n-Gram Statistics for Natural Language Understanding and Text Processing

CHING Y. SUEN, SENIOR MEMBER, IEEE

Abstract—n-gram (n = 1 to 5) statistics and other properties of the English language were derived for applications in natural language understanding and text processing. They were computed from a well-known corpus composed of 1 million word samples. Similar properties were also derived from the most frequent 1000 words of three other corpuses. The positional distributions of n-grams obtained in the present study are discussed. Statistical studies on word length and trends of n-gram frequencies versus vocabulary are presented. In addition to a survey of n-gram statistics found in the literature, a collection of n-gram statistics obtained by other researchers is reviewed and compared.

***Index Terms*—Character recognition, context, language understanding, n-gram statistics, positional distributions of letters, text processing, word length analysis.**

I. Natural Language Understanding and Text Processing

AMONG the numerous applications of pattern recognition techniques, the subject of character recognition has been nurtured intensively, mainly because of its practical value in data processing and in computer input of large volumes of data. Although many commercial OCR machines are available, their capabilities are often limited to single font or stylized font reading. Those which have been designed for multifont applications are not only prohibitively expensive (over $1 million), but are still in the process of refinement. Yet, unlike human beings, none of these machines seems to have made much use of contextual information. Factors like letter sequences, word dependencies, sentence structures and phraseology, style and subject matter, as well as comprehension, knowledge, inference, association, guessing, prediction, and imagination all take place very naturally during the process of human reading. These processes take place extremely effectively and efficiently in the human brain because they are the results of many years of trial, learning, and correction.

Many investigations on the process of human reading and comprehension, and effects of contextual information have been made by linguists and psychologists [1], [5], [8], [18], [40], [41], [49], [50], [54], [73], [94]. Since there are thousands of type fonts in the world, it may not be feasible to build an optical machine which is capable of recognizing all of them by shapes alone. The best solution seems to be "making machines more intelligent" like human beings—a major step towards successful artificial intelligence in the area of natural language understanding and processing. In order to do so, the use of contextual information is indispensable. It is only based on such intelligence that more advanced aspects in machine understanding of natural languages such as syntax, grammer, vocabulary, concordance, pronunciation, and other properties can be explored. Indeed, text processing, combined with techniques dealing with automatic correction of deletion, substitution, and insertion errors, has become a subject of great interest [35], [37], [56], [64], [93], [95].

The importance of text analysis can be illustrated by the following applications. Various properties of natural languages have been investigated, e.g., analyses of the frequency of occurrences of letters, words, and phonemes [13], [14], [22], [42], [89], information content and prediction of letter sequences [15], [36], [55], [67], [72], [73], [102], and language understanding and text analysis [5], [46], [48], [59], [63], [64], [99]-[101]. Apart from the numerous efforts towards error correction in characters and texts cited in the Reference section of this paper, the following applications have been explored: 1) automatic proofreading and correction of typographical errors and misspelled words produced either by machines or by humans [2], [9], [12], [19], [20], [32], [51], [52], [80], [88], [97], 2) handwritten and spoken computer programs [3], [25], [39], 3) name records [12], [21], 4) address reading by mailsorting machines [24], [29], [33], [65], [66], and 5) cryptanalysis [31], [57].

In the above applications, two main methods of processing natural language have been utilized, i.e., the dictionary and n-gram methods. In the dictionary method, words are compared with those stored in memory through string-to-string matching [2], [88], [98]. For small vocabularies such as computer statements, principles of both syntax and semantics can be added to enhance error correction and good results have been reported [3], [10], [20], [25], [81]-[84]. In the n-gram method, a number (n) of letters contained in the words are compared with their probability of occurrence stored in the computer. Owing to substantial saving in memory space and computing time, the n-gram method has been applied by many researchers, e.g., [12], [19], [23], [29], [34], [35], [38], [52], [53], [61], [69]-[71], [86], [95], [97]. Also, key letters have been tried [85].

In connection with the research program on handprint recognition and standardization, as well as on reading machines for the blind [75]-[79], the author has made an analysis of some interesting properties of natural languages. It is the purpose of this paper to present n-gram and other properties

Manuscript received March 30, 1978; revised September 6, 1978. This work was supported by the Education Department of Quebec and the National Research Council of Canada.

The author is with the Department of Computer Science, Concordia University, Montreal, P.Q., Canada, and the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139.

0162-8828/79/0400-0164$00.75 © 1979 IEEE

TABLE I
DISTRIBUTION OF UNIGRAMS IN DIFFERENT POSITIONS

	Starting Position										
	1	2	3	4	5	6	7	8	9	10	11
A	2.8087	2.2362	1.2958	.4973	.3734	.3337	.2804	.1048	.0899	.0405	.0154
B	1.0289	.0987	.0981	.0991	.0351	.0583	.0268	.0150	.0070	.0036	.0007
C	.9848	.1526	.5431	.4395	.2598	.2183	.1094	.0809	.0494	.0214	.0133
D	.6168	.0874	1.2489	.5185	.4067	.4530	.2231	.1881	.1079	.0524	.0212
E	.5151	2.7375	3.4775	2.0350	1.6164	.8776	.6589	.3914	.2245	.1077	.0536
F	.9432	1.1086	.2270	.1045	.0320	.0490	.0322	.0110	.0005		
G	.3477	.0625	.2753	.2131	.2337	.1758	.2061	.1062	.0759	.0292	.0152
H	1.2976	3.5196	.1153	.5587	.3195	.1413	.0683	.0353	.0162	.0102	.0016
I	1.6479	1.6459	.8815	.8552	.7624	.4708	.3605	.1834	.1321	.0516	.0334
J	.1132	.0000	.0250	.0126							
K	.0992	.0090	.1296	.2440	.0886	.0126	.0049	.0017	.0010		
L	.5190	.4571	.7315	.7945	.5254	.2329	.2149	.1603	.0887	.0533	.0379
M	.8715	.1430	.6472	.4198	.1192	.1173	.1115	.0348	.0114	.0068	.0016
N	.5023	2.2967	.9201	.8935	.5549	.6603	.3297	.3638	.2001	.1486	.0632
O	1.7886	3.4089	1.1016	.5086	.4234	.1540	.1608	.1034	.0921	.0619	.0313
P	.8036	.2304	.3209	.2910	.0985	.0452	.0385	.0085	.0045	.0072	.0014
Q	.0390	.0168	.0202	.0082	.0017	.0040	.0041				
R	.5232	1.1624	1.5615	.7186	.8893	.5548	.2165	.1578	.0460	.0317	.0080
S	1.4799	.5997	1.3045	.8174	.5830	.4968	.4320	.2223	.1371	.0767	.0405
T	3.9845	.7901	1.3825	1.3774	.6960	.5507	.4148	.2386	.1641	.1035	.0524
U	.2383	.8646	.5788	.4159	.2355	.1025	.0819	.0489	.0260	.0040	.0015
V	.1248	.1329	.3838	.1326	.0589	.0682	.0169	.0301	.0093	.0107	.0012
W	1.5080	.0969	.2528	.1438	.0302	.0420	.0033	.0032	.0018		
X	.0000	.1149	.0486	.0044	.0041	.0036	.0027	.0005	.0005		
Y	.2089	.2303	.2537	.2792	.1486	.1406	.1499	.1027	.0783	.0525	.0295
Z	.0006	.0000	.0098	.0081	.0069	.0050	.0081	.0063	.0012	.0004	.0005
-	.0000	.0000	.0008	.0008	.0064	.0009	.0023	.0004	.0000	.0000	.0004
'	.0000	.0185	.0197	.0445	.0329	.0137	.0129	.0042	.0013	.0007	.0015
Total	22.9953	22.2209	17.8552	12.4357	8.5423	5.9835	4.1713	2.6037	1.5666	.8746	.4255

