SOUND AND SYLLABLE DISTRIBUTION IN WRITTEN KANNADA AND THEIR APPLICATION TO SPEECH AND HEARING

M. JAYARAM*

The purpose of the present study was to compute the distribution of sounds and syllables in written Kannada. Simple frequency counts of different sounds (letters) and syllables were made from an analysis of more than 1 lakh words. The written material selected were news papers and popular magazines. The analysis showed the occurrence of 896 different syllables in Kannada with varying frequency of occurrence. The potential application of these statistics to (1) Communication engineering and (2) in the assesment, diagnosis, research and rehabilitation of individuals with speech, hearing and language pathology has been enumerated in the last section of the paper.

General purpose of such studies

Recent technical developments hold considerable promise of producing mechanically, in conventional Orthography, the sounds and words spoken into the machine. Also, machines that speak and recognise speech has long been a dream of commnication engineers. The realization of these machines, be it a phonetic typewriter or a computer which responds to spoken words, depends upon the availability of a great deal of information about the acoustical characteristics of sounds in each language spoken by its community. Furthermore, the possibility of translation from one language into another by computer, points out the need for the study of languages from the point of view appropriate to machine handling. Besides, there are the classical problems of designing efficient teleprinters in Indian languages, design of ciphers in Indian languages and the extent of secrecy possible with them. These data may also be of use in automatic generation of speech for a reading machine for the blind. A scientific study of such problems has to begin with the statistical analysis of a language in its spoken as well as written forms.

The present study is concerned with the statistical analysis of Kannada language for certain language parameters. Such statistics of language parameters

• NIMHANS Bangalore.

vary depending upon whether they were obtained form spoken form or written texts Statistics given here are the result of the analysis of written texts of Kannada.. These statistics, it is hoped, would be of interest and relevance not only to communication engineer but also speech-language pathologists and audiologists as well.

Relative Frequencies of Speech Sounds

Simple frequency counts of all speech sounds (Phonemes) occurring in a given written material were made. The material used for such frequency counts were-from books, magazines and news papers, representative of the language, and published during the period 1974-1976. Before proceeding further, it must be pointed out that the values of all the statistics are subject to the limitations which are inherent in the sampling methods employed. In particular, the statistics vary to some extent with the type of material employed for analysis (text books, magazines, news papers, journals); variety of speech employed (conversational speech or written materials) ; and period of publication of the material analysed. Another factor that could have influenced the results has to with the fact that all the printed materials analysed in the present study were all published in the south interior Karnataka. However, it must also be pointed out here that the statistics given in the succeeding pages have been arrived at by analysing a large material published over a period of 3 years (Table 1), The statistics arrived are the result of analysis of little more than 1 lakh words and the frequency counts of speech sounds have been made extended over nearly 8 lakh written speech letters. Thus, it is believed that the statistics given here for printed material have an inherent validity and reliability in them.

TABLE I

The total number of words, syllables and speech sounds analysed to arrive at the statistics reported in this paper.

Linguistic Unit	Total Number Analysed
1. Sentences	14,610
2. Words	1,00,078
3. Syllables	3,62,282
4. Speech Sounds	7,60,792
. 5. Syllables per Word	3.62
6. Speech Sounds per word	7.6
7. Speech Sounds per Syllable	2.1
8. Words per sentence	6.85

The relative frequencies of different speech sounds in Kannada are given in Table 2. As remarked earlier, these values should not be taken as absolute or final. Some departures from these figures should be expected depending upon the size and nature of the sample material. Another fact about Table 2 should be mentioned here : In column 2 are given the relative frequency of occurrence of speech sounds when Kannada was considered to have 50 letter alphabet. In column 4 are given the relative frequency of speech sounds when Kannada was considered to have a 51 letter alphabet. The extra letter in the alphabet occurred because of the inclusion of word space as equivalent to a speech sound. Considering a word space as equivalent to a speech sound would be resorted to by communication engineers. Such a scheme will have application in telegraphic transmission and design of a common telegraphic code for a number of languages (Ramakrishna et al, 1962).

TABLE 2

Frequency of occurrence of different sounds in the Kannada language. Very frequently it would be necessary to consider word space as equivalent to the letter for communication purposes. Therefore, the columns 4 & 5 give the frequency of occurrence of sounds when space was considered equivalent to a sound thus making a 51 letter alphabet for Kannada.

