coefficient is about 0.440.

HAYASHI's Quantification Method Type III does not literally change the positions of rows and columns of original data matrix. Instead, it gives the quantities to categories and samples in the original data matrix in the way that the given quantities yield the highest correlation coefficient between the two. Table 12 shows how Extended HAYASHI's Quantification Method Type III assigns quantities to categories and samples. The correlation coefficient attained for Table 12 is exactly the same as the correlation coefficient obtained for Table 11. As a matter of fact, Tables 10 and 11 have been worked out on the basis of Table 12 just for the sake of demonstration. Without using the present method, there is no easy way to carry out the rearrangement of rows and columns with properly adjusted distances.

After simultaneous quantification of both categories and samples, we can classify them according to the given quantities. However, Table 12 contains only one of the several sets of category weights and sample scores produced by this statistical method. Table 13 shows another set of category weights

Table 10. Rearranged Data Matrix with Category Weights and Sample Scores.

No	Text ID		YOU	I	WE	IT	THEY	HE	SHE
			(1.0)	(2.0)	(3.0)	(4.0)	(5.0)	(6.0)	(7.0)
165	B 666	(1.0)	1611.9	12.5	43.3	237.3	171.1	26.2	5.7
305	B 816	(2.0)	2999.3	81.5	54.8	616.4	426.7	17.8	2.5
202	B 704	(3.0)	2108.0	217.8	108.5	550.3	246.8	23.5	30.4
171	B 672	(4.0)	2089.4	18.7	17.0	285.1	239.3	62.8	95.0
163	B 664	(5.0)	2054.8	244.9	158.4	473.5	339.9	43.3	32.4
303	B 813	(6.0)	1750.0	82.2	444.8	454.5	307.4	17.3	18.4
114	B 615	(7.0)	193.3	53.0	23.8	63.5	42.4	0.0	0.0
101	B 602	(8.0)	1942.8	98.7	11.4	652.7	159.4	83.5	34.2
155	B 656	(9.0)	922.4	59.9	46.8	375.1	170.3	13.1	2.8
264	B 769	(10.0)	3383.6	110.1	488.8	597.7	384.1	184.5	66.9
242	B 745	(11.0)	1914.6	51.7	58.2	325.6	321.3	142.3	10.8
137	B 638	(12.0)	1401.0	63.3	61.3	409.8	261.1	25.1	52.2
278	B 785	(13.0)	868.8	6.8	17.8	401.9	244.5	2.5	1.7
308	B 821	(14.0)	2555.5	6.0	12.8	399.9	1309.2	25.2	6.7
159	B 660	(15.0)	1893.3	97.9	26.4	723.5	371.6	85.8	25.3
540	S 248	(364.0)	546.5	243.6	184.3	572.8	507.0	316.0	0.0
296	B 804	(365.0)	0.0	5.2	1.7	409.1	53.0	11.3	0.9
399	S 96	(366.0)	1158.7	1334.8	865.4	699.2	581.8	640.5	298.2
310	B 823	(367.0)	1053.0	904.5	576.0	589.5	166.5	351.0	427.5
320	S 1	(368.0)	1130.8	2005.1	388.9	1318.3	513.0	623.3	427.5
408	S 105	(369.0)	40.4	242.6	27.0	721.0	175.2	94.3	0.0
260	B 765	(370.0)	976.3	322.7	308.0	755.9	653.5	364.3	194.9
504	S 208	(371.0)	582.0	582.0	582.0	844.8	488.1	262.8	225.3
547	S 255	(372.0)	1525.2	2501.8	252.1	1171.9	511.2	687.1	644.1
314	B 9003	(373.0)	61.7	128.1	17.1	212.7	94.9	45.7	17.1
519	S 226	(374.0)	1387.3	2029.8	400.4	875.2	572.6	1089.4	102.4
461	S 162	(375.0)	1256.9	1904.4	342.8	624.6	491.3	1108.3	0.0
460	S 161	(376.0)	912.7	1636.1	409.0	615.2	608.5	875.5	6.8
594	S 327	(377.0)	1290.2	1912.0	427.2	972.9	554.1	1099.8	55.0
331	S 14	(378.0)	1222.2	1720.3	311.5	1528.9	516.0	779.9	307.9
133	B 634	(714.0)	92.4	51.7	59.4	144.1	132.0	542.2	23.1
466	S 167	(715.0)	173.4	279.0	263.9	542.9	203.6	1628.7	150.8
38	B 41	(716.0)	623.2	1013.3	107.0	772.6	326.8	2268.5	1485.7
299	B 809	(717.0)	8.3	103.6	70.4	455.7	165.7	923.9	0.0
106	B 607	(718.0)	20.3	91.2	125.3	481.1	229.5	554.8	614.7
61	B 96	(719.0)	326.0	375.9	184.6	596.3	924.7	2331.4	982.4
455	S 156	(720.0)	80.4	249.2	241.2	619.1	217.1	1993.9	16.1
458	S 159	(721.0)	140.6	140.6	140.6	309.4	421.9	393.8	1237.5
138	B 639	(722.0)	5.3	5.3	810.1	354.2	399.2	1218.7	1281.8
203	B 705	(723.0)	3.7	20.0	11.6	109.4	127.3	80.5	306.2
60	B 76	(724.0)	294.2	400.4	132.7	543.0	831.7	2453.1	1328.3
317	B 9006	(725.0)	19.2	9.6	60.7	372.5	139.4	782.2	189.4
287	B 795	(726.0)	52.7	18.3	41.2	215.2	563.2	1039.5	306.8
277	B 784	(727.0)	2.9	58.8	14.7	191.0	129.3	52.9	620.1
192	B 693	(728.0)	56.4	48.4	34.3	310.3	985.3	1968.6	479.6

R =0.3920361D+00

Table 11. Rearranged Data Matrix with Adjusted Category Weights and Sample Scores

No	Text ID		YOU	I	WE	IT	THEY	HE	SHE
			(1.000)	(1.500)	(1.666)	(1.882)	(2.130)	(3.989)	(4.061)
165	B 666	(100.000)	1611.9	12.5	43.3	237.3	171.1	26.2	5.7
305	B 816	(105.477)	2999.3	81.5	54.8	616.4	426.7	17.8	2.5
202	B 704	(119.183)	2108.0	217.8	108.5	550.3	246.8	23.5	30.4
171	B 672	(125.151)	2089.4	18.7	17.0	285.1	239.3	62.8	95.0
163	B 664	(127.911)	2054.8	244.9	158.4	473.5	339.9	43.3	32.4
303	B 813	(130.700)	1750.0	82.2	444.8	454.5	307.4	17.3	18.4
114	B 615	(130.894)	193.3	53.0	23.8	63.5	42.4	0.0	0.0
101	B 602	(131.385)	1942.8	98.7	11.4	652.7	159.4	83.5	34.2
155	B 656	(132.807)	922.4	59.9	46.8	375.1	170.3	13.1	2.8
264	B 769	(133.894)	3383.6	110.1	488.8	597.7	384.1	184.5	66.9
242	B 745	(136.874)	1914.6	51.7	58.2	325.6	321.3	142.3	10.8
137	B 638	(138.816)	1401.0	63.3	61.3	409.8	261.1	25.1	52.2
278	B 785	(139.437)	868.8	6.8	17.8	401.9	244.5	2.5	1.7
308	B 821	(144.624)	2555.5	6.0	12.8	399.9	1309.2	25.2	6.7
159	B 660	(145.291)	1893.3	97.9	26.4	723.5	371.6	85.8	25.3
540	S 248	(259.908)	546.5	243.6	184.3	572.8	507.0	316.0	0.0
296	B 804	(260.193)	0.0	5.2	1.7	409.1	53.0	11.3	0.9
399	S 96	(260.276)	1158.7	1334.8	865.4	699.2	581.8	640.5	298.2
310	B 823	(260.499)	1053.0	904.5	576.0	589.5	166.5	351.0	427.5
320	S 1	(261.596)	1130.8	2005.1	388.9	1318.3	513.0	623.3	427.5
408	S 105	(261.863)	40.4	242.6	27.0	721.0	175.2	94.3	0.0
260	B 765	(262.253)	976.3	322.7	308.0	755.9	653.5	364.3	194.9
504	S 208	(262.360)	582.0	582.0	582.0	844.8	488.1	262.8	225.3
547	S 255	(262.391)	1525.2	2501.8	252.1	1171.9	511.2	687.1	644.1
314	B 9003	(262.732)	61.7	128.1	17.1	212.7	94.9	45.7	17.1
519	S 226	(263.218)	1387.3	2029.8	400.4	875.2	572.6	1089.4	102.4
461	S 162	(264.683)	1256.9	1904.4	342.8	624.6	491.3	1108.3	0.0
460	S 161	(265.058)	912.7	1636.1	409.0	615.2	608.5	875.5	6.8
594	S 327	(265.087)	1290.2	1912.0	427.2	972.9	554.1	1099.8	55.0
331	S 14	(265.748)	1222.2	1720.3	311.5	1528.9	516.0	779.9	307.9
133	B 634	(484.785)	92.4	51.7	59.4	144.1	132.0	542.2	23.1
466	S 167	(487.959)	173.4	279.0	263.9	542.9	203.6	1628.7	150.8
38	B 41	(488.962)	623.2	1013.3	107.0	772.6	326.8	2268.5	1485.7
299	B 809	(496.351)	8.3	103.6	70.4	455.7	165.7	923.9	0.0
106	B 607	(509.890)	20.3	91.2	125.3	481.1	229.5	554.8	614.7
61	B 96	(513.613)	326.0	375.9	184.6	596.3	924.7	2331.4	982.4
455	S 156	(513.862)	80.4	249.2	241.2	619.1	217.1	1993.9	16.1
458	S 159	(522.423)	140.6	140.6	140.6	309.4	421.9	393.8	1237.5
138	B 639	(537.448)	5.3	5.3	810.1	354.2	399.2	1218.7	1281.8
203	B 705	(538.062)	3.7	20.0	11.6	109.4	127.3	80.5	306.2
60	B 76	(540.394)	294.2	400.4	132.7	543.0	831.7	2453.1	1328.3
317	B 9006	(540.954)	19.2	9.6	60.7	372.5	139.4	782.2	189.4
287	B 795	(541.340)	52.7	18.3	41.2	215.2	563.2	1039.5	306.8
277	B 784	(554.638)	2.9	58.8	14.7	191.0	129.3	52.9	620.1
192	B 693	(556.869)	56.4	48.4	34.3	310.3	985.3	1968.6	479.6

R = 0.4398024D+00

Table 12. Original Data Matrix with Category Weights and Sample Scores Supplied by Quantification (Axis 1).

No	Text ID		HE	SHE	IT	THEY	YOU	WE	I
			(1.9668)	(2.0390)	(-.1398)	(0.1079)	(-1.022)	(-.3561)	(-.5215)
1	B 4	(1.67171)	1681.1	718.9	708.8	295.9	584.7	116.2	833.7
2	B 5	(1.25071)	1594.9	443.6	845.3	461.7	650.3	231.2	840.9
3	B 6	(1.12993)	979.4	1054.7	731.2	763.2	686.3	378.0	921.8
4	B 7	(1.10195)	1559.0	304.8	731.4	428.2	653.3	196.9	1010.4
5	B 8	(1.33630)	1394.4	837.1	554.6	399.1	739.2	188.1	1059.9
6	B 9	(1.64941)	1738.1	891.4	637.5	433.0	600.5	240.8	962.5
7	B 10	(0.17582)	630.9	546.0	637.2	313.5	613.2	352.8	2183.4
8	B 11	(1.40477)	1492.8	773.4	689.2	451.8	642.4	219.9	959.2
9	B 12	(2.08055)	1337.7	1467.6	609.5	500.6	475.5	167.6	609.5
10	B 13	(1.90941)	1700.9	1074.2	766.4	452.7	531.1	125.5	723.1
11	B 14	(1.42082)	1294.5	1851.7	928.9	281.2	1067.7	144.9	1395.4
12	B 15	(1.53655)	1379.5	1743.9	788.8	444.7	876.2	283.2	1238.3
13	B 16	(1.38231)	1694.0	583.2	747.0	632.2	692.1	262.5	779.9
14	B 17	(1.32575)	935.5	1156.5	859.7	659.4	606.2	304.9	710.0
15	B 18	(0.42515)	851.2	608.1	677.9	403.9	591.2	282.0	2145.3
364	S 53	(-.98267)	114.4	59.5	1194.8	576.8	1643.4	398.3	1611.4
365	S 54	(-.91411)	191.4	0.0	1252.2	657.3	1501.8	545.0	1768.1
366	S 55	(-.62339)	447.5	98.0	993.1	724.6	1491.8	422.0	1981.9
367	S 56	(-.66559)	314.9	207.2	1222.4	505.5	1404.7	430.9	2328.7
368	S 57	(-.85521)	224.2	77.8	1084.4	654.3	1592.3	329.4	1821.1
369	S 58	(-.54195)	171.5	285.9	857.7	1108.4	1196.4	285.9	1860.5
370	S 59	(-.84168)	242.8	34.0	961.7	553.7	1282.2	602.3	2039.9
371	S 60	(-.72085)	357.1	12.3	1026.3	500.8	1153.5	324.3	2023.8
372	S 61	(-.74236)	192.3	259.3	871.9	778.0	1560.5	478.4	2338.5
373	S 62	(-.70215)	228.7	65.3	1306.7	702.4	1074.0	485.9	1874.3
374	S 63	(-.63970)	188.6	317.0	1183.7	501.6	1468.6	417.3	1681.2
375	S 64	(-.93750)	272.3	46.8	919.1	523.4	1795.6	485.1	1999.8
376	S 65	(-.47574)	391.8	195.9	779.2	797.0	1139.8	734.6	1812.1
377	S 66	(-.55784)	160.4	270.1	995.9	666.7	1021.2	460.0	1696.4
378	S 67	(-.25455)	574.1	181.7	875.6	607.2	974.8	380.0	1743.0
714	T 27791	(1.34935)	585.0	154.9	430.1	303.1	115.1	160.4	203.2
715	T 28491	(1.38652)	714.5	203.6	504.7	365.6	133.1	169.9	280.7
716	T 28591	(1.68317)	554.3	101.6	350.5	256.6	38.6	94.5	134.3
717	T 28691	(1.42419)	560.3	65.7	410.0	304.4	63.8	116.6	132.6
718	T 28791	(1.36985)	652.7	174.6	484.3	367.7	106.0	148.4	254.1
719	T 29491	(1.44145)	573.4	89.2	358.7	424.7	42.8	156.4	141.0
720	T 29591	(1.37983)	468.0	81.9	394.5	307.3	40.9	95.5	130.5
721	T 29691	(1.38114)	641.7	248.5	442.3	330.0	133.2	175.5	307.0
722	T 29791	(1.40993)	524.5	89.2	385.8	320.8	59.4	136.5	126.6
723	T 30491	(1.44106)	497.9	76.4	407.7	254.3	43.8	119.6	113.4
724	T 30591	(1.48840)	599.0	77.5	380.6	301.4	55.7	146.3	157.3
725	T 30691	(1.27017)	565.0	198.3	510.2	364.7	101.4	186.9	219.6
726	T 30791	(1.33638)	449.8	104.3	372.0	297.4	66.0	111.1	132.6
727	T 31591	(1.59757)	532.5	189.2	389.6	300.3	67.5	83.4	175.3
728	T 31791	(1.37421)	486.0	137.9	402.0	302.1	68.0	133.6	152.6

R = 0.4398024D+00

Table 13. Original Data Matrix with Category Weights and Sample Scores Supplied by Quantification (Axis 2).