of the English language obtained from a computational analysis of well-known corpuses, including the 1 million words of Kucera *et al.*'s data base [42], and the first 1000 most common words compiled by Dewey [22], Thorndike *et al.* [89], and Carroll *et al.* [14]. The analysis was based on the frequency of occurrence of English words obtained from running texts. For different data bases, the n-gram (for $n = 1$ to 5) statistics which occur in different positions (1 to 11) of the words have been computed. Statistical studies on word length and trends of n-gram frequency versus vocabulary are presented. In addition to an extensive review of n-gram statistics found in the literature, a collection of n-gram obtained by other researchers are reviewed and compared. The various figures and tables presented here should be useful to those researchers who are engaged in automatic processing of English texts, computational analysis of linguistics, and machine understanding of human languages.

II. n-GRAM STATISTICS

n-gram statistics were computed from well-known corpuses including the 1 million words compiled by Kucera *et al.*, and the first 1000 most frequent words compiled by the above authors, by Carroll *et al.*, Dewey, and Thorndike *et al.* These two sets of data will be called Data 1 and Data 2 in subsequent sections.

A. Data 1

These data consist of 1 million words collected by Kucera *et al.* from 500 pieces of carefully selected samples of natural language texts printed in 1961 and written by American writers. These samples were distributed among 15 categories including reportage, editorial and reviews of the press, texts on religion, skills and hobbies, humor, Popular Lore, Belles Lettres, biography, learned and scientific writings, different types of fiction, and miscellaneous topics.

As this corpus was intended for general use, special symbols and specific codes were included. In order to obtain the n-gram statistics, the corpus was preprocessed to eliminate special symbols, punctuation marks, numbers, diaeresis or umlaut, points of ellipsis, Greek alphabets, misspelled words, etc. The only punctuation marks retained were the hyphen and the apostrophe since they appear within regular words. The total number of samples selected was 922 000 words.

Positional distributions of characters is an important aspect in text processing and automatic correction of errors. Distribution of unigrams in different positions is shown in Table I. This distribution is dependent on the word length, which also plays an important role in text processing and natural language understanding. For example, using word length and the first or the second half of the word, a contextual postprocessing technique has been designed [24] for address reading in a postal system. Word length information was utilized in studies done by several researchers [10], [27], [33], [34].

As can be seen from Table I, the first four positions account for more than 75 percent of the total number of letters. This can be explained from histogram word distributions according to length presented in Fig. 1. Most of the words occupy two to five letters, yielding an average word length of about 4.5 letters. The letter E has the highest frequency of occurrence, followed by T, A, O, I, etc. Their ranks are highly dependent on words with high frequency, e.g., *the*, *of*, *and*, *to*, *a*, *in*, etc. It is interesting to note that the letter J occupies only four positions, Q, 7 and F and X, 9. The other letters occupy the entire span of 11 positions. A comparison of unigram distributions will be described later.

The percentage distribution of bigrams is shown in Table II. Apart from the great variations of occurrences, ranging from a high of 4.7818 percent for "*e*" to a low of 0.001 percent for some, there are also many nonexistent bigrams, notably the consonant–consonant pairs. It is based on the binary occurrences of bigrams that Riseman *et al.* [60], [61] have accomplished substantial savings in bigram storage and computing time for contextual word recognition and error correction.

The first 50 2-grams, 3-grams, 4-grams, and 5-grams, arranged in descending frequencies, are shown in Table III. Here again, one can see the dominance of high frequency words such as

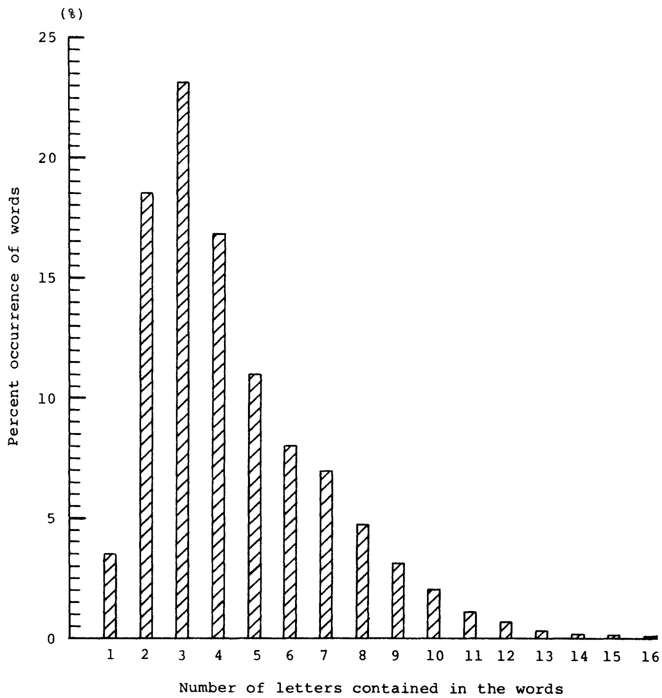

Fig. 1. Word distributions according to length.

TABLE II
PERCENT DISTRIBUTION OF BIGRAMS

First Letter	Second Letter																										
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	space
a		.154	.332	.351	.003	.054	.149	.005	.282	.008	.089	.766	.212	1.605	.003	.125		.819	.795	1.122	.076	.174	.050	.014	.216	.009	.659
b	.112	.006			.502				.066	.015		.166	.002		.165			.072	.024	.011	.170	.005			.138		.020
c	.385		.047		.488			.448	.174		.108	.110			.567		.003	.093	.012	.301	.086				.021		.095
d	.118		.001	.035	.554	.001	.020	.002	.314	.004		.023	.011	.023	.163		.002	.072	.089	.001	.097	.012	.003		.048		2.357
e	.588	.010	.328	.908	.340	.116	.080	.016	.135	.002	.018	.387	.283	1.084	.046	.121	.037	1.568	.891	.320	.010	.217	.109	.144	.136	.002	4.782
f	.135				.183	.109			.213			.038			.426			.176	.001	.070	.070				.003		1.038
g	.106			.001	.275	.001	.016	.224	.106			.034	.003	.044	.121			.146	.033	.013	.052				.008		.601
h	.863	.004	.003		2.965	.001			.742			.007	.004	.022	.425			.066	.007	.130	.057		.002		.029		.610
i	.155	.058	.480	.269	.251	.151	.208				.048	.341	.265	1.835	.515	.054	.009	.252	.893	.920	.005	.191		.015		.033	.134
j	.017				.032				.001						.046			.002			.052						.003
k	.007				.208	.002	.002	.003	.071			.006		.052	.001				.029	.001			.001		.005		.203
l	.343	.002	.004	.264	.611	.047	.001	.001	.431		.022	.491	.020	.003	.284	.015		.009	.094	.077	.084	.022	.012		.347		.713
m	.423	.066	.001		.648	.003			.224			.002	.061	.006	.269	.150		.034	.076	.001	.098		.001		.050		.385
n	.200		.288	1.127	.554	.033	.726	.004	.236	.007	.041	.061	.017	.061	.375	.003	.002	.005	.314	.727	.056	.030	.002	.002	.097	.001	2.000
o	.050	.071	.106	.144	.033	.967	.050	.017	.064	.005	.061	.251	.449	1.276	.209	.166	.001	.993	.214	.348	.818	.149	.296	.008	.029	.002	1.022
p	.226				.345	.001		.045	.083			.199	.014		.248	.102		.316	.033	.057	.072		.001		.005		.135
q																					.098						
r	.417	.012	.075	.142	1.436	.022	.065	.014	.460		.073	.075	.118	.126	.548	.025		.068	.304	.259	.080	.045	.008		.183		1.357
s	.176	.005	.088	.005	.684	.011		.270	.404		.032	.043	.042	.014	.307	.127	.006	.001	.282	.844	.217		.019		.031		2.681
t	.372	.001	.023	.001	.846	.004	.001	3.259	.835			.087	.023	.004	.937	.001		.276	.235	.144	.175		.063		.148	.001	2.223
u	.089	.062	.133	.064	.100	.012	.119		.074		.001	.267	.083	.300	.005	.111		.384	.345	.362	.001			.001	.005	.001	.088
v	.078				.669				.179						.043						.001				.004		.005
w	.433			.003	.317	.001		.358	.346		.001	.009	.001	.078	.214	.001		.025	.023	.004	.001				.002		.219
x	.018		.018		.012			.002	.020						.002	.052				.032	.002				.001		.025
y	.008	.006	.003	.002	.093		.001	.001	.024			.007	.014	.004	.149	.012		.004	.069	.017			.005			.001	1.278
z	.010				.030				.004			.001			.002										.001	.005	.003