Sound	(Space not counted)		(Space counted)	
	Frequency (Percentage)	Rank Order.	Frequency (Percentage)	Rank Order.
1	2	3	4	5
ʌ	19.04	1	16.81	1
	4.92	7	4.34	8
i	7.50	2	6.62	3
i:	0.85	26	0.75	27
u	5.61	5	4.95	6
u :	0.79	27	0.70	28
e	4.33	8	3.82	9
ae	1.53	18	1.34	19
ai	0.29	33.5	0.26	34.5
o	0.94	25	0.83	26
ou	0.49	31	0.43	32
au	0.07	37	0.06	38
k	2.52	14	2.22	15
	0.15	36	0.13	37
g	3.08	12	2.71	13
g^h	0.06	38	0.05	39
jn	0.03	39.5	0.03	40
tf	0.68	28	0.6	29

Sound	(Space not counted)		(Space counted)	
	Frequencp (Percentage)	Rank Order.	Frequency (Percentage)	Rank Order.
1	2	3	4	5
tf^h	0.03	39.5	0.02	42
j	0.65	29	0.58	30
j^h	0.002	43.5	0.002	44.5
t	1.08	24	0.95	25
t^h	0.02	41.5	0.02	42
d	1.24	21	1.09	22
d^h	0.002	43.5	0.002	44.5
n	0.55	30	0.49	31
t	3.78	9 .	3.33	10
t^h	0.23	35	0.20	36
d	5.54	6	4.89	7
d^h	0.46	32	0.41	33
n	6.99	3	6.17	4
P	1.25	20	1.10	21
P^h	0.02	41.5	0.02	42
b	1.13	22	0 .99	23
b^h	0.29	33.5	0.26	34.5
m	1.95	16	1.72	17
y	2.23	15	1.96	16
r	6.04	4	5.33	5
1	3.34	10	2.95	11
v	3.15	11	2.78	12
	1.10	23	0.97	24
s	2.65	13	2.34	14
h	1.29	19	1.13	20
	1.61	17	1.42	18
Space	—	-	12.09	2

We may draw attention to certain features concerning the frequencies of the speech sounds. Notice, for instance, that the short vowels a, i, u, etc. are by far the more frequent compared to their longer counterparts. A preponderance of unaspirated consonants like t, d, over their aspirated form like th; dh, etc. again illustrates the operation of the Zipf's principle of least effort (1949). The long vowels which consume more time and the aspirated consonants which demand more effort as well as more time are used less frequently. In fact, we can take advantage of this statistical result in testing the relationship between phonetic factors on the one hand and stuttering and misarticulation on the other hand. Similarly, this result would be pertinent in phonological analysis of the language of an aphasic patient.

Relative Frequencies of different types of Syllables :

The relative frequencies of the different syllable types and syllables would be considerable practical interest. Such statistics would be helpful for example, in the diagnosis and rehabilitation of speech-language pathologies, transmission of information etc.

TABLE 3

The relative frequency of different types of syllables in Kannada.

Types of Syllable	Frequency-Percentage
V	5.12
CV	72.78
VC	0.78
CVC	4.38
CCV	16.01
CCVC	0.75
Others	0.18

As in the case of relative frequencies of speech sound, a simple count of the-different syllable types was made. The results are given in Table 3. The syllables commonly occurring in practice may be classified into the following main categories : (i) Pure Vowels (V); (ii) Consonant-Vowels (CV) ; (iii) Vowel-Consonants (VC) ; (vi) Consonant-Vowel-Consonant (CVC) ; (v) Consonant-Consonant-Vowel-(CCV) ; (iv) Consonant-Consonant-Vewel-Consonant (CCVC) ; and (vii) others. In connected speech, a given linguistic utterance (whether in spontaneous speech or in written material) can be broken down into successive syllables in more than one way. In the computation cited here, we have followed the form of breakdown consistent with the transcription scheme employed in the Kannada language. As the orthography of Kannada is based on the principles that each syllable shall be represented by a single character, written material in Kannada lends itself to syllabic analysis in a straightforward manner. One can take a sample text, put a separate dot for each syllable in the proper category and count the percentages after a sufficient sample has been analysed. The percentages of the different types of syllables thus obtained from samples of about 3,65,000 characters in written kannada are given in Table 3.