No	Text ID		HE	SHE	IT	THEY	YOU	WE	I
			(-.1626)	(1.4710)	(-.7166)	(-1.007)	(-.0155)	(-1.481)	(1.4538)
1	B 4	(0.7477)	1681.1	718.9	708.8	295.9	584.7	116.2	833.7
2	B 5	(0.1391)	1594.9	443.6	845.3	461.7	650.3	231.2	840.9
3	B 6	(0.5769)	979.4	1054.7	731.2	763.2	686.3	378.0	921.8
4	B 7	(0.3047)	1559.0	304.8	731.4	428.2	653.3	196.9	1010.4
5	B 8	(1.0303)	1394.4	837.1	554.6	399.1	739.2	188.1	1059.9
6	B 9	(0.7774)	1738.1	891.4	637.5	433.0	600.5	240.8	962.5
7	B 10	(1.7828)	630.9	546.0	637.2	313.5	613.2	352.8	2183.4
8	B 11	(0.7033)	1492.8	773.4	689.2	451.8	642.4	219.9	959.2
9	B 12	(1.1550)	1337.7	1467.6	609.5	500.6	475.5	167.6	609.5
10	B 13	(0.7870)	1700.9	1074.2	766.4	452.7	531.1	125.5	723.1
11	B 14	(1.7669)	1294.5	1851.7	928.9	281.2	1067.7	144.9	1395.4
12	B 15	(1.4602)	1379.5	1743.9	788.8	444.7	876.2	283.2	1238.3
13	B 16	(0.0983)	1694.0	583.2	747.0	632.2	692.1	262.5	779.9
14	B 17	(0.5876)	935.5	1156.5	859.7	659.4	606.2	304.9	710.0
15	B 18	(1.6824)	851.2	608.1	677.9	403.9	591.2	282.0	2145.3
364	S 53	(0.2349)	114.4	59.5	1194.8	576.8	1643.4	398.3	1611.4
365	S 54	(0.0927)	191.4	0.0	1252.2	657.3	1501.8	545.0	1768.1
366	S 55	(0.5131)	447.5	98.0	993.1	724.6	1491.8	422.0	1981.9
367	S 56	(0.9096)	314.9	207.2	1222.4	505.5	1404.7	430.9	2328.7
368	S 57	(0.4916)	224.2	77.8	1084.4	654.3	1592.3	329.4	1821.1
369	S 58	(0.5868)	171.5	285.9	857.7	1108.4	1196.4	285.9	1860.5
370	S 59	(0.5236)	242.8	34.0	961.7	553.7	1282.2	602.3	2039.9
371	S 60	(0.7894)	357.1	12.3	1026.3	500.8	1153.5	324.3	2023.8
372	S 61	(0.9089)	192.3	259.3	871.9	778.0	1560.5	478.4	2338.5
373	S 62	(0.2576)	228.7	65.3	1306.7	702.4	1074.0	485.9	1874.3
374	S 63	(0.5630)	188.6	317.0	1183.7	501.6	1468.6	417.3	1681.2
375	S 64	(0.6059)	272.3	46.8	919.1	523.4	1795.6	485.1	1999.8
376	S 65	(0.2455)	391.8	195.9	779.2	797.0	1139.8	734.6	1812.1
377	S 66	(0.5245)	160.4	270.1	995.9	666.7	1021.2	460.0	1696.4
378	S 67	(0.6112)	574.1	181.7	875.6	607.2	974.8	380.0	1743.0
714	T 27791	(-0.7962)	585.0	154.9	430.1	303.1	115.1	160.4	203.2
715	T 28491	(-0.6052)	714.5	203.6	504.7	365.6	133.1	169.9	280.7
716	T 28591	(-0.9462)	554.3	101.6	350.5	256.6	38.6	94.5	134.3
717	T 28691	(-1.2747)	560.3	65.7	410.0	304.4	63.8	116.6	132.6
718	T 28791	(-0.7005)	652.7	174.6	484.3	367.7	106.0	148.4	254.1
719	T 29491	(-1.3813)	573.4	89.2	358.7	424.7	42.8	156.4	141.0
720	T 29591	(-1.2055)	468.0	81.9	394.5	307.3	40.9	95.5	130.5
721	T 29691	(-0.3274)	641.7	248.5	442.3	330.0	133.2	175.5	307.0
722	T 29791	(-1.2761)	524.5	89.2	385.8	320.8	59.4	136.5	126.6
723	T 30491	(-1.2816)	497.9	76.4	407.7	254.3	43.8	119.6	113.4
724	T 30591	(-1.1688)	599.0	77.5	380.6	301.4	55.7	146.3	157.3
725	T 30691	(-0.8395)	565.0	198.3	510.2	364.7	101.4	186.9	219.6
726	T 30791	(-1.0945)	449.8	104.3	372.0	297.4	66.0	111.1	132.6
727	T 31591	(-0.5468)	532.5	189.2	389.6	300.3	67.5	83.4	175.3
728	T 31791	(-0.9695)	486.0	137.9	402.0	302.1	68.0	133.6	152.6

R＝0.2732325D＋00

Table 14. Original Data Matrix with Category Weights and Sample Scores Supplied by Quantification (Axis 3).

No	Text ID		HE	SHE	IT	THEY	YOU	WE	I
			(-.0235)	(1.5862)	(-.3881)	(-.0712)	(1.6217)	(-1.130)	(-.9772)
1	B 4	(0.6369)	1681.1	718.9	708.8	295.9	584.7	116.2	833.7
2	B 5	(0.2131)	1594.9	443.6	845.3	461.7	650.3	231.2	840.9
3	B 6	(0.7755)	979.4	1054.7	731.2	763.2	686.3	378.0	921.8
4	B 7	(-0.0143)	1559.0	304.8	731.4	428.2	653.3	196.9	1010.4
5	B 8	(0.7552)	1394.4	837.1	554.6	399.1	739.2	188.1	1059.9
6	B 9	(0.6064)	1738.1	891.4	637.5	433.0	600.5	240.8	962.5
7	B 10	(-0.7064)	630.9	546.0	637.2	313.5	613.2	352.8	2183.4
8	B 11	(0.5577)	1492.8	773.4	689.2	451.8	642.4	219.9	959.2
9	B 12	(1.5165)	1337.7	1467.6	609.5	500.6	475.5	167.6	609.5
10	B 13	(0.9773)	1700.9	1074.2	766.4	452.7	531.1	125.5	723.1
11	B 14	(1.5284)	1294.5	1851.7	928.9	281.2	1067.7	144.9	1395.4
12	B 15	(1.3198)	1379.5	1743.9	788.8	444.7	876.2	283.2	1238.3
13	B 16	(0.4441)	1694.0	583.2	747.0	632.2	692.1	262.5	779.9
14	B 17	(1.0257)	935.5	1156.5	859.7	659.4	606.2	304.9	710.0
15	B 18	(-0.5634)	851.2	608.1	677.9	403.9	591.2	282.0	2145.3
364	S 53	(0.1584)	114.4	59.5	1194.8	576.8	1643.4	398.3	1611.4
365	S 54	(-0.2934)	191.4	0.0	1252.2	657.3	1501.8	545.0	1768.1
366	S 55	(-0.1812)	447.5	98.0	993.1	724.6	1491.8	422.0	1981.9
367	S 56	(-0.4094)	314.9	207.2	1222.4	505.5	1404.7	430.9	2328.7
368	S 57	(0.0547)	224.2	77.8	1084.4	654.3	1592.3	329.4	1821.1
369	S 58	(-0.1104)	171.5	285.9	857.7	1108.4	1196.4	285.9	1860.5
370	S 59	(-0.6539)	242.8	34.0	961.7	553.7	1282.2	602.3	2039.9
371	S 60	(-0.6472)	357.1	12.3	1026.3	500.8	1153.5	324.3	2023.8
372	S 61	(-0.1697)	192.3	259.3	871.9	778.0	1560.5	478.4	2338.5
373	S 62	(-0.7460)	228.7	65.3	1306.7	702.4	1074.0	485.9	1874.3
374	S 63	(0.1831)	188.6	317.0	1183.7	501.6	1468.6	417.3	1681.2
375	S 64	(0.0539)	272.3	46.8	919.1	523.4	1795.6	485.1	1999.8
376	S 65	(-0.5397)	391.8	195.9	779.2	797.0	1139.8	734.6	1812.1
377	S 66	(-0.3925)	160.4	270.1	995.9	666.7	1021.2	460.0	1696.4
378	S 67	(-0.4822)	574.1	181.7	875.6	607.2	974.8	380.0	1743.0
714	T 27791	(-0.2987)	585.0	154.9	430.1	303.1	115.1	160.4	203.2
715	T 28491	(-0.2731)	714.5	203.6	504.7	365.6	133.1	169.9	280.7
716	T 28591	(-0.4623)	554.3	101.6	350.5	256.6	38.6	94.5	134.3
717	T 28691	(-0.5837)	560.3	65.7	410.0	304.4	63.8	116.6	132.6
718	T 28791	(-0.3503)	652.7	174.6	484.3	367.7	106.0	148.4	254.1
719	T 29491	(-0.6251)	573.4	89.2	358.7	424.7	42.8	156.4	141.0
720	T 29591	(-0.5780)	468.0	81.9	394.5	307.3	40.9	95.5	130.5
721	T 29691	(-0.1682)	641.7	248.5	442.3	330.0	133.2	175.5	307.0
722	T 29791	(-0.5339)	524.5	89.2	385.8	320.8	59.4	136.5	126.6
723	T 30491	(-0.6226)	497.9	76.4	407.7	254.3	43.8	119.6	113.4
724	T 30591	(-0.6557)	599.0	77.5	380.6	301.4	55.7	146.3	157.3
725	T 30691	(-0.3343)	565.0	198.3	510.2	364.7	101.4	186.9	219.6
726	T 30791	(-0.4037)	449.8	104.3	372.0	297.4	66.0	111.1	132.6
727	T 31591	(-0.0921)	532.5	189.2	389.6	300.3	67.5	83.4	175.3
728	T 31791	(-0.3707)	486.0	137.9	402.0	302.1	68.0	133.6	152.6

R =0.2565148D+00

and sample scores with a correlation coefficient of about 0.273. Table 14 has yet another set of category weights and sample scores, yielding a lower correlation coefficient than those listed in the previous two tables. In general, if there are *n* categories and more than *n* samples, HAYASHI's Quantification Method Type III can produce *n*-1 sets (or "axes," to use a statistical term) of category weights and sample scores. So, in the present case, there are six ways to assign numerical values to categories and samples, each of which yields a correlation coefficient lower than that of the set produced in the previous stage.

Table 15 indicates that correlation coefficients for the first three axes are 0.44, 0.27 and 0.26. The seemingly small correlation coefficients are actually quite significant, considering that there are 2,619,476 tokens of pronouns involved in the calculation. "Proportion Accounted for," associated with each correlation coefficient, is the amount of information contained in the original data matrix explained by the axis in question. "Cumulative Proportion" indicates the percentage of the information in the original data explained up to the axis in question.

Ordinarily, not all sets of category weights and sample scores are used for

Table 15. Correlationship Coefficients, Proportion Accounted for, and Cumulative Proportion Accounted for for Each Axis.

Axis	Correlationship Coefficients	Proportion Accounted for	Cumulative Proportion
1	0.4398024D+00	0.4470188D+02	0.4470188D+02
2	0.2732325D+00	0.1725343D+02	0.6195530D+02
3	0.2565148D+00	0.1520671D+02	0.7716201D+02
4	0.2178696D+00	0.1096993D+02	0.8813194D+02
5	0.1824095D+00	0.7689627D+01	0.9582157D+02
6	0.1344625D+00	0.4178430D+01	0.1000000D+03

Table 16. Normalised Category Weights Given to Pronouns.

Category	Axis 1	Axis 2	Axis 3
HE	0.1966771D+01	-.1625917D+00	-.23460690−01
SHE	0.2038963D+01	0.1470989D+01	0.1586175D+01
IT	-.1397855D+00	-.7165918D+00	-.3880612D+00
THEY	0.1079346D+00	-.1007389D+01	-.7124307D−01
YOU	-.1021682D+01	-.1548340D−01	0.1621680D+01
WE	-.3561173D+00	-.1480572D+01	-.1129848D+01
I	-.5215427D+00	0.1453846D+01	-.9771822D+00

analysing given data, because after the first several axes, the proportion accounted for usually becomes small and the rest of the axes can be discarded. In the present case, up to the third axis, "Cumulative Proportion Accounted for" becomes 77%, still leaving unaccounted about 23% of the information contained in the frequency table. This is natural because when the number of samples or the number of categories is large (as in the present case of 728 samples), it is difficult to observe a clear tendency. Although Axis 4 may be significant, only three axes are included in the following analysis, mainly because of the practical consideration that the figures presented later can deal with a maximum of three axes if they are to be grasped easily.

The three sets of category weights and sample scores calculated for producing the first three correlation coefficients are then normalised, with the means equal to 0, and the variances equal to 1.0. Normalised category weights and some of the sample scores thus calculated are given in Table 16 and Table 17.

Table 17. Normalized Sample Scores.