the, *of*, *and*, *that*, *with*, *to*, *this*, etc. Frequency prefixes and suffixes are also visible in these tables, e.g., *re*, *ing*, *ion*, *ent*, etc.

The percentage occurrence of n-grams in different positions is exhibited in Fig. 2. The unigram and bigram curves are almost identical. Higher n-grams start at a higher position. The sharp drop in n-grams which occurs between letter positions 2 and 4 is due to the high concentration of word samples in this region.

It might be interesting to note that n-gram statistics are strongly influenced by the vocabulary which is defined as the number of different words in descending order of frequencies. As the vocabulary increases, longer words and rarely used words begin to come in, and one would expect the word length to increase, as well as the total number of different n-grams. This effect is illustrated in Figs. 3 and 4. For $n = 2$, the number of different 2-grams increases from 350, for a vocabulary of

TABLE III
PERCENT OCCURRENCE OF THE TOP 50 2-GRAMS, 3-GRAMS, 4-GRAMS, AND 5-GRAMS

	2-grams		3-grams		4-grams		5-grams	
1	e	4.7818	the	2.9275	the	3.1360	that	.6971
2	th	3.2586	he	2.6470	and	1.3822	tion	.6500
3	he	2.9649	of	1.1734	ing	.9256	with	.4796
4	s	2.6807	nd	1.1554	tion	.6418	ation	.4735
5	d	2.3565	and	1.0549	hat	.5685	ould	.3422
6	t	2.2228	ed	1.0530	ion	.5642	this	.3386
7	n	2.0003	to	.9033	his	.5441	here	.3123
8	in	1.8353	er	.8231	that	.4831	ther	.3061
9	an	1.6045	in	.8035	was	.4398	from	.2874
10	er	1.5681	ng	.7474	for	.4252	have	.2601
11	re	1.4357	is	.7449	ther	.3951	ment	.2501
12	r	1.3572	on	.7440	ent	.3783	they	.2380
13	y	1.2781	ing	.7264	with	.3688	hich	.2343
14	on	1.2760	re	.6772	ere	.3648	which	.2343
15	nd	1.1273	as	.6589	her	.3594	were	.2161
16	at	1.1224	at	.6371	ith	.3365	other	.2138
17	en	1.0843	ion	.5785	atio	.3249	ions	.2117
18	f	1.0375	es	.5590	ted	.2964	there	.2097
19	o	1.0222	or	.5525	ould	.2531	ight	.2022
20	or	.9925	ent	.4994	nce	.2509	would	.1870
21	of	.9674	her	.4763	here	.2395	tions	.1824
22	to	.9367	for	.4707	are	.2363	ction	.1805
23	it	.9201	tio	.4639	ment	.2358	ting	.1801
24	ed	.9080	en	.4390	uld	.2331	their	.1770
25	is	.8932	ly	.4337	this	.2306	heir	.1757
26	es	.8906	hat	.4232	had	.2300	ence	.1729
27	ha	.8630	tha	.4083	ter	.2283	been	.1626
28	te	.8455	his	.4083	not	.2181	when	.1534
29	st	.8436	an	.4038	all	.2175	ally	.1534
30	ti	.8352	al	.3983	one	.2150	ough	.1502
31	ar	.8186	ere	.3780	ave	.2117	more	.1498
32	ou	.8183	st	.3746	out	.2091	hing	.1490
33	as	.7948	nt	.3739	from	.1971	will	.1476
34	al	.7656	th	.3620	but	.1969	ning	.1441
35	hi	.7418	ll	.3530	rom	.1958	thing	.1371
36	nt	.7269	it	.3367	ght	.1888	what	.1339
37	ng	.7258	was	.3323	have	.1800	ding	.1338
38	l	.7129	ce	.3288	ore	.1775	said	.1290
39	se	.6844	ter	.3281	een	.1748	ical	.1247
40	ve	.6685	ut	.3163	ight	.1744	ever	.1236
41	a	.6589	ve	.3151	hen	.1727	state	.1215
42	me	.6476	se	.3053	ver	.1699	ents	.1199
43	le	.6112	ati	.3001	ers	.1689	ring	.1199
44	h	.6101	le	.2961	they	.1685	about	.1194
45	g	.6010	me	.2926	ich	.1646	bout	.1194
46	ea	.5882	all	.2883	ons	.1646	ound	.1185
47	co	.5673	ith	.2856	hey	.1628	into	.1178
48	ne	.5542	ts	.2825	ill	.1614	than	.1177
49	de	.5535	thi	.2808	ough	.1600	them	.1177
50	ro	.5484	ch	.2806	hich	.1596	only	.1168

1000 words, to about 550 for 10 000 words. Higher n-grams increase more rapidly with the vocabulary. In Fig. 4 one can see that the use of 4-grams and 5-grams requires a substantial amount of memory and is, in fact, larger than the number of words from which they are derived. Apart from the regularity of length (n) of the n-grams, there is hardly any saving in storage of n-grams for $n \geqslant 4$ compared with the dictionary.

B. Data 2

n-gram analyses were made on the first 1000 most common words of the various corpuses cited below. They were included in the present study because the first 1000 words form the most important vocabulary of the language and they normally occupy 65–75 percent of the total occurrences of words in the language. They form the basic vocabulary of elementary texts and every day usage of the language and were used in studies on error correction [56].