The figures given in Table 3 are comparable to those of Ramakrishna et al (1962) which were obtained from sample of about 10,000 characters of written Kannada. The similarity of the results suggests that length of the sample analysed may not be that crucial as it has been made out to be. In Kannada, ' CV ' and CCV ' types of syllables have a much higher percentage of occurrence than ' VC ' and ' CVC' Types of syllables. However, the higher frequency of occurrence of CVC ' syllables may have been influenced by the particular form of breakdown of syllables that was followed here.

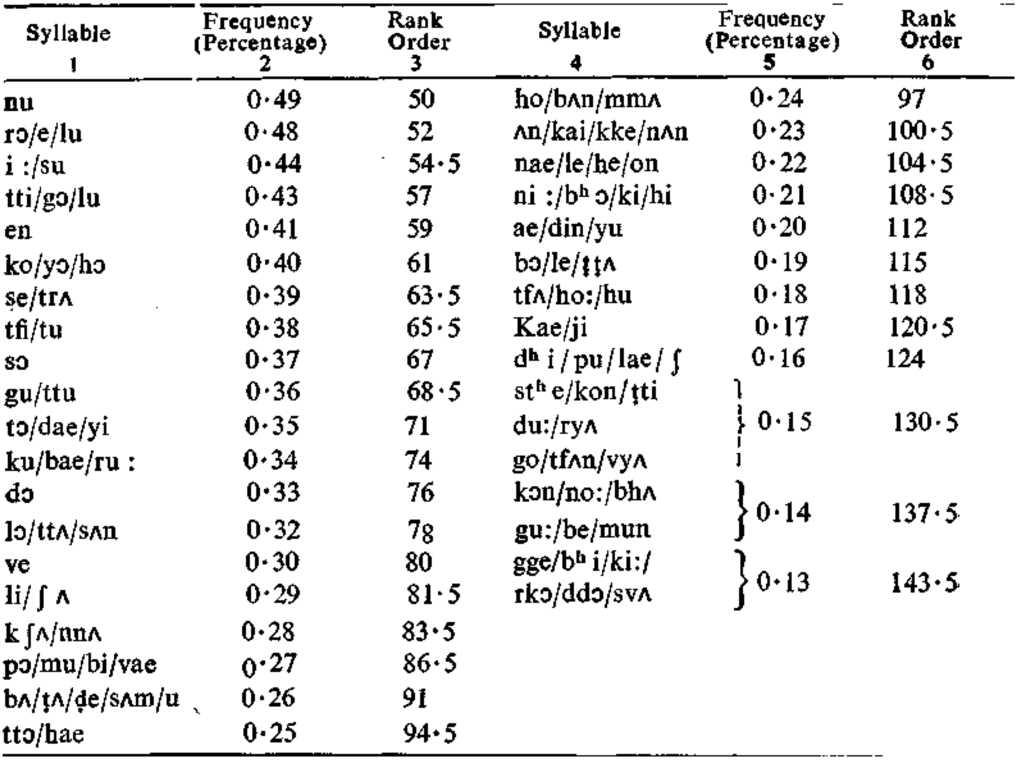
TABLE 4

50 most frequently occurring syllables in written Kannada. 3,62.282 syllables in 1,00,078 words have been analysed to arrive at these statistics. The syllables with rank order 1 to 20 had occurrence figures of 1,000 or more in every 1 lakh syllables analysed (shown in columns 1 to 3) while syllables with rank orders from 21-5 to 49 had occurernce figures of 500 to 1,000 in every 1 lakh syllables analysed (shown in colums 4 to 6).

Syllable 1	Frequency (Percentage) 2	Rank Order 3	Syllable 4	Frequeney (Percentage) 5	Rank Order 6
dʌ	4·85	1	ɔ/si	0·97	21·5
rʌ	2·82	2	ni/re	0·95	23·5
vʌ	2·65	3	i	0·89	25
ʌ	2·24	4	li	0·87	26
nʌ	2·13	5	pʌ/lv	0·81	27·5
gʌ	2·09	6	ḍi	0.80	29
yʌ	1·9	7	vɔ/vi	0·79	30·5
ru	1·64	8	nɔ	0·78	32
kʌ	1·58	9	bʌ/vu	0·77	33·5
du	1·55	10	mɔ	0·76	35
sʌ	1·52	11	de	0·75	36
ri	1·38	12	ti	0·73	37
tʌ	1·36	13	hʌ	0·71	38
lli	1·35	14	jʌ/llʌ	0·69	39·5
nnu	1·33	15	ḍu/ne	0·67	41·5
ge	1·29	16	di	0·66	43
gi	1·16	17	ḍʌ	0·65	44
lʌ	1·12	18	te	0·59	45
mʌ	1·07	19	ɚʌ	0·58	46
kɔ	1·02	20	prʌ	0·56	47
			ke	0·55	48
			ddʌ	0·54	49