No	Text ID	Axis 1	Axis 2	Axis 3
1	B 4	0.1671706D+01	0.7476942D+00	0.6368700D+00
2	B 5	0.1250713D+01	0.1390707D+00	0.2131021D+00
3	B 6	0.1129926D+01	0.5769061D+00	0.7755322D+00
4	B 7	0.1101951D+01	0.3046798D+00	-.1431608D-01
5	B 8	0.1336304D+01	0.1030321D+01	0.7551526D+00
6	B 9	0.1649405D+01	0.7774227D+00	0.6064387D+00
7	B 10	0.1758155D+00	0.1782830D+01	-.7064183D+00
8	B 11	0.1404772D+01	0.7033367D+00	0.5576722D+00
9	B 12	0.2080547D+01	0.1154993D+01	0.1516522D+01
10	B 13	0.1909413D+01	0.7869950D+00	0.9772967D+00
11	B 14	0.1420820D+01	0.1766901D+01	0.1528414D+01
12	B 15	0.1536553D+01	0.1460201D+01	0.1319844D+01
13	B 16	0.1382311D+01	0.9833750D-01	0.4441311D+00
14	B 17	0.1325749D+01	0.5876216D+00	0.1025750D+01
15	B 18	0.4251521D+00	0.1682359D+01	-.5633707D+00
364	S 53	-.9826708D+00	0.2348588D+00	0.1583998D+00
365	S 54	-.9141112D+00	0.9265133D-01	-.2934221D+00
366	S 55	-.6233939D+00	0.5130976D+00	-.1812096D+00
367	S 56	-.6655867D+00	0.9096036D+00	-.4094200D+00
368	S 57	-.8552101D+00	0.4916342D+00	0.5469141D-01
369	S 58	-.5419510D+00	0.5868184D+00	-.1104138D+00
370	S 59	-.8416771D+00	0.5235725D+00	-.6539027D+00
371	S 60	-.7208538D+00	0.7894480D+00	-.6471856D+00
372	S 61	-.7423566D+00	0.9088861D+00	-.1696928D+00
373	S 62	-.7021492D+00	0.2576381D+00	-.7459557D+00
374	S 63	-.6397027D+00	0.5630459D+00	0.1831411D+00
375	S 64	-.9374996D+00	0.6058675D+00	0.5389464D-01
376	S 65	-.4757370D+00	0.2455155D+00	-.5397419D+00
377	S 66	-.5578376D+00	0.5245104D+00	-.3924945D+00
378	S 67	-.2545523D+00	0.6112347D+00	-.4822063D+00
714	T 27791	0.1349354D+01	-.7962455D+00	-.2987230D+00
715	T 28491	0.1386524D+01	-.6052462D+00	-.2730577D+00
716	T 28591	0.1683171D+01	-.9462120D+00	-.4622777D+00
717	T 28691	0.1424190D+01	-.1274695D+01	-.5836737D+00
718	T 28791	0.1369855D+01	-.7005071D+00	-.3502858D+00
719	T 29491	0.1441450D+01	-.1381343D+01	-.6251231D+00
720	T 29591	0.1379831D+01	-.1205539D+01	-.5780207D+00
721	T 29691	0.1381136D+01	-.3274357D+00	-.1681889D+00
722	T 29791	0.1409928D+01	-.1276064D+01	-.5338625D+00
723	T 30491	0.1441064D+01	-.1281610D+01	-.6225939D+00
724	T 30591	0.1488398D+01	-.1168806D+01	-.6556686D+00
725	T 30691	0.1270174D+01	-.8395467D+00	-.3342720D+00
726	T 30791	0.1336383D+01	-.1094549D+01	-.4036862D+00
727	T 31591	0.1597570D+01	-.5467847D+00	-.9209008D-01
728	T 31791	0.1374208D+01	-.9695278D+00	-.3707063D+00

2.2. Plotting the Quantities along Each Axis

2.2.1. Category Weights Given to Pronouns

Although these numerical values indicate the relative positions of each category and each sample along three axes, it is not so easy to grasp relationships at a glance. Mere numerical values are always difficult to process. Therefore, the values given in these tables are next plotted in figures like Fig. 2, showing the relative position of each category along the three axes.[7]

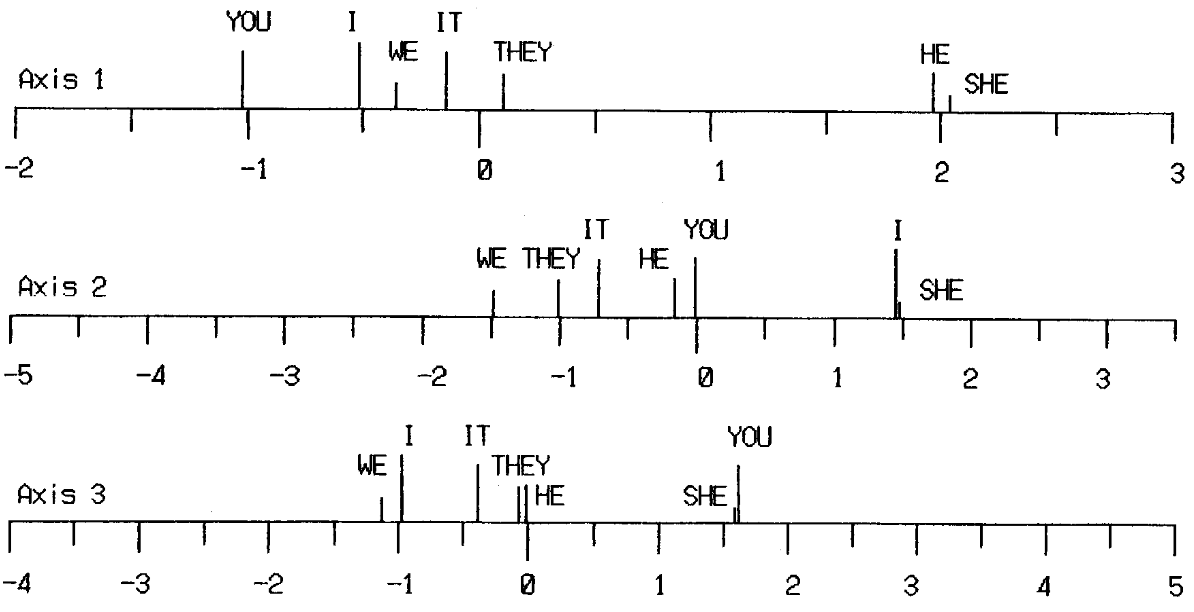

Fig. 2. Distribution of Pronominal Categories along Each Axis.

However, these figures give two different kinds of information: 1) the horizontal positions of categories showing the category weights calculated by the present method, and 2) the length of the vertical line for each pronominal category showing the ratio of the number of occurrences of the category in question to the total number of pronominal forms. As quantities are normalised, categories with their relative amounts indicated by the

[7] In this figure, the range covered by each axis is made different, reflecting the range covered by the same axis for the sample distribution. So, one could directly compare category distribution and sample distribution for each axis.

vertical lines balance at the original point of the axis. By referring to these figures, relationships among categories themselves are obvious.

These figures show the pronouns neatly grouped along Axis 1. Third person singular masculine and feminine pronouns, "HE" and "SHE," are located around 2.0 in the positive region while second person and first person pronouns, "YOU" and "I," are located in the negative region of the axis. "YOU" is the extreme, with a quantity close to -1.0, followed by "I" with the quantity around -0.5. "WE" with a quantity of -0.36 is between "I" and the origin of the axis. Third person plural and third person singular neuter pronouns, "THEY" and "IT," appear around the origin, indicating more or less even use across 728 texts of the corpus.

Along Axis 2 are three groups of pronouns. One group made up of "I" and "SHE" is located around 1.5 in the positive region and, in contrast, another group of "WE," "THEY" and "IT" is in the negative range of the axis. "YOU" and "HE" are very close to the origin, indicating that they are neutral on this axis. Actually, the quantity given to YOU is very close to 0.

Three groups appear in the case of Axis 3, too. "YOU" and "SHE" form a group close to 1.5 in the positive region. "WE" and "I" form another group in the negative range around -1.0. The rest of the categories, "IT," "THEY" and "HE," are close together around the origin of the axis. Values of "HE" and "THEY," especially "HE," are very close to 0.

On the whole, "HE" and "SHE" contrast with "YOU" along Axis 1, "I" and "SHE" contrast with "WE" along Axis 2 and "YOU" and "SHE" contrast with "WE" and "I" along Axis 3. "IT" and "THEY" assume neutral positions along Axis 1, "YOU" and "HE" along Axis 2, and "THEY" and "HE" along Axis 3. "THEY" and "IT" seem to be the most neutral of all pronominal categories.

2.2.2. Sample Scores Given to Texts

Table 17 provides sample scores for some of the 728 texts, but the scores given to all the texts are plotted along each axis as in the top part of Figs. 3, 4 and 5 labeled "Total." In these figures, as in the case of category weights, horizontal positions of samples are determined by given sample scores. The vertical length for each sample represents the ratio of occurrences of pronouns in that particular sample to total number of pronouns in the entire corpus. Since all 728 text samples are plotted along the axes, there are many overlaps, making the figures difficult to grasp. (There is a program by which we can pick out a particular range of the axis and enlarge that part in order to identify the samples, but the enlarged versions of these figures are omitted here because they require much space and the overall tendency is clear from these figures.)

In addition to "Total" showing distribution of all the samples, additional figures extracting all the samples in the Book Corpus, Spoken Corpus and Times Corpus from the total distribution are prepared as shown in the lower part of Figs. 3, 4 and 5.

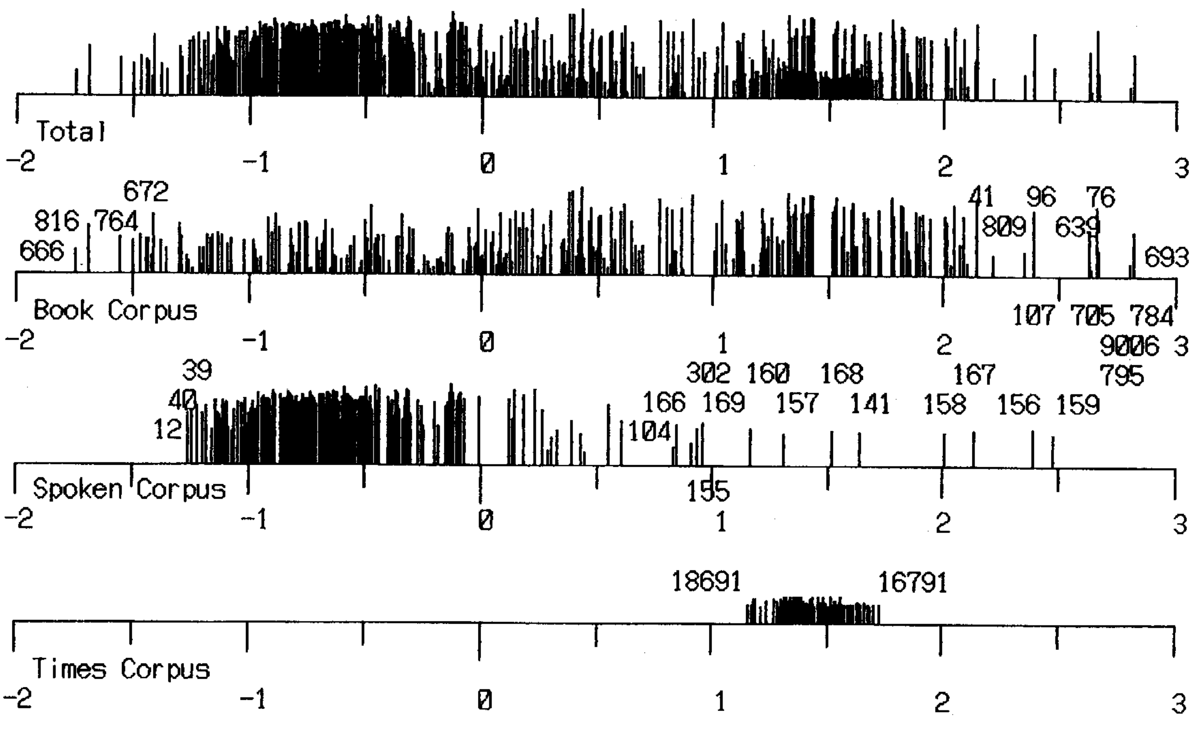

Fig. 3. Distribution of Textual Samples along Axis 1.

The top part of Fig. 3 (Total) shows the total distribution of texts along Axis 1, revealing text samples to be scattered all along this axis and varied according to the relative number of pronouns used in the texts. The rest of Fig. 3 (Book Corpus, Spoken Corpus and Times Corpus) shows the distribution of texts in an individual corpus. The tendency of dispersion of texts and variations of the amount of pronominal forms are mainly ascribed to the Book Corpus. Some of the outstanding samples in the Book Corpus are B693, B784, B795, B9006, B76, B705, and B639 in the positive range and B666 and B816 in the negative range.[8]

In the cases of the Spoken Corpus and the Times Corpus, a clear pattern specific to them is observed. For example, text samples of the Spoken Corpus are most concentrated in the range of around -1.2 to -0.1 and are characterised by the high and relatively constant ratio of the use of pronouns. Text samples scattered around the positive range of the axis, such as S159, S156, S167, and S158, are mainly from the university lectures.

Much clearer patterns can be observed in the case of the Times Corpus, where all of the text samples are distributed in the amazingly restricted range of around 1.15 to 1.75. Text samples pertaining to the Times Corpus are also characterised by the low and constant ratio of the use of pronouns. Newspaper English has a definite tendency toward the use of pronominal forms, both in the categories and the amounts.

Distribution of texts along Axis 2 is shown in Fig. 4. Again, the texts in the Book Corpus are dispersed over a wide range, and the texts in the Times Corpus are located in a very narrow range. Texts in the Spoken Corpus are between these two. In the case of the Book Corpus, there is a high frequency of pronouns in the positive region and a low frequency of pronouns in the negative region. However, there are exceptions. B784 has a low frequency of pronouns at the positive extremity; B735, B782 and B568 have a relatively high frequency of pronouns at the negative extremity. Texts in the Spoken Corpus are most concentrated in the area between the origin and 1.0. All texts in the Times Corpus are in the narrow band around -1.0.

Along Axis 3, the most concentrated area is in the negative region close to the origin of the axis. Axis 3 is similar to Axes 1 and 2 in so far as texts

[8] The capital letters "B," "S," and "T" placed at the beginning of text identifiers (indicating the Book Corpus, the Spoken Corpus and the Times Corpus respectively) are omitted in the figures below when the figures themselves are labelled as such.