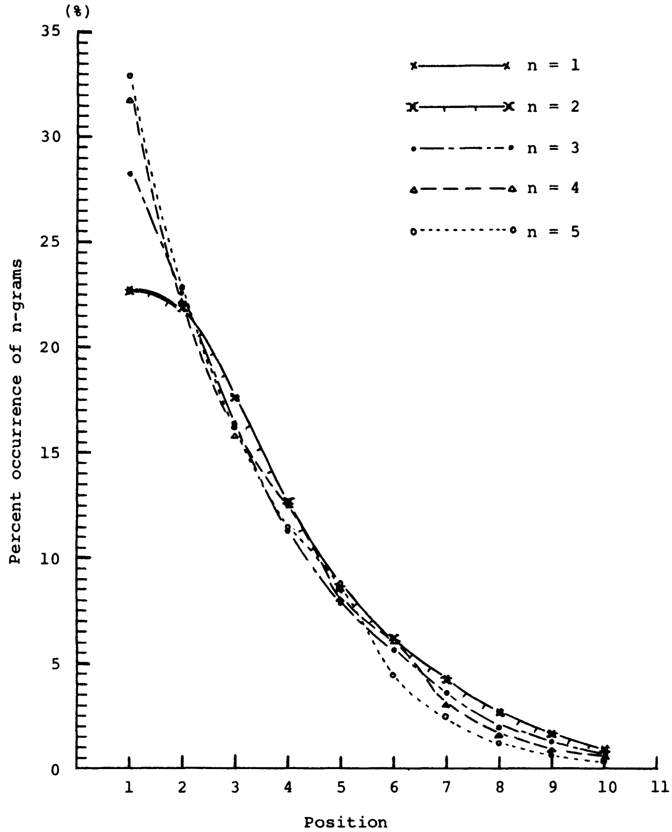

Fig. 2. Percent occurrence of n-grams in different positions.

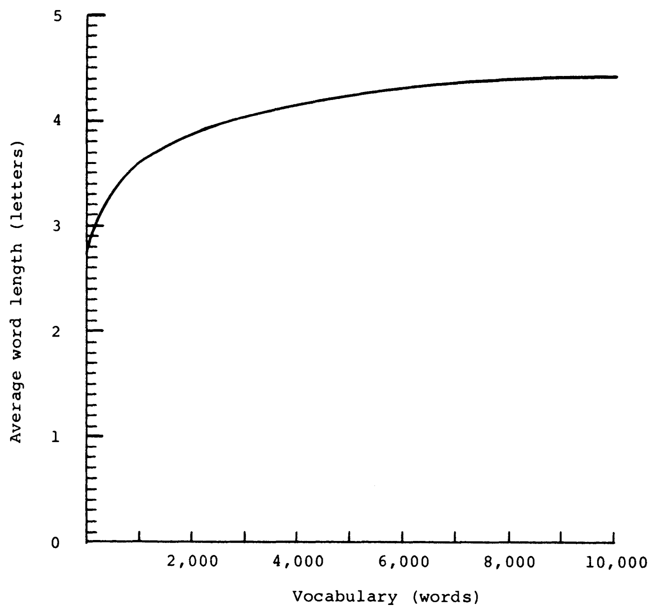

Fig. 3. Variations of word length as a function of vocabulary.

1) Carroll *et al.*

Sample size: 5 088 721 words.

This corpus is composed of 500-word samples taken from 1045 texts selected in 1969. These samples were chosen from

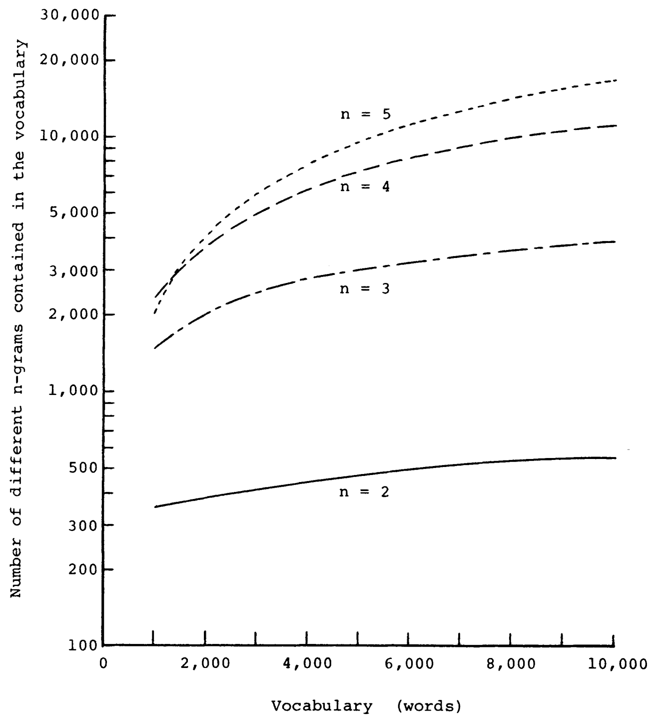

Fig. 4. Variations of total number of different n-grams as a function of vocabulary.

a range of required and recommended texts covering different subjects for students between grade three and grade nine.

2) Dewey

Sample size: 100 000 words.

These words were collected around 1920. They were selected from English as spoken and written at that time including editorials, news, fiction, speeches, correspondence, religious and scientific English, etc. Numerals and proper names were excluded.

3) Kucera *et al.*

See description on Data 1.

4) Thorndike *et al.*

Sample size: 9 million words.

It consists of approximately 4.5 million words of the Thorndike general count of 1931 and the Lorge magazine count, the Thorndike count of 120 juvenile books, and the Lorge-Thorndike semantic count.

Unigram distribution of the first 1000 most frequent words in the above four corpuses are shown in Table IV. Despite the difference in material and date at which the samples were collected, the figures obtained from all four corpuses seem to agree well. The ranks of the different letters are almost identical.

TABLE IV

DISTRIBUTION OF UNIGRAMS OF THE FIRST 1000 MOST FREQUENT WORDS IN (a) CARROLL *et al.*, (b) DEWEY, (c) KUCERA *et al.*, (d) THORNDIKE *et al.*

	(a)	(b)	(c)	(d)
A	8.23	8.23	8.39	8.84
B	1.45	1.58	1.53	1.42
C	1.74	1.78	1.98	1.73
D	3.93	3.44	3.75	3.68
E	13.00	12.40	12.75	12.41
F	2.63	3.08	3.06	2.63
G	1.58	1.47	1.36	1.60
H	8.17	7.70	8.23	8.11
I	5.98	6.49	6.42	6.28
J	0.09	0.07	0.08	0.11
K	0.77	0.55	0.51	0.96
L	3.37	3.41	3.26	3.67
M	2.40	2.37	2.37	2.53
N	6.66	6.93	6.87	6.68
O	8.78	8.89	8.78	8.78
P	1.13	1.29	1.18	1.14
Q	0.04	0.04	0.05	0.05
R	5.06	5.04	5.08	5.13
S	5.52	5.34	5.35	4.73
T	10.75	11.16	11.15	10.46
U	2.60	2.60	2.38	2.77
V	0.70	0.90	0.81	0.85
W	3.04	2.79	2.70	3.12
X	0.07	0.08	0.08	0.07
Y	2.09	2.22	1.73	2.12
Z	0.01	0.01	0.01	0.02
'	0.17	0.13	0.12	0.11

III. AVAILABILITY OF n-GRAM STATISTICS

During the process of analyzing the data contained in the above well-known corpuses, the author also conducted a survey of the n-gram statistics available in the literature. It was noticed that n-grams were computed from written texts as well as from transcriptions of speech. Since this paper is more concerned with text rather than speech processing, only statistical distributions obtained from written texts are presented. Some analyses were made on general texts and some were made under certain constraints. Both kinds of analyses will be described below. Unigrams derived from them, or with permission to publish them, are tabulated in Tables V and VI, respectively, for the two categories.