TABLE 5

List of the most frequently occurring syllables in written Kannada. The syllables with rank orders from 50 to 94.5 had a frequency of occurrence of 250 to 500 in every 1 lakh syllable analysed (shown in columns 1 to 3) and syllables with rank: order from 97 to 143.5 had a frequency of occurrence of 125 to 250 in every 1 lakh: syllables analysed (shown in columos 4 to 6). The analysis of written Kannada for: the frequency of occurrence of different syllables yielded 896 different syllables with various percentages of occurrence, but all these syllables have not been shown in Tables 4 & 5 (Table 5 should be construed as a continuation of Table 4). The syllables which have not been shown here had occurrence figures of 10 to 123 per every 1 lakh syllables analysed.



Syllable 1	Frequency (Percentage) 2	Rank Order 3	Syllable 4	Frequency (Percentage) 5	Rank Order 6
nu	0·49	50	ho/bʌn/mmʌ	0·24	97
rɔ/e/lu	0·48	52	ʌn/kai/kke/nʌn	0·23	100·5
i :/su	0·44	54·5	nae/le/he/on	0·22	104·5
tti/gɔ/lu	0·43	57	ni :/b^h ɔ/ki/hi	0·21	108·5
en	0·41	59	ae/din/yu	0·20	112
ko/yɔ/hɔ	0·40	61	bɔ/le/ṭṭʌ	0·19	115
ṣe/trʌ	0·39	63·5	tfʌ/ho:/hu	0·18	118
tfi/tu	0·38	65·5	Kae/ji	0·17	120·5
sɔ	0·37	67	d^h i / pu / lae / ʃ	0·16	124
gu/ttu	0·36	68·5	st^h e/kon/ṭti	0·15	130·5
tɔ/dae/yi	0·35	71	du:/ryʌ		
ku/bae/ru :	0·34	74	go/tfʌn/vyʌ		
dɔ	0·33	76	kɔn/no:/bhʌ	0·14	137·5
lɔ/ttʌ/sʌn	0·32	78	gu:/be/mun		
ve	0·30	80	gge/b^h i/ki:/	0·13	143·5
li/ ʃ ʌ	0·29	81·5	rkɔ/ddɔ/svʌ		
k ʃʌ/nnʌ	0·28	83·5			
pɔ/mu/bi/vae	0·27	86·5			
bʌ/ṭʌ/ḍe/sʌm/u	0·26	91			
ttɔ/hae	0·25	94·5			

Relative frequencies of the different syllables were also calculated and the 150 most frequent syllable are given in Tables 4 and 5. The sample was the same used for computing the frequency counts of different types of syllables. The 3,65,000 syllables analysed yielded 896 different syllables. The 146 syllables reported in Tables 4 and 5 amount to 83% of the syllables analysed.

It can be seen from Tables 4 and 5 that the most frequently occurring syllables / dA / had occurrence rate of 4853 per every 1 lakh syllables and the least frequent syllable / SVA / had occurrence rate of 128 for every 1 lakh syllables. The remaining 746 syllables, which have not been shown in the Tables, had occurrence rate ranging from 5 to 123 per every 1 lakh syllables analysed.

The occurrence rate of different types of syllable among the 150 syllables shown in Tables 4 and 5 are as follows: V-2236 (most frequent) and 195 (least frequent) ; CV-4853 and 128 ; VC-416 and 216 ; CVC-319 and 139 ; and CCV-1352 and 128 (The occurrence rate is in every 1 lakh syllables analysed).

Though the relationship is not perfect, it appears from Table 4 that the syllables which occur most frequently are the CV combinations of the most frequently occurring consonants and vowels. These statistics serve to show the operation of the Zipf's priniciple of least effort in verbal behaviour also.