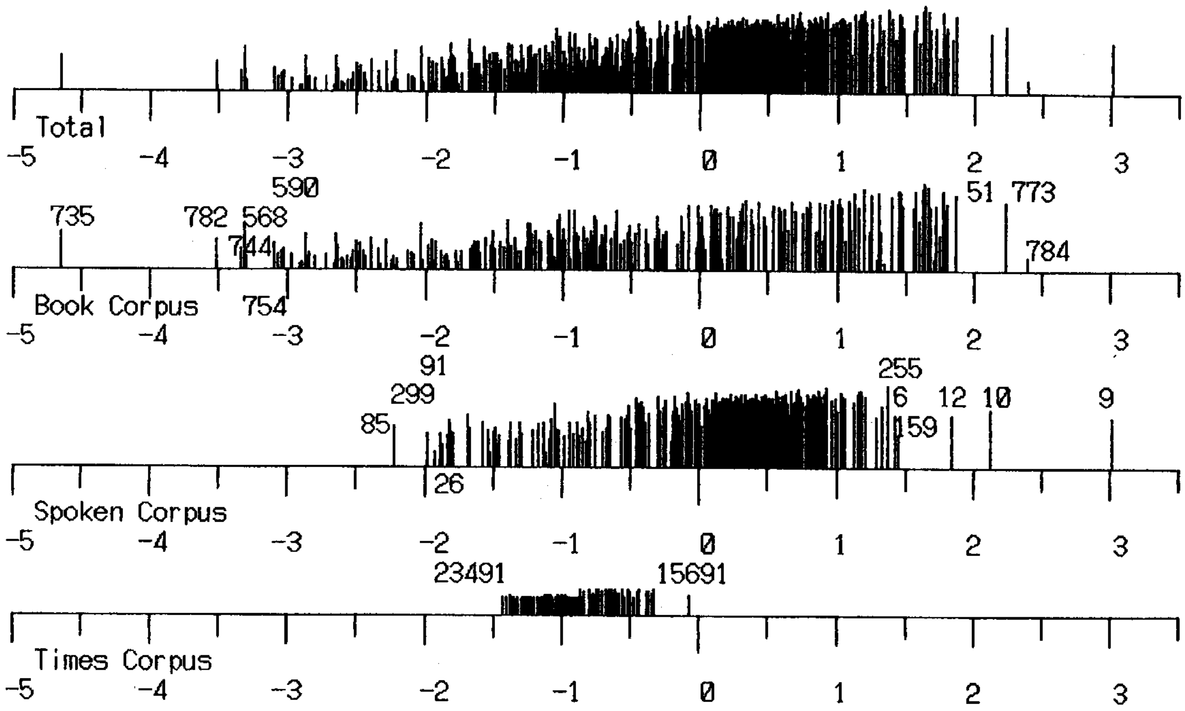

Fig. 4. Distribution of Textual Samples along Axis 2.

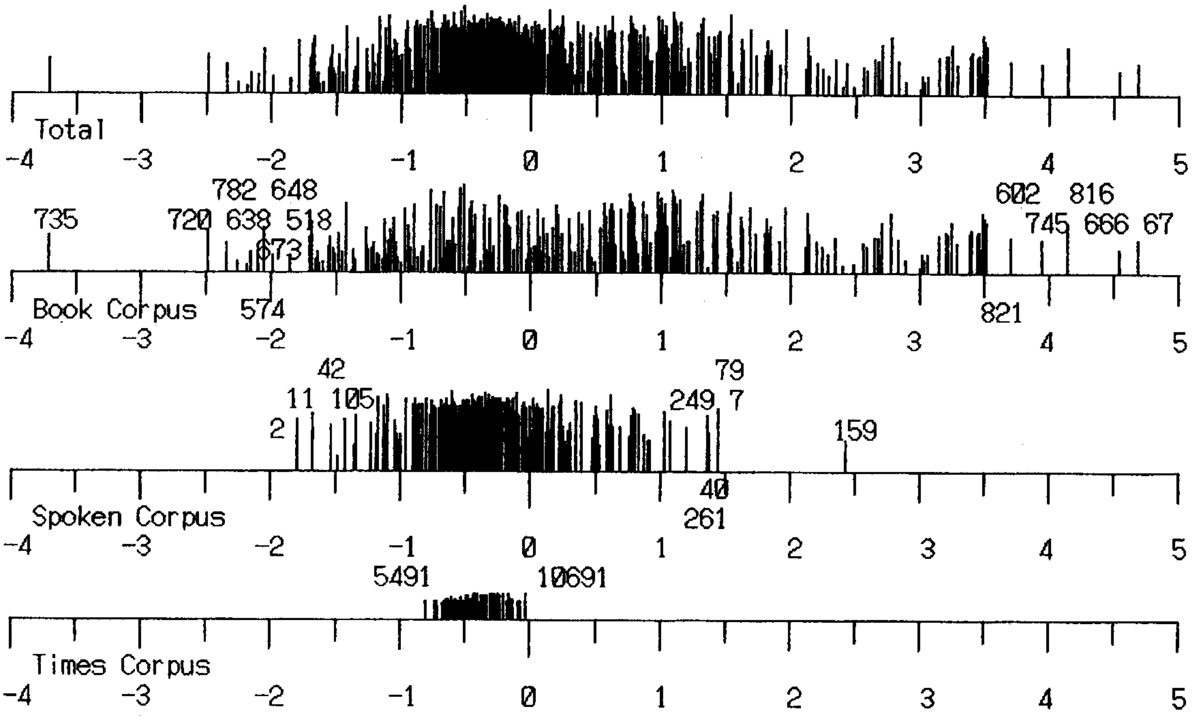

Fig. 5. Distribution of Textual Samples along Axis 3.

in the Book Corpus are spread over a wide range, texts in the Times Corpus over a narrow band and texts in the Spoken Corpus are in between.

There is a definite tendency of dispersion along three axes as far as the texts in the Book Corpus are concerned. The reason may be that the books cover a great variety of genres including fiction, biography, business, hobbies, cooking, health and fitness, astrology and occult sciences. Consequently, the perspectives assumed and the amount of pronouns used in these books may vary greatly.

In contrast, the amounts of pronouns used in texts of the Times Corpus are small and constant. Thus, newspaper English is definitely characterised by few pronouns, especially in comparison with texts in the Spoken Corpus. The Times Corpus is also characterised by the distribution of texts over a narrow range along three axes. Axes 2 and 3 show that text distribution is close to the origin. It may be that certain topics require the use of certain pronouns, but when consecutive issues of a general periodical are collected, there is a tendency to neutralise differences in the use of specific pronominal forms. However, Axis 1 still retains uniqueness as described below.

The amount of pronouns used in texts of the Spoken Corpus is very large and level of concentration is between the Book Corpus and the Times Corpus. The most concentrated area is easy to locate along three axes. However, some texts in this corpus tend to spread over a wide range, especially along Axis 1, but they are chiefly university lectures.

2.2.3. Interpretation of the Axes

Figs. 2, 3, 4 and 5 outline relationships among pronominal categories and relationships among textual samples, and there is a clear correspondence between the figures showing the category distribution and figures showing the sample distribution. The reason why a particular pronominal category assumes a certain position in Fig. 2 can be explained by referring to the sample distributions of Figs. 3, 4 and 5. Conversely, the reason why a particular sample assumes certain positions in Figs. 3, 4 and 5 can be explained by referring to the category distribution in Fig. 2. Thus, plotting the categories and the samples along these axes will eventually make the interpretations of the quantities given to them a relatively easy and simple matter.

For example, as shown in Fig. 3, text samples of the Spoken Corpus are most concentrated in the negative range between around -1.2 to -0.1 along Axis 1. This area coincides with the area of "YOU," "I" and "WE" in the category distribution along Axis 1 in Fig. 2.[9] So, most texts in the Spoken Corpus are oriented toward the first person or the second person pronouns. In contrast, text samples in the Spoken Corpus scattered around the positive range of the axis such as S159, S156, S167, and S158, are mainly from the university lectures, characterised by the high frequency of "HE" and "SHE," indicating the tendency toward personal detachment in the academic lectures.

Quantities are given along these axes so that categories and samples may be grouped or classified, but often given quantities indicate the reasons for certain classification. Thus, because of the distribution of categories and samples along the same axis, it is often possible to explain why they are so distributed or arranged along that axis. In other words, an axis can sometimes be given a straightforward interpretation explaining why categories and samples are so arranged.

In this study, Axis 1 places "SHE" and "HE" at the far end of the positive range. In contrast, "YOU," "I" and "WE" are placed in the opposite negative range. Axis 1 can thus be interpreted as a measure of the degree of personal detachment or personal involvement, a fact confirmed either by the sample distribution along this axis or by superimposing Fig. 2 onto Fig. 3. To repeat, "YOU" and "I" are important pronouns that contribute greatly to the distributions of texts in the Spoken Corpus. The most concentrated area of this corpus corresponds to the area where YOU and I are located in the category distribution. Therefore, many texts in the Spoken Corpus are examples of texts with a high degree of personal involvement. However, some texts in the Spoken Corpus tend to spread toward "HE" and "SHE" along Axis 1, indicating personal detachment.

The distribution of categories and samples along Axis 1, whose "Propor-

[9] This is clearly seen when Fig. 2 is superimposed onto this figure. Precisely for this reason, the ranges of the three axes in Fig. 2 were determined as they were.

tion Accounted for" is about 45%, is determined by writer's/speaker's intention to involve personally the reader/hearer in what he is writing or speaking about. Personal involvement or detachment is the determining factor in the distribution of pronouns and texts.

As mentioned above, all text samples pertaining to the Times Corpus are distributed in an amazingly restricted range along Axis 1. Clearly, personal detachment is an important factor in newspaper English, causing the textual samples close to "HE" and "SHE."

In contrast to Axis 1, it is hard to interpret Axes 2 and 3. Interpretation of these two axes is not at all clear at present, mainly because of the behaviour of "SHE," which is grouped with "I" along Axis 2 and with "YOU" along Axis 3. To interpret these axes, it is necessary to examine the distribution of texts in greater detail or to examine the content of texts to see how different contexts determine the use of different pronouns. Such examination is practically impossible considering the amount of texts involved.

2.3. Plotting the Quantities in a Three-Dimensional Space

Plotting quantities for categories and samples along three axes (Fig. 2 and Figs. 3, 4 and 5) facilitates interpretation of the given quantities. An additional way to enhance the treatment of these quantities is to choose either the first two or three axes for analysis (considering how 2 or 3 axes account for cumulative proportion) and to collapse or combine these axes to draw two-dimensional or three-dimensional figures.

The quantities in Table 16, plotted in a three-dimensional space (Fig. 6) can be thought of as an assemblage of the three axes in Fig 2. The relative amounts for categories in the corpus are indicated by the volume of a ball-shaped object placed at the end of the vertical line which shows the quantity for Axis 3. The pronominal structure of the entire corpus appears in this figure.

As in the case of category distribution, all samples are plotted as in Fig. 7 in a three-dimensional space according to the quantities in Table 17. The relative amount of pronouns used in a particular sample is indicated by the size of a ball-shaped object placed at the end of the vertical line which

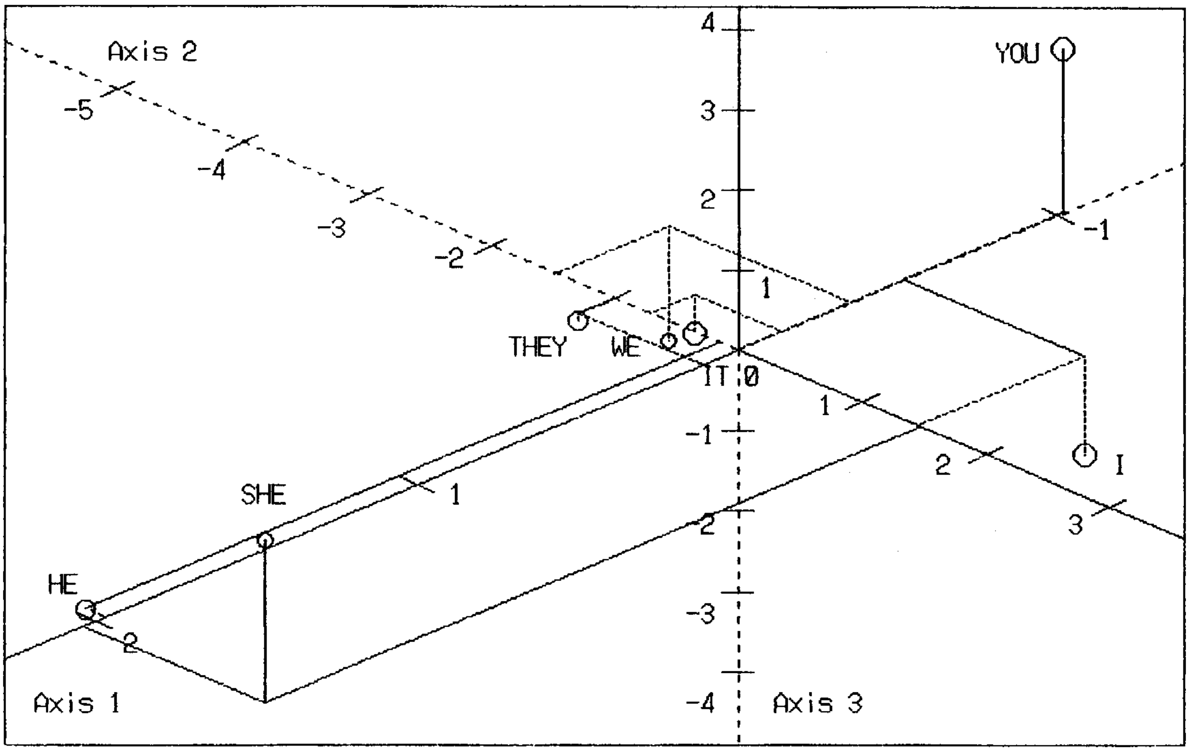

Fig. 6. Three-Dimensional Distribution of Pronominal Categories.

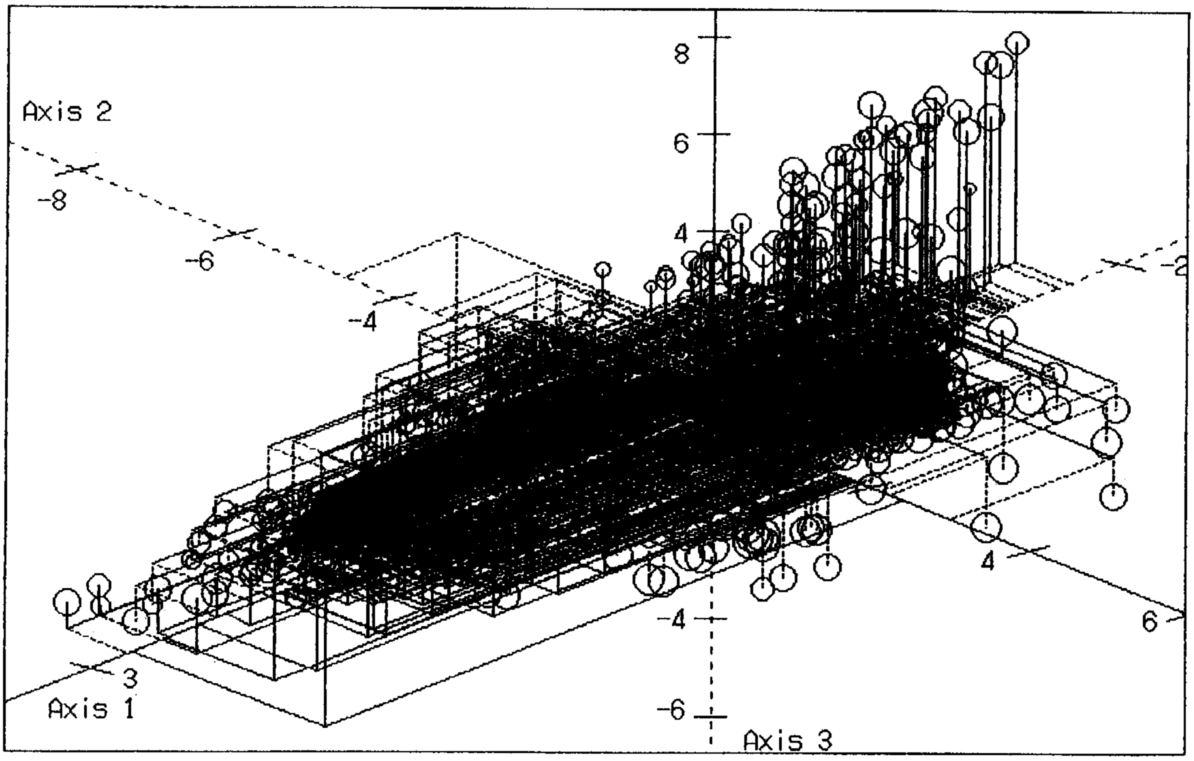

Fig. 7. Three-Dimensional Distribution of Textual Samples.

indicates the quantity given for the third axis.