A. General Texts

Various statistical data on n-grams can be found in the literature. This section provides a brief description of the data found or processed by the author.

1) Baddeley and Conrad [6].

Sample size: 13 000 words (76 150 letters and spaces).

This data base consists of unigrams and bigrams of the 26 letters plus space computed from all words which appeared in the editorial columns of *The Times* newspaper for five successive days in 1960. Names of persons and foreign place names were excluded.

2) Casey and Nagy [16]

Sample size: 600 000 characters.

Bigram probabilities observed from 600 000 characters of identified text were presented.

3) Dewey [22]

Sample size: 100 000 words.

Unigram frequencies derived from 100 000 words were presented.

4) Gaines [31]

Sample size: 10 000 letters.

TABLE V
DISTRIBUTIONS OF UNIGRAMS ON GENERAL TEXTS: (1) BADDELEY *et al.*, (2) CASEY *et al.*, (3) AFTER DEWEY, (4) AFTER GAINES, (5) NEWMAN *et al.*, (6) PRATT, (7) SHINGHAL, (8) SUEN, (9) UNDERWOOD *et al.*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
A	7.81	7.74	7.88	7.81	8.63	8.15	7.80	8.07	7.98
B	1.72	1.35	1.56	1.28	1.45	1.44	1.51	1.47	1.55
C	2.78	3.92	2.68	2.93	1.66	2.76	3.25	2.94	2.79
D	3.28	3.96	3.89	4.11	4.64	3.79	3.72	3.96	3.97
E	13.08	12.68	12.68	13.05	12.53	13.11	12.47	12.70	12.57
F	2.29	2.71	2.56	2.88	2.52	2.92	2.47	2.46	2.20
G	1.76	1.52	1.87	1.39	1.51	1.99	1.79	1.79	2.11
H	5.60	4.89	5.73	5.85	8.61	5.26	5.36	5.94	5.30
I	7.25	7.57	7.07	6.77	5.82	6.35	7.68	7.10	7.36
J	0.11	0.27	0.10	0.23	0.17	0.13	0.19	0.15	0.18
K	0.47	0.30	0.60	0.42	0.74	0.42	0.54	0.59	0.73
L	4.26	3.50	3.94	3.60	5.01	3.39	4.05	3.90	4.07
M	2.58	1.94	2.44	2.62	2.37	2.54	2.50	2.50	2.58
N	7.00	7.47	7.06	7.28	6.48	7.10	7.10	7.03	7.10
O	7.63	7.76	7.76	8.21	7.10	8.00	7.65	7.80	7.34
P	2.05	2.48	1.86	2.15	1.58	1.98	2.09	1.88	1.93
Q	0.12	0.11	0.09	0.14	0.04	0.12	0.14	0.10	0.09
R	6.03	6.40	5.94	6.64	5.93	6.83	6.05	5.92	6.21
S	6.36	6.36	6.31	6.46	5.63	6.10	6.67	6.29	6.75
T	10.26	10.27	9.78	9.02	10.27	10.47	9.35	9.68	8.82
U	2.50	2.73	2.80	2.77	2.67	2.46	2.71	2.61	3.07
V	1.16	0.98	1.02	1.00	0.72	0.92	1.09	0.98	1.01
W	1.93	1.43	2.14	1.49	2.14	1.54	1.75	2.04	2.03
X	0.20	0.27	0.16	0.30	0.04	0.17	0.23	0.18	0.19
Y	1.75	1.34	2.02	1.51	1.71	1.98	1.74	1.71	1.96
Z	0.03	0.06	0.06	0.09	0.04	0.08	0.11	0.06	0.10

TABLE VI
UNIGRAM DISTRIBUTIONS DERIVED FROM SPECIFIC WORD SAMPLES: (1) BOURNE *et al.* ON SUBJECT WORDS, (2) BOURNE *et al.* ON PROPER NAMES, (3) MAYZNER *et al.*, (4) AFTER SOLSO *et al.*

	(1)	(2)	(3)	(4)
A	9.07	10.74	8.10	7.61
B	1.62	2.21	1.63	1.54
C	4.73	3.20	2.36	3.11
D	2.65	3.72	4.32	3.95
E	9.90	11.29	13.32	12.62
F	1.51	0.99	1.79	2.34
G	1.76	2.10	2.18	1.95
H	2.51	4.24	7.72	5.51
I	9.93	6.31	5.15	7.34
J	0.15	1.89	0.15	0.15
K	0.46	1.38	1.07	0.65
L	5.19	7.39	4.47	4.11
M	3.26	3.32	2.48	2.54
N	6.50	7.61	6.01	7.11
O	9.36	6.51	6.63	7.65
P	3.61	1.31	1.53	2.03
Q	0.15	0.03	0.08	0.10
R	8.17	9.93	5.89	6.15
S	4.81	4.66	6.07	6.50
T	8.16	4.30	9.78	9.33
U	3.32	2.18	3.09	2.72
V	0.97	0.85	0.99	0.99
W	0.61	1.94	2.87	1.89
X	0.26	0.03	0.14	0.19
Y	1.04	1.59	2.12	1.72
Z	0.29	0.24	0.06	0.09

This data base consists of unigrams, bigrams, and some trigrams obtained by O. P. Meaker from 10 000 letters.

5) Newman and Gerstman [54]

Sample size: 10 000 letters, spaces, and periods.

Unigrams and trigrams were derived from the *Bible* and consisted of the larger part of Isaiah XXIX–XXXI in the King James version. No space was used following a period; other punctuations were disregarded.

6) Pratt [57]

Sample size: varied.

Unigram and bigram frequencies were obtained from 1000 words while trigram frequencies were obtained from 20 000 words. In bigram statistics, divisions between words were respected.

7) Shinghal [69]

Sample size: approximately 530 000 words.

Frequencies on unigrams were presented. The 530 000 samples were composed of texts taken from ten different subjects including children's literature, law, music, medicine, psychology, religious scriptures, newspapers and periodicals, military science, management science, and philosophy. Punctuations, special symbols, and numerics were omitted. Word length distributions were also presented.

8) Underwood and Schulz [96]

Sample size: approximately 15 000 words.

Unigrams, bigrams and trigrams were hand tabulated from approximately 15 000 words of written passages. All words were used, including contractions. Apostrophes were deleted. Hyphenated words were considered to be two independent words. The passages were selected from novels, short stories, advertisements, magazines, newspapers, encyclopedias, and so on.

Unigram distributions of the above sources and those obtained from Data 1 are exhibited in Table V. It can be seen that the distributions vary more than those shown in Table IV. These perturbations were due to many reasons such as the type of material used, the date the samples were collected, as well as methods of analyses. Distributions of high frequency letters such as E, T, A, O, I, N, and S are more stable, while those of the lower ones such as K, Q, X, and Z very more.

The reader is referred to the references cited for statistics of higher n-grams.

B. Specific Samples

Apart from the above, n-gram statistics have been obtained by others using more specific samples of constraints.

1–2) Bourne *et al.* [11]

This data base consists of unigrams and bigrams obtained from 2082 subject words (16 913 letters) and 8207 proper names (141 190 letters). Distributions of letters by positions

(1-10) were tabulated. Plots of word lengths were also presented.