TABLE 6

Frequency of occurrence of different sounds in the sentence—initial position (columns 1 to 3) and word-initial position (columns 4 to 6), respectively. Sentence —initial position refers to the first sound of the first word of each sentence whereas the word-initial position refers to the first sound of each word.

Sound	Frequency (Percentage)	Rank Order	Sound	Frequency (Percentage)	Rank Ordei
1	2	3	4	5	6
ʌ	12.60	1	m	10.64	1
i	9.22	2	k	9.98	2
	8.57	3	b	8.06	3
n	7.47	4	s	7.98	4
k	6.39	5	h	7.48	5
m	5.83	6	ʌ	7.4	6
i:	5.43	7	n	6.9	7
s	5.34	8	P	6.73	8
h	4.55	9	v	4.41	9
p	3.96	10	t	4.32	10
b	3.10	11		4.07	11
r	3.09	12	e	2.99	12
t	2.85	13	i	2.83	13
v	2.45	14	d	2.66	14
o	2.31	15	b^h	2.49	15
j	2.12	16	r	1.91	16
J	2.13	17	j	1.83	17
e	2.19	18	u	1.75	18
d	1.70	19	o	1.49	19
tf	1.29	20	ae	1.08	20
y	1.23	21		0.91	21
g	1.13	22	t^f	0.83	22
b^h	1.05	23	g	0.75	23
ae	0.9	24	y	0.5	24
l	0.68	25			
u	0.67	26			

TABLE 7

Frequency of occurrence of the most frequently occurring syllables in the sentence - initial (colums 1 to 3) and word - initial positions (columns 4 to 6), respectively. Sentence - initial position refers to the first syllable of the first word of all sentences while word - initial position refers to the first syllable of all words.

Syllable 1	Frequency (Percentage) 2	Rank Order 3	Syllable 4	Frequency (Percentage) 5	Rank Order 6
Λ	10.98	1	Λ	6.98	1
	8.92	2		4.07	2
i	7.13	3	mΛ	3.74	3
i:	5.65	7	kΛ	3.33	4
nΛ	2.54	5	**SΛ**	2.99	5
mΛ	2.25	6	nΛ	2.91	6
kΛ	2.08	7	i	2.74	7
SΛ/no	1.97	8.5	**mΛ**	2.24	8
r	1.84	10	vi/bΛ	2.16	9 5
ni	1.68	11	pΛ/b^h	1.91	11.5
e	1.59	12	u/e	1.66	13.5
tΛ	1.38	13	i :	1.58	15
in	1.24	14	tΛ/pU	1.5	16.5
vi	1.21	15	ke/s	1.33	18.5
JΛ	1.17	16	ni/prΛ	1.25	20.5
prΛ	1.14	17	mu/hΛ/h	1.16	23
ke	1.11	18	ae/pr /mae	1.08	26
Λn	1.06	19	en/jΛ/n /bae/vΛ	0.98	30
bΛ	1.03	20	do/bi	0.91	33.5
h	1.02	21	r	0.83	35
hΛ	0.98	22	on/rΛ/hae/hin	0.75	37.5
pΛ/ae	0.92	23.5	o/ku/kai/t		
s	0.83	25	ni :/mi:/he/hou	}o.67	43.5
k	0.79	26	ko/tum/bΛn/vai	0.58	49.5
on	0.78	27	tfi/ji: /nae		
y	0.76	28	be/b^h Λ/sΛm	}0.50	55.5
ri:	0.71	29	hi/hu:		
o	0.70	30	Λn/kru/tΛn/d /dae/		
u/ni:	0.68	31.5	hu/ho	}0.42	63
tu	0.63	33	ou/em/di/me		

Syllable 1	Frequency (Percentage) 2	Rank Order 3	Syllable 4	Frequency (Percentage) 5	Rank Order 6
mu	0.6	34	mo/su : spΛ/sΛn/ hru	0.33	71
m	0.57	35			
he	0.56	36			
t /dae/pu	0.54	38			
b^h	0.52	40			
ko/p	0.51	41.5			
rΛ	0.49	43			
jΛ/tfi	0.48	44.5			
mo	0.43	46			
mi :/rou	0.41	47.5			
mun	0.4	49			
bi/hin/vai	0.38	51			
be	0.36	53			
di/hi:/ ho/bae/mu v / :	0.35	57			
SΛn	0.32	61			
SΛm/tfΛ i/si	0.3	63.5			