When numerical values are transformed into a point in a 3-dimensional space, it is easy to grasp visually the relationships binding categories and samples. In Figs. 6 and 7, categories sharing similar tendencies across samples or samples showing similar distribution across categories either have the same position or are close together. If a particular category is distributed evenly across samples, it will be placed at the origin of coordinates, showing that the distribution of this category is not influenced by samples. Similarly, if a particular sample exhibits an even distribution across categories, it will also be placed at the origin of coordinates, showing no preference as to categories. In general, if an item is close to the origin of coordinates, it is considered to be rather neutral, whereas an item located far from the origin is somehow unique. So, these figures make it easy to grasp relationships among categories and relationships among samples.

As in the cases of Fig. 2 and Figs. 3, 4 and 5, what may be more important in drawing these figures is that the comparison of Figs. 6 and 7 explains: 1) the distribution of categories on the basis of sample distribution, or 2) the distribution of samples on the basis of category distribution. A certain category in Fig. 6 is located in that particular place because of the location of samples in Fig. 7, samples characterised by the high frequency of that particular category. Conversely, the location of a certain sample in Fig. 7 corresponds to categories in Fig. 6, categories characterised by the outstanding frequency in that particular sample.

2.4. **Summary**

Given a frequency distribution table made of a number of categories and a number of samples, the present technique reveals: 1) relationships among categories, 2) relationships among samples and (perhaps more importantly), 3) relationships between categories and samples. The technique operates by providing category weights and sample scores, which maximise the correlation coefficient between categories and samples. Figures plotting category weights and sample scores make it easy to grasp visually these relationships. Often, figures provide interpretations of axes.

3. Results and Discussion

3.1. Three-Dimensional Distribution of Pronominal Categories

"HE" and "SHE" are close together on Axis 1, but rather separated along Axes 2 and 3. "I," "YOU" and "WE" are located in the negative range along Axis 1, but are widely separated along Axes 2 and 3. "IT" and "THEY" are close to the origin of the coordinates, indicating that they are categories used more or less evenly in the entire corpus. Fig. 6, which considers three axes simultaneously, thus illustrates and summarises the facts discussed in Section 2.2.1.

3.2. Three-Dimensional Distribution of Textual Samples

3.2.1. Distribution of the Entire Textual Samples

The sample structure of the entire corpus appears in Fig. 7, reflecting the category distribution depicted in Fig. 6. Samples characterised by high frequency of a particular pronoun are placed around the area assumed by that particular pronoun in Fig. 6. Superimposing Fig. 7 on Fig. 6 confirms the positioning, although axes are drawn somewhat differently in the two figures. Because 728 samples are plotted in this tiny figure, it is difficult to see tendencies due to overlaps. Therefore, some samples are analysed separately, namely, the Book Corpus, the Spoken Corpus, and the Times Corpus, and the samples of particular interest in the Book Corpus and the Spoken Corpus.

3.2.2. Distribution of the Samples in the Book Corpus

Fig. 8 shows distribution of samples in the Book Corpus, except for samples whose quantities for the three axes are between -1.0 and 1.0. Figure 8 shows, therefore, samples unique in that their uses of pronouns are somehow conspicuous. Samples excluded are those whose distribution of pronouns is somewhat even across categories and distribution of those samples is shown and discussed in Fig. 10.

Fig. 8 indicates patterns of the use of pronouns in various kinds of texts in the Book Corpus. The most outstanding pattern is seen in the cases of textual samples located in the upper right-hand part, coinciding with the area of YOU in category distribution. Among YOU-oriented books, some of the conspicuous samples and titles are:

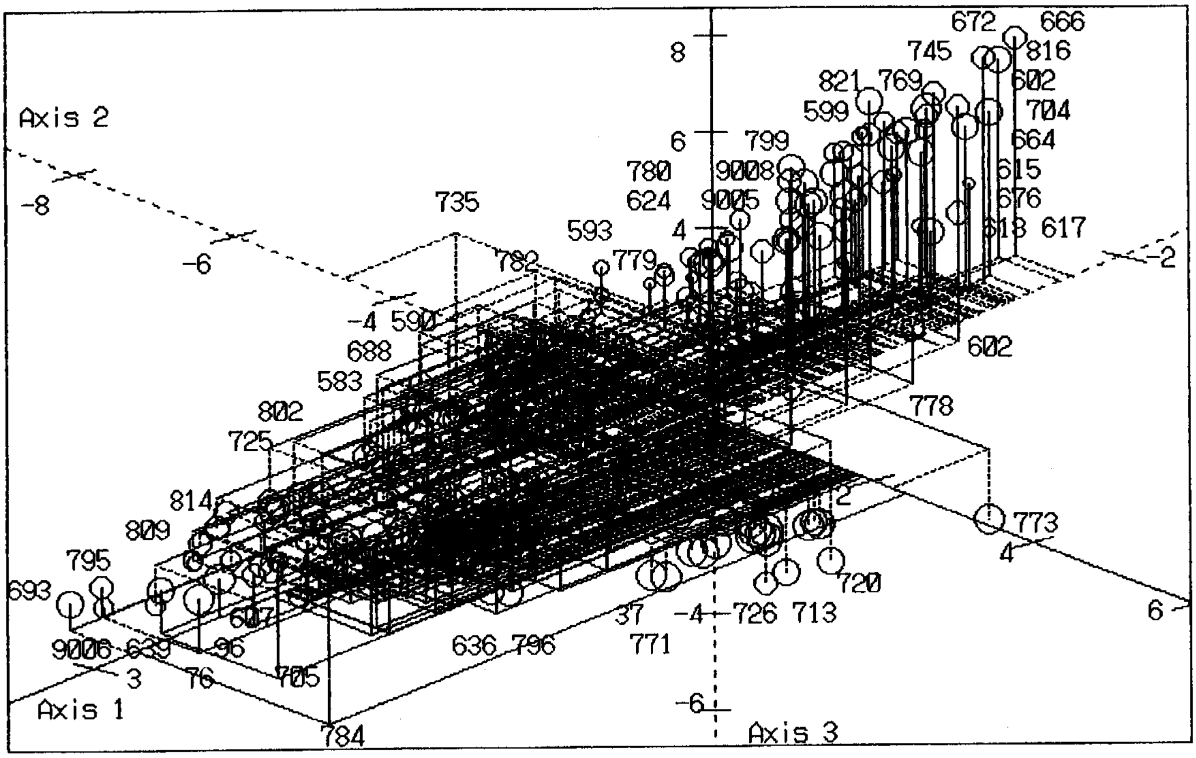

Fig. 8. Distribution of Textual Samples in the Book Corpus Except for Those Close to the Origin of Coordinates.

B666: *Beat Jet Lag.*
B816: (Details not known at the time of this research.)
B672: *Shape your Body, Shape your Life.*
B602: *The Crystal Oracle.*
B704: *The Natural Face Book.*
B664: *Get out of Debt and Prosper.*
B676: *The Playing Card Workbook.*
B617: *Me and You: A self-help plan for managing myalgic encephalomyelitis (post-viral fatigue syndrome).*
B615: *Diabetic Desserts.*
B618: *The Diabetic Microwave Cookbook.*
B745: *Diabetes.*
B769: *How to Stay in Love: the happiness for couples workbook.*
B821: *The Wheel of Fortune: How to control your future.*

As these titles suggest, most are "how-to-do" books and it is clear that such books assume a particular perspective relative to the reader, realised by the abundant use of "YOU." The above list of "YOU"-oriented books differs from the list of books characterised by the high frequency of "YOU" obtained from the adjusted frequency table. The high frequency in the adjusted frequency table does not necessarily mean that "YOU" is the only pronominal form used abundantly in the texts. Other forms of pronouns may also be used abundantly. But the books listed above according to quantification use "YOU" abundantly in comparison with other pronominal forms. Furthermore, even though the absolute frequency of pronouns is low (as in the cases of B615 and B618), the text will be located in the region of "YOU"-oriented books because "YOU" shows a proportionally high frequency compared with other pronouns. This kind of result is obtained only by quantification, not by ordinary descriptive statistics dealing with one variable at a time.

B773, *Inspirational Woman,* a biography of a female occult scientist, is the only book characterised by the high frequency of "I." Incidentally, the fact that B773 has the high frequency of "I" agrees with the results of descriptive statistics of the adjusted frequency.

At the lower left-hand side of the figure, there is a group of several books characterised by the abundance of "HE," including:

B693: *The Gods of Asgard* (Norse mythology).
B795: *The Norse Tarot.*
B9006: *Tarot for Relationships.*
B639: *Relationships, Astrology and Karma.*
B809: *The Spirit of Masonry.*
B96: *Coup d'Etat* (adventure and thriller).
B76: *Chaplin: His life and his art.*

Both "SHE"-oriented books and "WE"-oriented books are hard to see in Fig. 8 because of overlaps, so the distribution of the same texts in the Book Corpus is shown in Fig. 9 from a different perspective.

"SHE"-oriented books include:

B784: *Mauritania.*
B705: *Compania and Lucania.*
B692: *The Second Mauritania.*
B607: *The Female Pope.*
B52: *Miss Melville Regrets.*

B14: *Death of a Nymph.*
B15: *Pretty Maids all in a Row.*

Interestingly, the first three books listed above concern ships or shipping history; they concern figurative women, not real women. Texts identified by two-digit numbers are novels in which, as their titles indicate, a protagonist is probably or definitely a woman. B76, Chaplin's biography, is located close to the region of "HE" and to the region of "SHE," possibly reflecting the fact that many women were involved in his life.

"WE"-oriented books are placed on the lower left-hand corner of Fig. 9. Among them are:

B735: *The Astrology of Karma.*
B782: *Market-Led Strategic Change: Making marketing happen in your organization.*
B673: *Music and the Elemental Psyche: A practical guide to music and changing consciousness.*
B744: *The Crisis of Life on Earth: Our legacy from the second millen-*

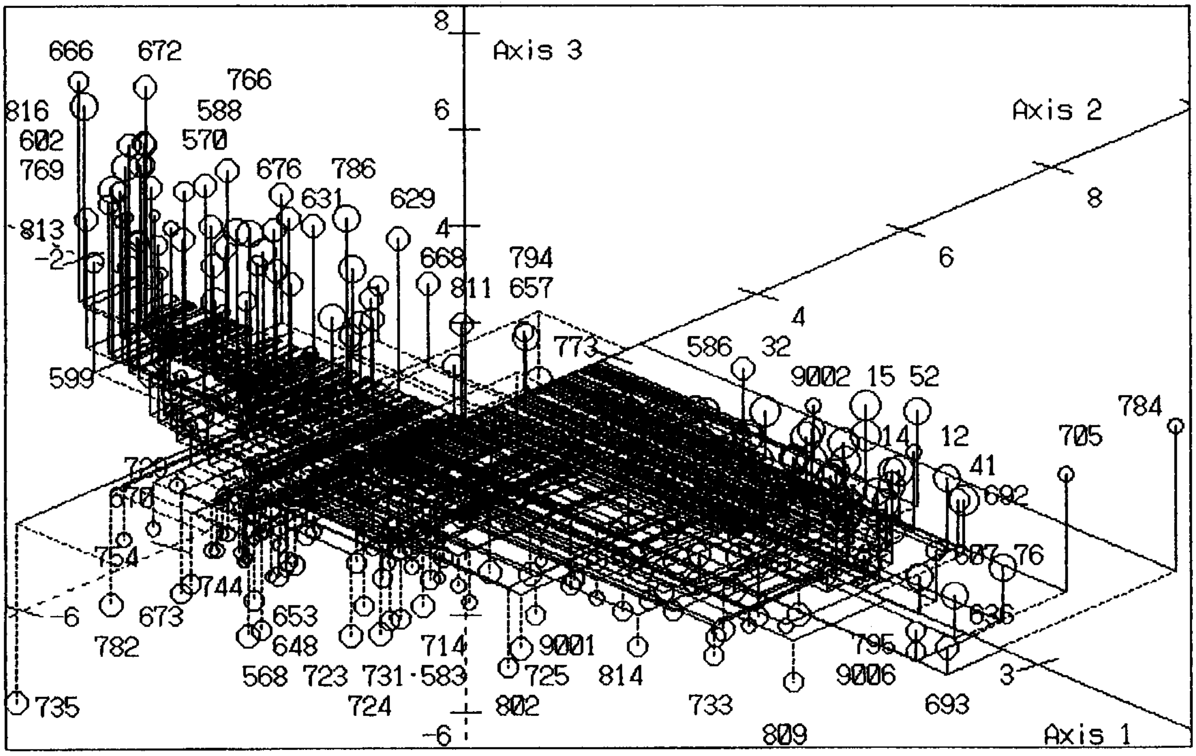

Fig. 9. Distribution of Textual Samples in the Book Corpus from a Different Perspective.

nium.
B754: *Finance for the Non-Financial Manager.*
B670: *Minerals: What they are and why we need them.*

"WE"-oriented books involve both reader and writer in discourse, making situations or problems applicable not only to "you" but also to "me." Such perspective often occurs in books concerning religion,[10] and in books concerning crises or medicine and health. "How-to-do" books also appear here.

Many text samples are in the lower part of the figure, between the areas of "WE"-oriented texts and "HE"-oriented texts, and such texts heavily use either "WE" or "HE" or both. The proportion of the use of "WE" or "HE" is shown by the position each text assumes in this area. The closer to the left the text is located, the more often it uses "WE," and the closer to the right the text is located, the more often it uses "HE." Texts located in the middle of "HE" and "WE" use both pronouns rather equally.