3) Mayzner *et al.* [44], [45]

This data base provides unigram, bigram, and trigram frequencies of 20 000 words composed of 3, 4, 5, 6, and 7 letters. Frequency distributions were tabulated according to word lengths and letter positions.

4) Solso *et al.* [74]

Unigram statistics for the 26 letters of the alphabet were prepared using Kucera *et al.*'s 1 million word data base. However, all hyphenated words and words containing apostrophes were omitted.

5) Topper *et al.* [90]

This data base comprised bigram statistics obtained from 1000 words taken from Thorndike and Lorge's lists. Corresponding frequencies for bigrams were selected from Underwood *et al.* [96].

Unigram distributions for these samples are presented in Table VI. Since very different materials were used in these analyses, substantial variations are observed in the distributions, e.g., 1) *A* occurs 10.74 percent in proper names computed in (2) and 7.61 percent in (4), 2) *E* occurs 13.32 percent in (3) and only 9.90 percent in (1), etc.

IV. Concluding Remarks

Statistics of the combination of letters (n-grams) and dictionaries have been employed by many researchers dealing with automatic correction of deletion, substitution, and insertion errors in text processing. This type of application is very useful because it offers substantial savings in human efforts. Even if the machine could simply indicate the possible errors, it would still be a great help in character and speech recognition, machine understanding and translation systems, dictionary and textbook preparations, automatic mail sorting, patent and reference searches, and various information retrieval applications. The use of n-gram statistics for these applications is a very attractive tool compared with the dictionary method which requires much greater storage and computing time. The tables and figures presented to show the distributions of n-grams, n-grams against letter position, word length, and vocabulary can be used by designers to find out the kind of statistics most suitable for their applications.

As demonstrated by the figures shown in Tables IV-VI, n-gram statistics vary from text to text. Also, statistics on high frequency words tend to be more reliable than on low frequency ones since the statistics on the latter can be easily altered when a slightly different type of text sample has been chosen for analysis. Thus, one cannot optimize the solution by plugging in a random set of n-gram statistics for processing any kind of texts and languages. In order that n-gram statistics may be used effectively in text processing and natural language understanding systems, the "context" of the text involved, whether novel, proper names, addresses, news items, computer programs, etc., must be taken into consideration.

During the past two decades, there has been considerable interest in computational analyses and machine understanding of natural languages. However, due to the complexity of the problem, present machines which make use of semantics and syntax are still limited. It appears that more effects should be put in this direction. Error-correction techniques using contextual information, combined with OCR technology, can form a very powerful tool for the collection of copious language units, such as texts, phrases, lexical, and grammatical units, the lack of which is one of the major impediments to the historical study of languages. The effect of such techniques are far-reaching. With joint efforts of computer scientists, linguists, psychologists, engineers, and others interested in interdisciplinary studies, there is little doubt that machine intelligence in the understanding of natural languages and text processing will come closer and closer to that of human beings.

Acknowledgment

The author would like to thank the following persons and organizations: 1) Dover Publications, Inc., for permission to publish the percent occurrence of English letters in H. F. Gaines' book, 1956; 2) Harvard University Press for permission to publish the percent occurrence of letters in English in G. Dewey's book, 1923; 3) Prof. R. L. Solso for permission to publish the percent occurrence of English letters in Solso *et al.*'s article, 1976; 4) Harvard University Press for permission to use the frequency data published in G. Dewey's book, 1923; 5) Brown University Press for permission to use the frequency data published in Kucera *et al.*'s book, 1967; 6) Houghton Mifflin Company, for permission to use the frequency data published in Carroll *et al.*'s book, 1971; and 7) Teacher College Press of Columbia University for permission to use Thorndike *et al.*'s book, 1944. Thanks are also due to L. Thiel and G. W. Gillespie for programming assistance.

References

[1] M. Aborn, H. Rubenstein, and T. D. Sterling, "Sources of contextual constraint upon words in sentences," *J. Exp. Psych.*, vol. 57, pp. 171-180, 1959.

[2] C. N. Alberga, "String similarity and misspellings," *Commun. Ass. Comput. Mach.*, vol. 10, pp. 302-313, May 1967.

[3] R. Alter, "Utilization of contextual constraints in automatic speech recognition," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 6-11, Mar. 1968.

[4] R. N. Ascher, G. M. Koppelman, M. J. Miller, G. Nagy, and G. L. Shelton, Jr., "An interactive system for reading unformatted printed text," *IEEE Trans. Comput.*, vol. C-20, pp. 1527-1543, Dec. 1971.

[5] R. Attar, Y. Choueka, N. Dershowitz, and A. S. Fraenkel, "KEDMA–Linguistic tools for retrieval systems," *J. Ass. Comput. Mach.*, vol. 25, pp. 52-66, Jan. 1978.

[6] A. D. Baddeley and R. Conrad, "Letter structure of the English language," *Nature*, vol. 186, pp. 414-416, Apr. 1960.

[7] L. R. Bahl and F. Jelinek, "Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 404-411, July 1975.

[8] L. Baker and J. L. Santa, "Context, integration, and retrieval," *Memory and Cognition*, vol. 5, pp. 308-314, May 1977.

[9] C. R. Blair, "A program for correcting spelling errors," *Inform. Contr.*, vol. 3, pp. 60-67, 1960.

[10] W. W. Bledsoe and I. Browning, "Pattern recognition and reading by machine," in *Proc. Eastern Joint Comput. Conf.*, 1959, pp. 225-232.

[11] C. P. Bourne and D. F. Ford, "A study of the statistics of letters in English words," *Inform. Contr.*, vol. 4, pp. 48-67, 1961.

[12] G. Carlson, "Techniques for replacing characters that are garbled on input," in *Proc. Spring Joint Comput. Conf.*, 1966, pp. 189-192.

[13] J. B. Carroll, "Word-frequency studies and the lognormal distri-

bution," in *Lanugage and Language Behavior*, E. M. Zale, Ed. New York: Appleton-Century-Crofts, 1968, pp. 213-235.

[14] J. B. Carroll, P. Davis, and B. Richman, *Word Frequency Book*. New York: American Heritage, 1971.

[15] E. C. Carterette and M. H. Jones, "Redundancy in children's texts," *Science*, vol. 140, pp. 1309-1311, 1963.

[16] R. G. Casey and G. Nagy, "An autonomous reading machine," *IEEE Trans. Comput.*, vol. C-17, pp. 492-503, May 1968.

[17] C. S. Christensen, "An investigation of the use of context in character recognition using graph searching," Center for Applied Math., Cornell Univ., Ithaca, NY, Tech. Rep. AFOSR 68-2470, Nov. 1968.

[18] C. Conrad, "Context effects in sentence comprehension: A study of the subjective lexicon," *Memory and Cognition*, vol. 2, pp. 130-138, 1974.

[19] R. W. Cornew, "A statistical method of spelling correction," *Inform Contr.*, vol. 12, pp. 79-93, 1968.

[20] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. Ass. Comput. Mach.*, vol. 7, pp. 171-176, Mar. 1964.

[21] L. Davidson, "Retrieval of misspelled names in an airlines passenger record system," *Commun. Ass. Comput. Mach.*, vol. 5, pp. 169-171, 1962.

[22] G. Dewey, *Relative Frequency of English Speech Sounds*. Cambridge: Harvard Univ. Press, 1923.