Frequency of occurrence of sounds and syllables in word-initial and sentence— initial position :

Simple freqency counts of all sounds and syllables in the sentence - initial and word-initial positions were made and the results are tabulated in Tables 6 and 7, respectively. It can be seen from Table 6 that it is more likely that sentences in written Kannada texts begin with a vowel, whereas individual words are more likely in Tables to begin with consonants. The same observation holds good for syllables too (Table 7). This observation might be of relevance in the study of stuttering where it is well known that most of the stuttering occurs on the initial sound or syllable of a word or initial word of a sentence and so on. Further, it also suggests that past attempts at relating stuttering on sounds and syllables to the phonetic complexity in the execution of such sounds may have to be reviewed. (Jayaram, 1983).

The present study was concerned with obtaining statistics on different symbols and their clusters in Kannada language. Once in possession of such statistics, we can construct a very convenient mathematical model of language called the Markoff process in statistics which can be utilised for artificially generating messages according to certain laws of probability. However, the primary significance of such data is for the communication theory in the sense that such statistical data may be used to difine the information content of a message. Information content of a message is a measurable quantity without reference to the meaning or its comprehension. It is possible to estimate the entropies of messages in a language on the basis of the frequencies of letters, speech sounds or their diagrams as suited to the particular situation.

Application on Speech and Hearing :

As has been mentioned several times above, though the primary significance of the statistics given here is for the communication theory, nevertheless, the statistics find application in the field of Speech and Hearing too. Armed with a list of the most frequently occurring sounds and syllables, it should be possible to plan and develop effective therapy material for children as well as adults with speech and language disorders. For example, in articulation disorders, a list of the most frequent sounds and syllables would guide the therapist to select the sound to be corrected first among all the sounds misarticulated. Obviously, correcting the sound most frequently occurring among the sounds misarticulated could improve the intelligibility to a greater extent than otherwise. Again, once the sound has been corrected, a list of the most frequently occurring syllables would guide the therapist in selecting the consonant - vowel combination for syllable drill.

Similarly, these lists would also guide in speech therapy for the deaf and hard of hearing. The deaf and hard of hearing children have limited language and it is also true that their ability to acquire language is limited. Therefore, it is better to teach them what occurs most frequently in the language, be it words or syllables or sounds or sentence patterns.

The statistics pertaining to sounds and syllables would also be helpful in developing stimulus material for determining speech discrimination scores. It can be ensured that the PB word lists developed would have the same distribution of sounds and syllables as in the natural language. By same logic, such statistics would be helpful in the construction of speech tests for central auditory disorders. Because material for speech discrimination scores are used in modified forms in tests for central auditory disorders. It is also believed that material here would be useful in developing test material for hearing aid trial purposes.

By the same token, it can be said that the statistics given here would be useful in research as well as diagnosis of speech and language disorders. One can certainly measure the severity of articulation disorders or put to test several hypotheses pertaining to phonetic and linguistic aspects of stuttering. The data enable to put into test a number of hypothesis on child's speech and language development too.

In babbling stage, during speech development it is the CV syllables which appear first. It would be intersting to study, utilising the statistics given in the preceding pages, whether the sounds which have a higher percentage of distribution are the ones to appear first in the speech-language development of a child ?; On similar lines, whether children develop cues to identify these most frequently occurring syllables first in their perceptual development ? ; On the dissolution of language, whether the syllables least uttered are the first ones to be lost when there is a damage to the speech production and perception system ? All possible applications of the above statistics are not given here for the simple reason that the full potentials of such statistics have not been understood. With little imagination a clinician or a researcher should be able to use these statistics more effectively.

REFERENCES

Jayaram, M., Phonetic inflluences on stuttering in monolingual and bilingual stutterers. 1983) Journal of Communication Disorders, 16,287-297.

Ramakrishna, B. S., Nair, K. K., Chiplunkur, V. N., Atal, B. S.

Ramachandran, V., and Subramanyam, R., (1962). Some aspects of the relative efficiencies of Indian Languages. Bangalore Indian Institute of Sceince.

Zipf, G. K., (1949) Human behaviour and Principle of Least Effort. Addison—Cambridge, Mass., Wesley Press Inc.,