As Fig. 6 shows, "THEY" and "IT" are close to the origin of coordinates, indicating that they are used more or less commonly across text samples. So, the text samples close to the regions of "THEY" and "IT" are certainly characterised by their relatively high frequency, but they are not really unique in the sense that they use one of the two pronouns quite abundantly. However, such texts are rather unique in the sense that their use of pronouns is more or less even across categories.

Fig. 10 shows text samples close to the origin of coordinates, i.e., samples whose quantities given to the three axes are all between -1.0 and 1.0. There are 30 such samples. Those closest to the origin use pronominal categories evenly, including:

B99: *The DeLorean Tapes* (non-fictional account of drug smuggling).
B721: *Biochemical Tissue Salts: A natural way to prevent and cure illness.*
B700: *Greek Vegetarian Cookery.*
B9003: *The Railway Puzzle Book.*
B823: *You and Your Aura.*
B130: *The Right Dose: How to take vitamins and minerals safely.*

[10] This is shown in NAKAMURA (1989b), which studied the structure of the Brown Corpus based upon the pronominal forms.

The remaining samples in Fig. 10 show a slight preference for one particular pronoun or another, depending upon the position they assume in the figure.

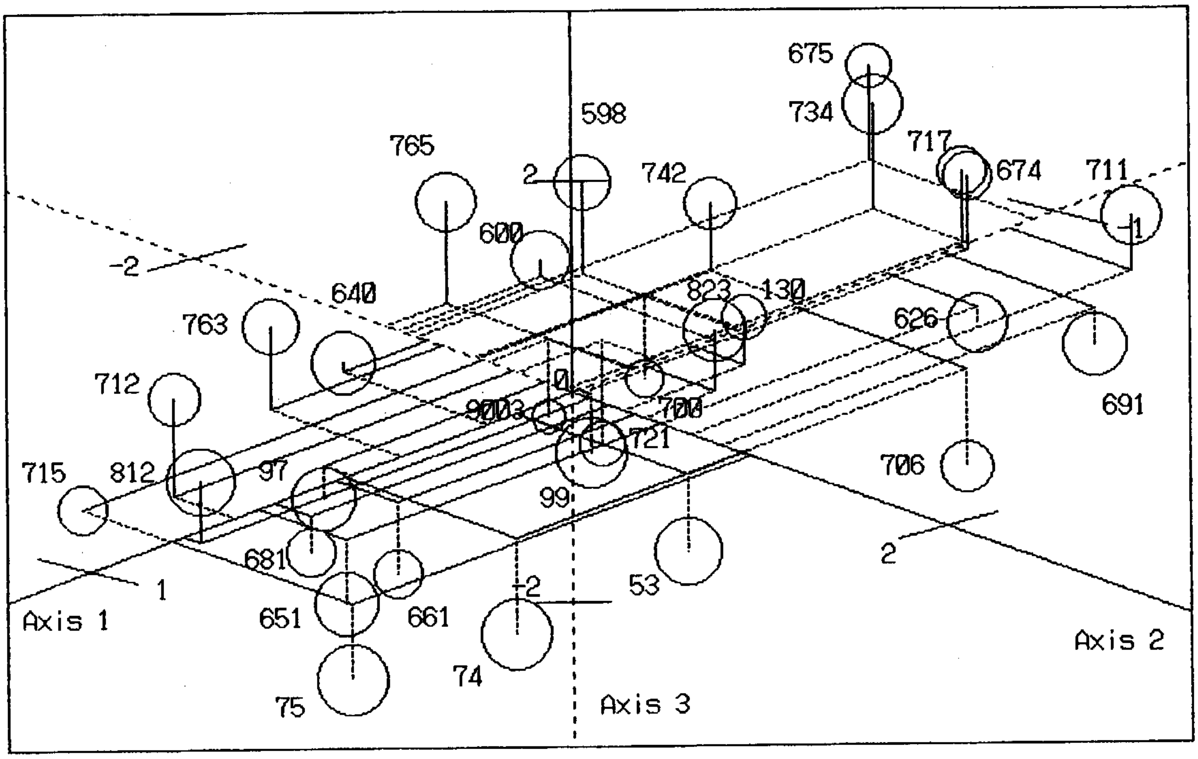

Fig. 10. Distribution of Textual Samples in the Book Corpus Close to the Origin of Coordinates.

Sixty-two text samples classified as fiction in the Book Corpus reveal particular tendencies in their choice of pronouns (see Fig. 11.).[11] The distribution of these works of fiction is restricted in the positive region along Axes 1 and 2. There are two exceptions, B96 (*Coup d'Etat*) and B97 (*San Andreas*), wherein the quantities along Axis 2 are negative, although their values are almost nil. B96 assumes that position because of its conspicuously frequent use of "HE," whereas B97, which is characterised by relatively frequent uses of "IT," "THEY," "YOU" and "WE" and by low occurrence of "SHE," is an outlier in the group of fiction.

All of these texts use pronouns abundantly as indicated by rather large

[11] Those not labelled are all located in the centre of the figure, including B4, B7, B8, B9, B11, B20, B24, B26, B29, B30, B34, B35, B39, B42, B44, B49, B54, B62, B71, B72, and B650.

balls placed at the ends of the vertical lines. There are two types of fiction: one type closer to the area of "I" and the other type closer to the areas of "HE" and "SHE," as indicated in Fig. 11.

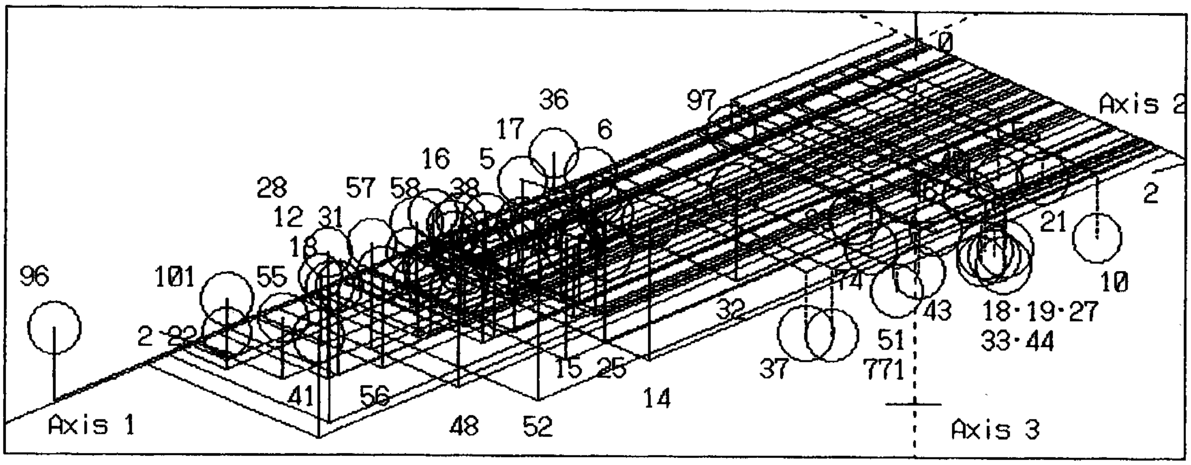

Fig. 11. Distribution of Fictions in the Book Corpus.

Samples located on the right-hand side are closer to the region of "I," meaning that "I" is more often used than "HE" or "SHE." Such samples include:

B10:	*The Tartan Ringers.*
B21:	*The Gwen John Sculpture.*
B18:	*The Disposal of the Living.*
B19:	*The Dutch Blue Error.*
B33:	*C. B. Greenfield: A Little Madness.*
B44:	*Forests of the Nights.*
B27:	*Follow the Sharks.*

Samples on the left-hand side of the figure show high frequency of either "HE" or "SHE." B96 is the sample closest to the position of "HE," as briefly mentioned above. B22 (*Man's Loving Family*) and B101 (*Night Birds* (children's fiction)) are also close to that area. B52 (*Miss Melville Regrets*) is the sample closest to the region of "SHE" and B14 (*Death of a Nymph*) is also close to that area. Texts located between the area of "HE" and the area of "SHE" use "HE" and "SHE" more or less equally and more often than "I." Such texts as B12 (*The Quest for K*), and B31 (*Last Shot*) are examples.

Texts located in the middle of "I"-oriented fiction and "HE"-or-"SHE"-oriented fiction, of course, use these pronouns rather evenly. There is only one text of that sort in the Bank of English, B32 (*The Unorthodox Murder of Rabbi Moss*).

On the whole, the distribution of samples in the Book Corpus is complicated, but there is a definite pattern of distribution. "IT" and "THEY" are quite neutral in the Bank of English, but other pronouns seem to be preferred in some text types, and not preferred in others. Samples in the Book Corpus, placed peripherally in Figs. 8 and 9, show a clear preference for one or two types of pronouns over other types. The outstanding type of pronoun in the Bank of English is probably "YOU," considering its distribution and its frequency of use. Many of the texts in the Book Corpus, especially the "how-to-do" books, contribute to its uniqueness in the Bank of English. Other conspicuous types of pronouns are "HE" and "SHE," but their use appears heavily dependent upon the topics of individual books, such as books about ships which are characterised by high frequency of "SHE." There are relatively few "I"-oriented books, but "I," "HE" and "SHE" are the three pronouns which characterise works of fiction in the Book Corpus.

3.2.3. Distribution of the Samples in the Spoken Corpus

As discussed in Section 2.2.2, text samples in the Spoken Corpus are not as scattered as in the Book Corpus, a fact clearly shown in Fig. 12. Dispersion of samples is on a much smaller scale than in the case of the Book Corpus, but Fig. 12 shows clear tendencies in the use of pronominal forms among text samples.

On the right-hand side of Fig. 12, "I"-oriented texts are identified by the vertical dotted lines descending. Typical examples are S9, S10, S12 and S11, all telephone conversations. (As a matter of fact, there are only 5 telephone conversations if radio phone-in programmes are excluded, and of these five, four of them are located in this area.) Texts characterised with the high frequency of "YOU" are identified by the solid vertical lines ascending in the right-hand side of the figure. The examples such as S40, S39, S7, S261, S79, S80, S23, S87, S249 and S263 are all samples of the

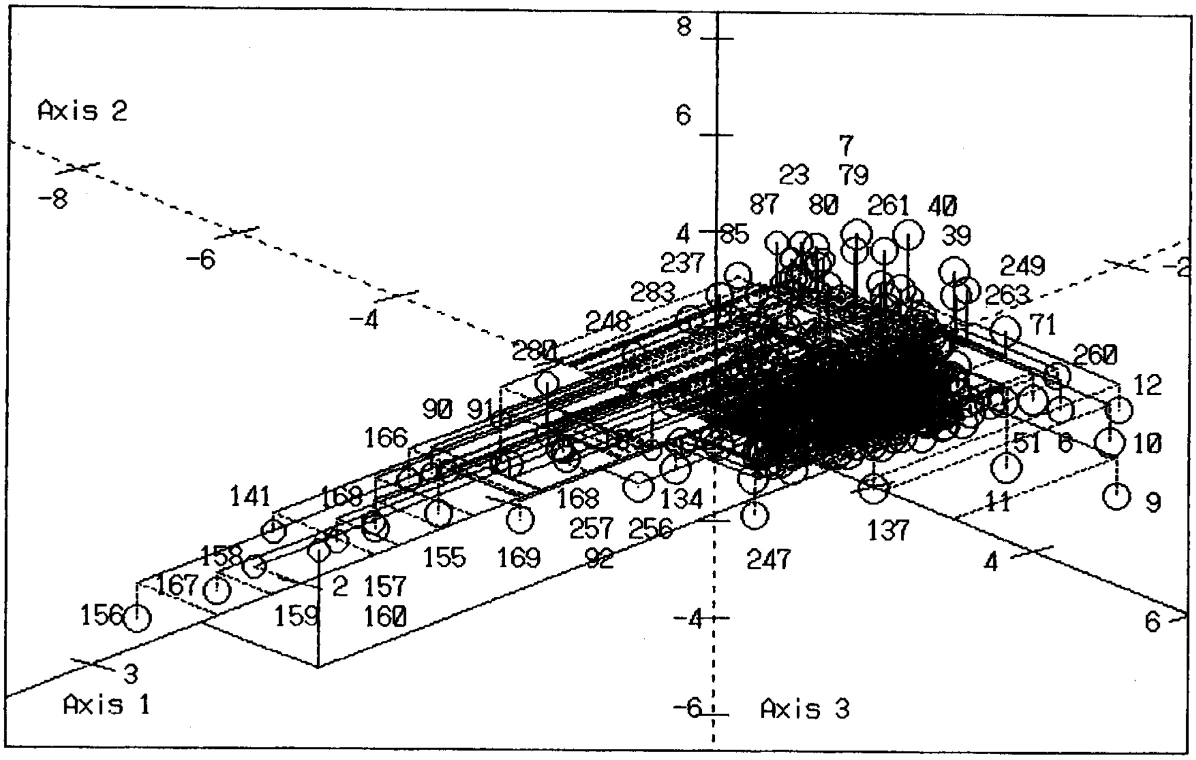

Fig. 12. Distribution of Textual Samples in the Spoken Corpus.

face-to-face mode such as lectures, seminars, talks or guided tours. Only one sample, S105, is "WE"-oriented and that is part of the lecture series concerning architecture in Birmingham. Most samples scattered in the positive region of Axis 1 are lectures, a fact to be elaborated upon later.

As Fig. 12 shows, the most concentrated area of text samples in the Spoken Corpus is quite restricted. Many samples are located in the negative range, from the origin to around -1.0, along Axis 1, and in the positive range, from the origin to around 1.0, along Axis 2. Incidentally, this area is where text samples classified as "various" or "various (mainly current affairs)" are located (see Fig. 13). Here 152 texts are found, 22 described as "various" and 130 described as "various (mainly current affairs)." All of the radio phone-in programmes belong to the latter text type.