[23] R. W. Donaldson and G. T. Toussaint, "Use of contextual constraints in recognition of contour-traced handprinted characters," *IEEE Trans. Comput.*, vol. C-19, pp. 1096-1099, Nov. 1970.

[24] W. Doster, "Contextual postprocessing system for cooperation with a multiple choice character recognition systems," *IEEE Trans. Comput.*, vol. C-26, pp. 1090-1101, Nov. 1977.

[25] R. O. Duda and P. E. Hart, "Experiments in the recognition of hand-printed text: Part II–Context analysis," in *Proc. Fall Joint Comput. Conf.*, 1968, pp. 1139-1149.

[26] A. W. Edwards and R. L. Chambers, "Can *a priori* probabilities help in character recognition?" *J. Ass. Comput. Mach.*, vol. 11, pp. 465-470, Oct. 1964.

[27] R. W. Ehrich and K. J. Koehler, "Experiments in the contextual recognition of cursive script," *IEEE Trans. Comput.*, vol. C-24, pp. 182-194, Feb. 1975.

[28] R. W. Ehrich, "A contextual post-processor for cursive script recognition–A summary," in *Proc. 1st Int. Joint Conf. Pattern Recognition*, 1973, pp. 169-171.

[29] E. G. Fisher, "The use of context in character recognition," Ph.D. dissertation, Dep. Comput. Sci., Univ. Massachusetts, Amherst, Mar. 1976.

[30] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268-278, Mar. 1973.

[31] H. F. Gaines, *Cryptanalysis*. New York: Dover, 1956, pp. 218-221.

[32] J. J. Giangardella, J. F. Hudson, and R. S. Roper, "Spelling correction by vector representation using a digital computer," *IEEE Trans. Eng., Writing, Speech*, vol. EWS-10, pp. 57-62, Dec. 1967.

[33] P. M. Hahn and N. C. Randall, "A best-match file search procedure for postal directions," in *Proc. Int. Conf. Cybern. Soc.*, Oct. 1972, pp. 512-518.

[34] A. R. Hanson, E. M. Riseman, and E. Fisher, "Context in word recognition," *Pattern Recognition*, vol. 8, pp. 35-45, 1976.

[35] L. D. Harmon and E. J. Sitar, "Method and apparatus for correcting errors in mutilated text," U.S. Patent 3 188 609, June 1965.

[36] G. Herdan, *The Advanced Theory of Language as Choice and Chance*. Berlin: Springer-Verlag, 1966.

[37] Y. Hoshino, "Word recognition apparatus," U.S. Patent 4 010 445, Mar. 1977.

[38] A. B. S. Hussain and R. W. Donaldson, "Suboptimal sequential decision schemes with on-line feature ordering," *IEEE Trans. Comput.*, vol. C-23, pp. 582-590, June 1974.

[39] E. B. James and D. P. Partridge, "Adaptive correction of program statements," *Commun. Ass. Comput. Mach.*, vol. 16, pp. 27-37, Jan. 1973.

[40] P. A. Kolers and M. T. Katzman, "Naming sequentially presented letters and words," *Language and Speech*, vol. 9, pp. 84-95, 1966.

[41] E. Korolev, "On automatic recognition of context," in *Proc. Int. Conf. Comput. Ling.*, 1971.

[42] H. Kucera and W. N. Francis, *Computational Analysis of Present-Day American English*. Providence: Brown Univ. Press, 1967.

[43] V. I. Levenstein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Phys.–Doklady*, vol. 10, pp. 707-710, 1966.

[44] M. S. Mayzner and M. E. Tresselt, "Tables of single-letter and digram frequency counts for various word-length and letter position combinations," *Psychonomic Monograph Suppl.*, vol. 1, pp. 13-32, 1965.

[45] M. S. Mayzner, M. E. Tresselt, and B. R. Wolin, "Tables of trigram frequency counts for various word-length and letter position combinations," *Psychonomic Monograph Suppl.*, vol. 1, pp. 33-78, 1965.

[46] G. I. McCalla and J. R. Sampson, "MUSE: A model to understand simple English," *Commun. Ass. Comput. Mach.*, vol. 15, pp. 29-40, Jan. 1972.

[47] C. K. McElwain and M. B. Evens, "The degarbler–A program for correcting machine-read Morse code," *Inform. Contr.*, vol. 5, pp. 368-384, 1962.

[48] J.-G. Meunier, S. Rolland, and F. Daoust, "A system for text and content analysis," *Comput. and the Humanities*, vol. 10, pp. 281-286, Sept./Oct. 1976.

[49] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test materials," *J. Exp. Psych.*, vol. 41, pp. 329-335, 1941.

[50] G. A. Miller and E. A. Friedman, "The reconstruction of mutilated English texts," *Inform. Contr.*, vol. 1, pp. 38-55, 1957.

[51] H. L. Morgan, "Spelling correction in systems programs," *Commun. Ass. Comput. Mach.*, vol. 13, pp. 90-94, Feb. 1970.

[52] R. Morris and L. L. Cherry, "Computer detection of typographical errors," *IEEE Trans. Prof. Commun.*, vol. PC-18, pp. 54-64, Mar. 1975.

[53] D. L. Neuhoff, "The Viterbi algorithm as an aid in text recognition," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 222-226, Mar. 1975.

[54] E. B. Newman and L. J. Gerstman, "A new method for analyzing printed English," *J. Exp. Psych.*, vol. 44, pp. 114-125, 1952.

[55] E. B. Newman and N. C. Waugh, "The redundancy of texts in three languages," *Inform. Contr.*, vol. 3, pp. 141-153, 1960.

[56] T. Okuda, E. Tanaka, and T. Kasai, "A method for the correction of garbled words based on the Levenshtein metric," *IEEE Trans. Comput.*, vol. C-25, pp. 172-178, Feb. 1976.

[57] F. Pratt, *Secret and Urgent–The Story of Codes and Ciphers*. Indianapolis: Bobbs-Merril, 1939, pp. 252-278.

[58] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-3, pp. 536-551, Oct. 1967.

[59] C. Rieger, "An organization of knowledge for problem solving and language comprehension," *Artificial Intelligence*, vol. 7, pp. 89-127, 1976.

[60] E. M. Riseman and R. W. Ehrich, "Contextual word recognition using binary bigrams," *IEEE Trans. Comput.*, vol. C-20, pp. 397-403, Apr. 1971.

[61] E. M. Riseman and A. R. Hanson, "A contextual postprocessing system for error correction using binary *n*-grams," *IEEE Trans. Comput.*, vol. C-23, pp. 480-493, May 1974.

[62] W. S. Rosenbaum and J. J. Hilliard, "Multifont OCR postprocessing systems," *IBM J. Res. Develop.*, vol. 19, pp. 398-421, July 1975.

[63] R. Rustin, Ed., *Natural Language Processing*. New York: Algorithmics, 1973.

[64] G. Salton and A. Wong, "On the role of words and phrases in automatic text analysis," *Comput. and the Humanities*, vol. 10, pp. 69-87, Mar./Apr. 1976.

[65] J. Schurmann, "Multifont word recognition system with application to postal address reading," in *Proc. 3rd Int. Joint Conf. Pattern Recognition*, Nov. 1976, pp. 658-662.

[66] V. Seth, "An approach to address identification from degraded address data," in *Proc. AFIPS*, June 1977, pp. 779-783.

[67] C. Shannon, "Prediction and entropy of printed English," *Bell Sys. Tech. J.*, vol. 30, pp. 50-64, 1951.