Conversations without a specific topic fall in this area and their use of pronouns is neutral if texts are placed close to the origin of coordinates, "YOU"-oriented if they are placed close to -1.0 on Axis 1 and "I"-oriented if they are close to -0.5 on Axis 1 and 1.0 on Axis 2. Fig. 14 shows that many texts described as "YOU"-and-"I"-oriented are located between the

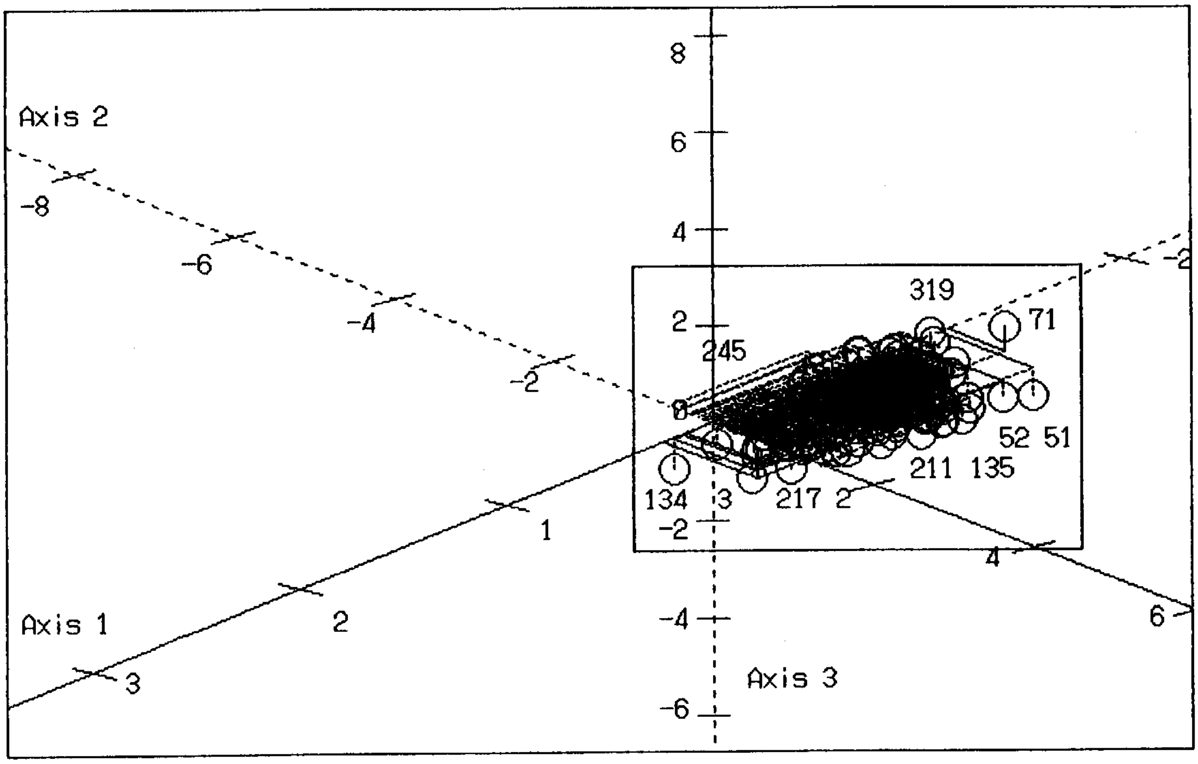

Fig. 13. Distribution of Textual Samples in the Spoken Corpus, the Topics of Which are Either "Various" or "Various (Mainly Current Affirs)."

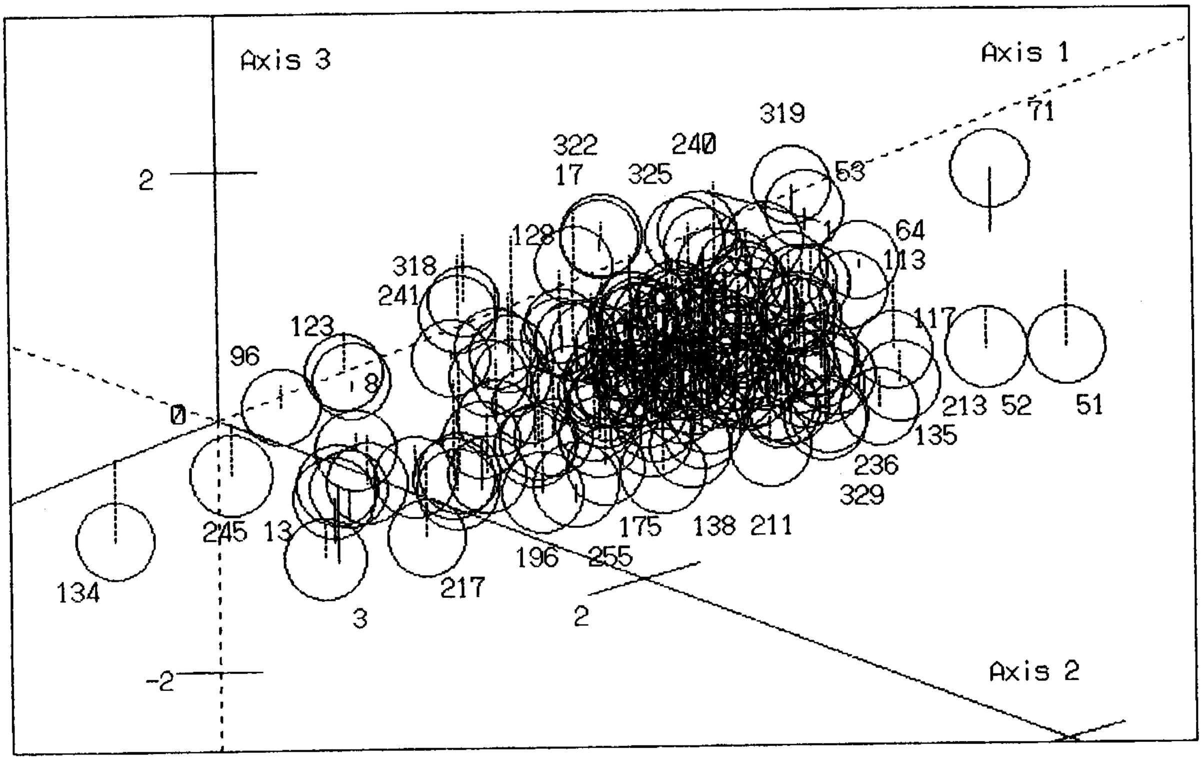

Fig. 14. Enlarged Extract from Fig. 13.

area of "I" and the origin.[12]

In contrast to samples discussed immediately above, many other text samples, mainly texts taken from lectures, talks, seminars and course sessions, are scattered throughout the whole figure. Of 288 texts in the Spoken Corpus, there are 51 lectures, 16 talks, 15 seminars and 7 course sessions for a total of 89. Fig. 15 shows the distribution of texts, suggesting that their distribution is highly dependent on topic in contrast to texts with topics described as "various."

The idea of topic-dependent distribution is confirmed in Fig. 16, in which two series of lectures, 19 on "English Literature" and 15 on "Development Finance,"[13] are extracted from Fig. 15. These two series of lectures differ greatly in their uses of pronouns. The lectures on "Development Finance" are all located on the upper left-hand side of the figure, where "YOU"-oriented texts are supposed to be located and their distribution along Axis 2 suggests that the second frequent category is "WE." The same lecturer talking to the same students on a general topic related to "Development Finance" may cause distribution of the texts in a limited area. This particular lecturer seems to involve listeners and lecturer in ongoing discourse, a fact indicated by the abundant use of "YOU" and "WE."

In contrast to the lectures on "Development Finance," those on "English Literature" are scattered throughout the whole figure except where "I"-oriented samples should be located. As Fig. 16 shows, lecture topics on "English Literature" are highly specific, and such specificity causes dispersed distribution of texts. Lectures on male writers are all located in the positive range along Axis 1, indicating that texts use "HE." In the case of a female writer, Mary Wollstonecraft, the tendency toward a particular use of pronoun is more obvious. The text is located exactly in the area where "SHE"-oriented texts are supposed to be. "Beginning of English Jacobin-

[12] For the sake of clarity, horizontal lines of each sample along Axis 1 and Axis 2 are deleted in this figure.

[13] The source information of the Spoken Corpus does not contain a specific topic for each lecture of this series. Textual samples located on the upper right-hand side of the figure and not labelled are all from this series.

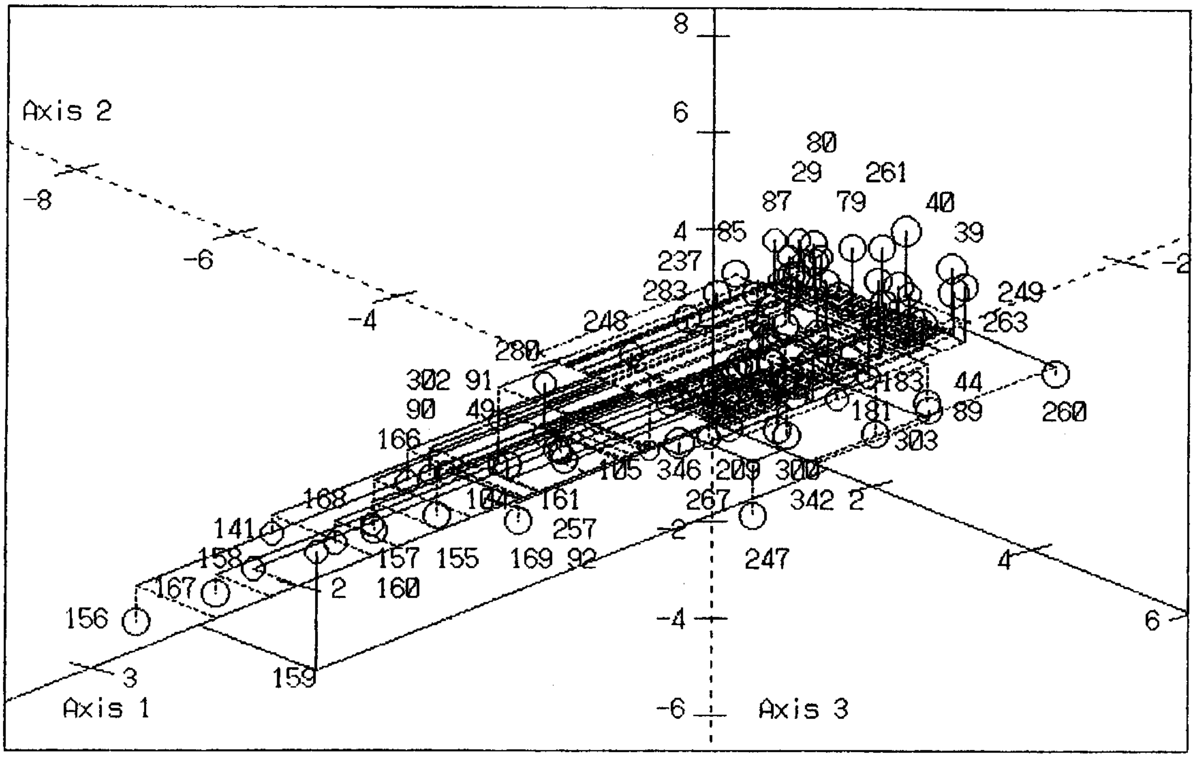

Fig. 15. Distribution of Textual Samples from Lectures, Talks, Seminars and Course Sessions in the Spoken Corpus.

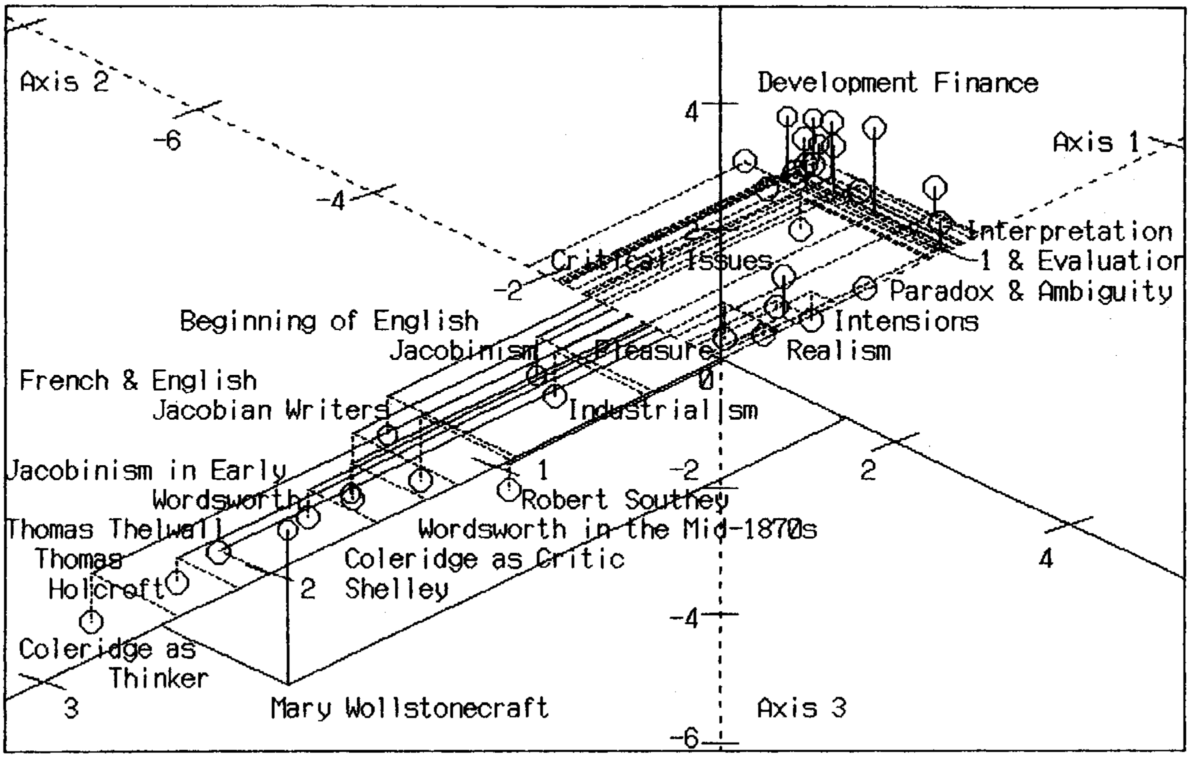

Fig. 16. Distribution of Textual Samples from the Lectures "Development Finance" and "English Literature."

ism" and "Industrialism" are two lectures which show frequent use of "THEY." These lectures probably refer to many writers. When topics are rather abstract, like "Pleasure," "Evaluation," "Realism," "Critical Issues," and "Intentions," texts are located close to the origin of coordinates. Only one lecture is "YOU"-oriented and as its title "Interpretation and Evaluation" suggests, the audience is probably supposed to interpret and evaluate, a supposition confirmed by the positioning of the text in the distributional figure.

Text samples in the Spoken Corpus are roughly divided into two kinds: 1) conversations on unspecified topics and 2) lectures, talks, seminars and course sessions with clearly identified topics. Samples of the first kind are distributed around the areas where "I"-oriented and "YOU"-oriented samples are supposed to be. Samples of the second kind are scattered throughout the figure, except in the area where "I"-oriented samples are supposed to be. The distribution of samples of the second kind looks like the distribution of the Book Corpus. But the Spoken Corpus lacks samples more or less "I"-oriented, such as biography in the Book Corpus and samples corresponding to those classified as fiction.

3.2.4. Distribution of the Samples in the Times Corpus

As discussed above in Section 2.2.2, all text samples in the Times Corpus appear in a very narrow range along each axis, and Fig. 17 shows this fact. Fig. 18 is the enlarged detail of Fig. 17, without the coordinates of each sample.[14] The concentration of samples in the Times Corpus is nearly the same as in cases of text samples classified as "various" in the Spoken Corpus. Texts in the Times Corpus are collected issue by issue, making topics "various," and ultimately cancelling the effects of particular usage of pronouns according to text topics. The Times Corpus and the texts classified as "various" in the Spoken Corpus differ because the Times Corpus is

[14] Horizontal and vertical lines indicating coordinates are deleted since all are very close together. The last two digits of the text identifiers in the above tables, i.e., "91," are all deleted in this figure since they do not help distinguish one text from another.

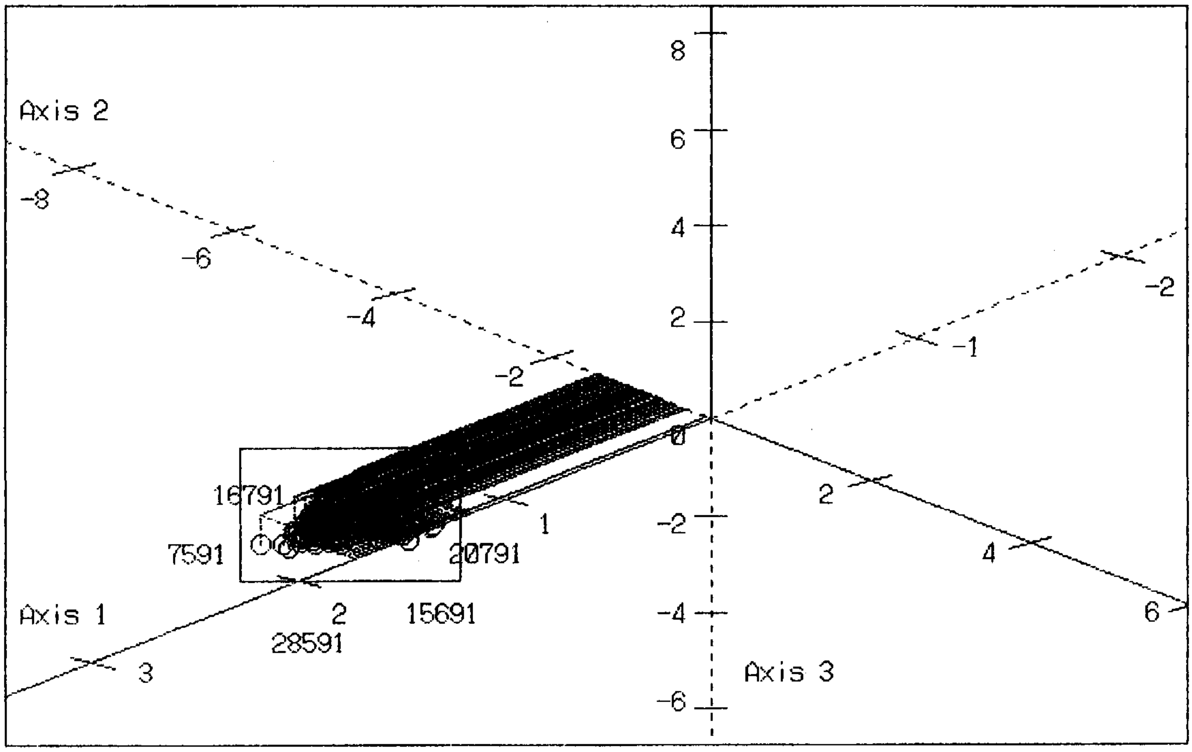

Fig. 17. Distribution of Textual Samples in the Times Corpus.

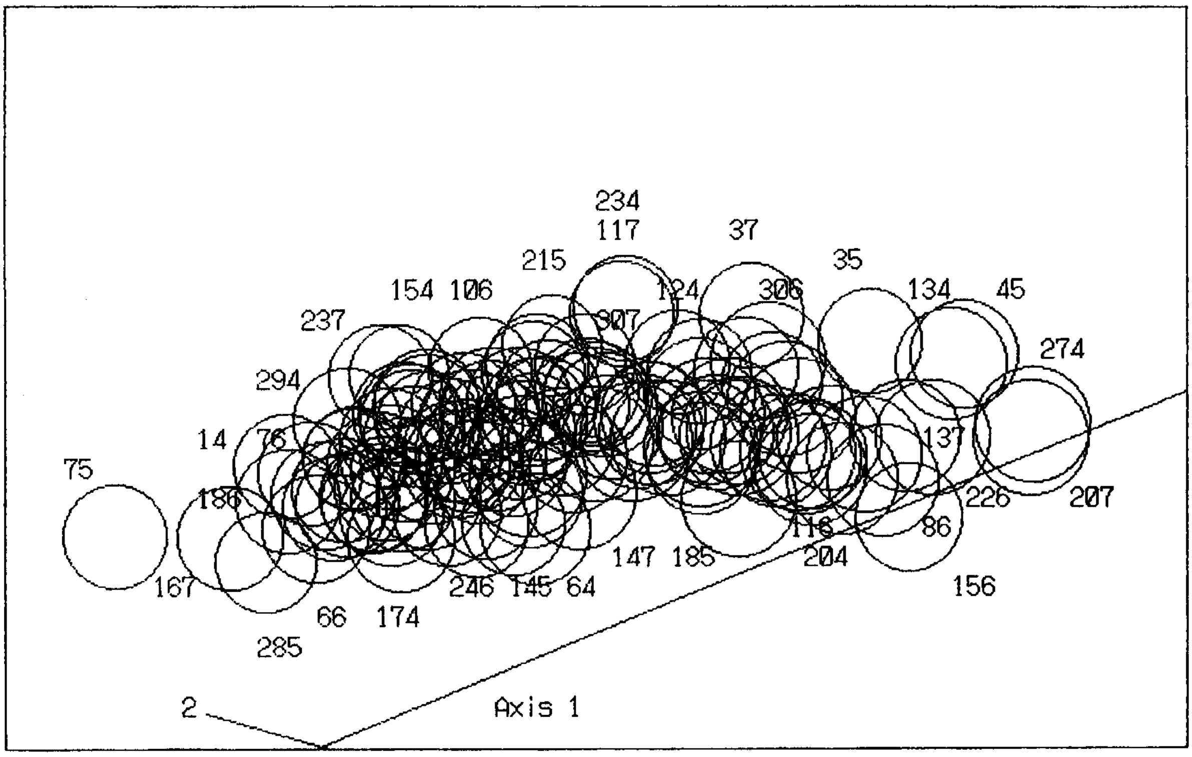

Fig. 18. Enlarged Extract from Fig. 17.

close to the areas of "HE" and "SHE" and the texts classified as "various" in the Spoken Corpus are close to "YOU" and "I." Such differences reflect the fact that newspaper English is of a public nature and the texts classified as "various" in the Spoken Corpus are usually of a private or personal nature.

The concentration of texts in the Times Corpus also suggests that collecting data issue by issue may not be sufficient to analyse newspaper English stylistically. There needs to be a further breakdown of the texts into genres as in the Brown and the LOB Corpora, where (at least) reportage, editorials and reviews are treated separately. The text sizes are already fairly large, especially in comparison with some texts in the Spoken Corpus, so breakdown of each issue into several genres may not be so difficult. Each section of newspaper can be treated and classified quite mechanically according to its specifically defined purpose.

4. Conclusions

The extended version of HAYASHI's Quantification Method Type III was applied to the adjusted frequency list of the seven pronominal categories across 728 textual samples in the Book Corpus, the Spoken Corpus and the Times Corpus. The structure of the Bank of English from the viewpoint of the use of pronouns was determined and shown visually in the figures above. Thus, the extended HAYASHI's Quantification Method Type III proves to be a useful tool for exploring: 1) relationships among pronominal categories themselves, 2) relationships among text samples themselves, and (perhaps more importantly) 3) relationships between pronominal categories and text samples. The most important factor in determining the structure is the degree of personal involvement or personal detachment in the discourse, a fact revealed through interpretation of the distribution of categories and samples along Axis 1. Interpretations of the other two axes are, however, not yet clear.

Distribution of most text samples depends heavily upon topics, if topics are specified. If topics are not clearly specified, texts tend to concentrate together in a small area reflecting their use of pronouns. This fact is illustrated by texts in the Times Corpus and texts with topics classified as "various" in the Spoken Corpus.

This methodology objectively corroborates some of our intuitive knowledge concerning the use of pronominal forms in various kinds of texts. Such corroboration proves the validity and usefulness of this method. In fact, the extended version of HAYASHI's Quantification Method Type III is a powerful tool for dealing with the frequency distribution of various categories across various text types, determining the typology of text types based on those categories.

Textual samples can be classified according to parts of speech, such as verbs, nouns, adjectives and adverbs. NAKAMURA (1993b) uses three semantic classes of verbs to see relationships among the Book Corpus, Spoken Corpus, Times Corpus and BBC Corpus in the Bank of English. Grammatical tags are chosen in NAKAMURA (1990, 1991b and 1992) to see relationships among genres in the Brown and the LOB Corpora. Even phrasal categories or syntactic categories may be used to classify textual types, if such terms are clearly defined and if data are prepared in a machine-readable format.

If the frequency list of collocates of a particular word across several types of texts is provided, text typology concerning collocates of that particular word can also be determined. So this method is also useful in the study of collocations.[15] The extended version of HAYASHI's Quantification Method Type III could also be applied to variation studies such as the comparison of British and American English (see NAKAMURA (1992 and 1993a)). Actually, there is no limit in applying this methodology if there is a clear framework for research and if an appropriate mass of data to support that framework is obtained.

[15] Incidentally, I am presently studying the collocates of the word "woman" in the Bank of English in collaboration with John Sinclair. We hope that our work will be published sometime in the not so distant future.

References

AARTS, Jan and Willem MEIJS, eds. 1990. *THEORY AND PRACTICE IN CORPUS LINGUISTICS.* Amsterdam and Atlanta: Rodopi.

AIJMER, Karin and Bengt ALTENBERG, eds. 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* London and New York: Longman.

BAKER, Mona, Gill FRANCIS and Elena TOGNINI-BONELLI, eds. 1993. *TEXT AND TECHNOLOGY: In Honour of John Sinclair.* Philadelphia and Amsterdam: John Benjamins.

BIBER, Douglas. 1988. *Variations across speech and writing.* Cambridge: Cambridge University Press.

———. 1989. A typology of English texts. *Linguistics* 27:3-43.

FRANCIS, W. Nelson and Henry KUČERA. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton & Mifflin.

HOFLAND, Knut and Stig JOHANSSON. 1982. *Word Frequencies in British and American English.* Bergen: The Norwegian Computing Center for the Humanities.

JOHANSSON, Stig. 1979. CORPUS-BASED STUDIES OF BRITISH AND AMERICAN ENGLISH. In *PAPERS FROM THE SCANDINAVIAN SYMPOSIUM ON SYNTACTIC VARIATION,* ed. by Sven JACOBSON. 85-100. Stockholm: Almqvist & Wiksell.

JOHANSSON, Stig and Anna-Brita STENSTRÖM, eds. 1991. *English Computer Corpora: Selected Papers and Research Guide.* Berlin and New York: Mouton de Gruyter.

JOHANSSON, Stig and Knut HOFLAND. 1988. *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus Vol. 1: Tag Frequencies and Word Frequencies.* Oxford: Clarendon Press.

———. 1989. *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus Vol. 2: Tag Combinations and Word Combinations.* Oxford: Clarendon Press.

KUČERA, Henry and W. Nelson FRANCIS. 1967. *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

MIYAJIMA, T. 1970. *Goi no Ruijido* (Similarity of Vocabulary). *Kokugogaku* (Japanese Linguistics) 82:42-96.

MIZUTANI, Shizuo. 1977a. *Keiryo-goiron kara Mita Myojoha to Negishi-ha* (*Myojo* School and *Negishi* School from the view point of quantitative analysis of vocabulary). *Keiryo-Kokugogaku* (Quantitative Japanese Linguistics) 11:30-37.

———. 1997b. *Suri-Gengogaku* (Mathematical Linguistics). Tokyo: Baifu-kan.

MOSTELLER, F. and R. E. K. ROURKE. 1973. *Sturdy Statistics: Nonparametrics and Order Statistics.* Reading, Mass.: Addison-Wesley Pub. Co.

NAKAMURA, Junsaku. 1985. On the Methodologies of Quantitative Groupings of English Texts. *JACET* (The Japan Association of College English Teachers) *BULLETIN* 16:133-148.

———. 1986. Classification of English Texts by Means of HAYASHI's Quantification Method Type III. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 21:71-86.

———. 1987. Notes on the Use of HAYASHI's Quantification Method Type III for Classifying English Texts. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 22:127-145.

———. 1989a. Creation of a Vocabulary Frequency Table from the Brown Corpus. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 24:171-182.

———. 1989b. A Quantitative Study on the Use of Personal Pronouns in the Brown Corpus. *JACET BULLETIN* 20:51-71.

———. 1990. A Study on the Structure of the Brown Corpus Based upon the Distribution of Grammatical Tags. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 1:13-35.

———. 1991a. A Study on the Structure of the Brown Corpus Based upon the Distribution of its Vocabulary Items. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 2:27-47.

———. 1991b. The Relationships among Genres in the LOB Corpus Based upon the Distribution of Grammatical Tags. *JACET BULLETIN* 22:55-74.

———. 1992. The Comparison of the Brown and the LOB Corpora Based upon the Distribution of Grammatical Tags. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 3:43-58.

———. 1993a. Quantitative Comparison of Modals in the Brown and the LOB Corpora. *ICAME Journal* 17: 29-48.

———. 1993b. Statistical Methods and Large Corpora: A new tool for describing text types. In *Text and Technololgy: In Honour of John Sinclair,* ed. by Mona BAKER, Gill FRANCIS and Elena TOGNINI-BONELLI. 293-312. Amsterdam and Philladelphia: John Benjamins.

NAKAMURA, Junsaku and John M. SINCLAIR. "The World of 'Woman' in the Bank of English: An application of extended HAYASHI's quantification method type III in the study of collocations." In preparation.

SINCLAIR, John M., ed. 1987. *Looking Up: An account of the COBUILD Project in lexical computing.* London and Glasgow: Collins ELT.

SINCLAIR, John M. 1991. *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

———. 1992. "Lexicographers' Needs," Unpublished MS, Pisa Workshop on Text Corpora.

SVARTVIK, Jan, ed. 1990. *The London-Lund Corpus of Spoken English: Description* *and Research.* Lund:Lund University Press.

———. 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991.* Berlin and New York: Mouton de Gruyter.

YASUDA, S. and M. UMINO. 1977. *Shakai Tokeigaku* (Social Statistics). 2nd rev. ed. Tokyo: Maruzen.