[68] M. Shimura, "Recognition machines with parametric and nonparametric learning methods using contextual information," *Pattern Recognition*, vol. 5, pp. 149-168, 1973.

[69] R. Shinghal, "Using contextual information to improve performance of character recognition machines," Ph.D. dissertation,

School of Comput. Sci., McGill Univ., Montreal, P.Q., Canada, Mar. 1977.
[70] R. Shinghal, D. Rosenberg, and G. T. Toussaint, "A simplified heuristic version of Raviv's algorithm for using context in text recognition," in *Proc. 5th Int. Joint Conf. Artificial Intelligene*, vol. 1, Aug. 1977, pp. 179-180.
[71] R. Shinghal, D. Rosenberg, and G. T. Toussaint, "A simplified heuristic version of a recursive Bayes algorithm for using context in text recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, pp. 412-414, May 1978.
[72] G. Silva and H. Love, "The identification of variant texts by computer," *Inform. Storage and Retrieval*, vol. 5, pp. 89-108, 1969.
[73] F. Smith, "The use of featural dependencies across letters in the visual identification of words," *J. Verb. Learn. Verb. Behav.*, vol. 8, pp. 215-218, 1969.
[74] R. L. Solso and J. King, "Frequency and versatility of letters in the English language," *Behav. Res. Meth. & Instrum.*, vol. 8, pp. 283-286, 1976.
[75] C. Y. Suen and M. P. Beddoes, "Development of a digital spelled-speech reading machine for the blind," *IEEE Trans. Bio-Med. Eng.*, vol. BME-20, pp. 452-459, Nov. 1973.
[76] C. Y. Suen, M. P. Beddoes, and J. C. Swail, "The Spellex system of speech aids for the blind," in *Proc. AFIPS*, June 1976, pp. 217-220.
[77] C. Y. Suen, M. Berthod, and S. Mori, "Advances in recognition of handprinted characters," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, Nov. 1978.
[78] C. Y. Suen, C. Shiau, R. Shinghal, and C. C. Kwan, "Reliable recognition of handprinted characters," in *Proc. Joint Workshop Pat. Recog. & Art. Intel.*, June 1976, pp. 98-102.
[79] C. Y. Suen and R. J. Shillman, "Low error rate optical character recognition of unconstrained handprinted characters based on a model of human perception," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 491-495, June 1977.
[80] M. Symonds, "Computer detection of misspelled words," Nat. Res. Council of Canada, Radio and Elec. Eng. Div., Ottawa, Tech. Rep. ERB-843, Aug. 1970.
[81] A. J. Szanser, "Error-correcting methods in natural language processing," in *Proc. IFIP*, vol. 2, 1969, pp. 1412-1416.
[82] —, "Resolution of ambiguities by contextual word repetition," *Rev. Appl. Ling.*, vol. 7, pp. 49-56, 1970.
[83] —, "Automatic error correction in natural languages," *Inform. Storage and Retrieval*, vol. 5, pp. 169-174, 1970. Also, *Statis. Meth. Ling.*, vol. 6, pp. 52-59, 1970.
[84] —, "Automatic error correction in natural texts," Nat. Physical Lab., England, Tech. Rep. Com Sci 53 and 63, Dec. 1971 and Jan. 1973.
[85] E. Tanaka and T. Kasai, "A correcting method of garbled languages using ordered key letters," *Electron. Commun. Japan*, vol. 55-D, pp. 127-133, 1972.
[86] R. B. Thomas and M. Kassler, "Character recognition in context," *Inform. Contr.*, vol. 10, pp. 43-64, 1967.
[87] R. A. Thompson, "Language correction using probabilistic grammers," *IEEE Trans. Comput.*, vol. C-25, pp. 275-286, Mar. 1976.
[88] L. E. Thorelli, "Automatic correction of errors in text," *BIT*, vol. 2, pp. 45-62, 1962.
[89] E. L. Thorndike and I. Lorge, *The Teacher's Word Book of 30,000 Words.* New York: Teachers College Press, 1944.
[90] G. E. Topper, W. H. Macey, and R. L. Solso, "Bigram versatility and bigram frequency," *Behav. Res. Meth. & Instrum.*, vol. 5, pp. 51-53, 1973.
[91] G. T. Toussaint and R. W. Donaldson, "Some simple contextual decoding algorithms applied to recognition of handprinted text," in *Proc. Canadian Comput. Conf.*, 422102-422116, 1972.
[92] G. T. Toussaint, "Recent progress in statistical methods applied to pattern recognition," in *Proc. 2nd Int. Joint Conf. Pattern Recognition*, Aug. 1974, pp. 479-488.
[93] —, "The use of context in pattern recogntion," in *Proc. Conf. Pattern Recognition and Image Processing*, June 1977, pp. 1-10.
[94] E. Tulving and C. Gold, "Stimulus information and contextual information as determinants of tachistoscopic recognition of words," *J. Exp. Psych.*, vol. 66, pp. 319-327, 1963.
[95] J. R. Ullmann, "A binary *n*-gram technique for automatic correction of substitution, deletion, insertion and reversal errors in words," *Comput. J.*, vol. 20, pp. 141-147, May 1977.
[96] B. J. Underwood and R. W. Schulz, *Meaningfulness and Verbal Learning.* Chicago: Lippincott, 1960.
[97] C. M. Vossler and N. M. Branston, "The use of context for correcting garbled English text," in *Proc. Nat. Ass. Comput. Mach. Conf.*, D2.4-1-D2.4-13, Aug. 1964.
[98] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *J. Ass. Comput. Mach.*, vol. 21, pp. 168-173, 1974.
[99] J. Weizenbaum, "Contextual understanding by computer," *Commun. Ass. Comput. Mach.*, vol. 10, pp. 474-480, 1967.
[100] Y. Wilks, "A preferential, pattern-seeking, semantics for natural language inference," *Artifical Intelligence*, vol. 6, pp. 53-74, 1975.
[101] T. Winograd, *Understanding Natural Languages.* New York: Academic, 1972.
[102] G. K. Zipf, *Human Behavior and the Principle of Least Effort.* Cambridge, MA: Addison-Wesley, 1949.

Ching Y. Suen (M'66-SM'78) received the M.Sc. (Eng.) degree from the University of Hong Kong and the Ph.D. degree from the University of British Columbia, both in electrical engineering.

In 1972 he joined the Engineering Faculty of Concordia University, Montreal, P.Q., Canada, where currently he is an Associate Professor in the Department of Computer Science. He had been a visitor of the Character Recognition Group of the Research Laboratory of Electronics of M.I.T. on several occasions and a visitor of the Pattern Recognition Group of the Institut de Recherche d'Informatique et d'Automatique, France. He has participated in the development of several Canadian Standards on optical character recognition and is currently Chairman of the Character Recognition Committee of the Canadian Standards Association. He is the author/coauthor of 60 papers dealing with transistor circuits and electronics, speech analysis and synthesis, perception and psychophysics, electronic aids for the visually handicapped, handwriting education, and character recognition. His current interests include character analysis and recognition, language properties and text processing, speech analysis and synthesis, and mini- and microcomputer applications.

Dr. Suen served in the capacities of Chairman and Program Chairman of the IEEE Montreal Chapter on Biomedical Engineering and was the Program Chairman of two Symposiums on Biomedical Engineering.