J. Indian Inst. Sci., May–June 1988, 68, 199–205

Prediction and entropy of printed Kannada

M. Jayaram
Department of Speech Pathology, National Institute of Mental Health & Neurosciences (NIMHANS), Bangalore 560 029, India.

Received on July 4, 1986; Revised on September 18, 1987.

Abstract

Upper and lower entropy bounds have been calculated for Kannada text. The method employed depends on experimental results in prediction of the next letter when the preceding text is known. Results of experiments in prediction are given as also on redundancy. The last section of the paper deals with a comparison of the information entropies of Kannada and English texts.

Key words: Entropy, information, redundancy, digram, trigram, language, phonetics.

1. Introduction

Entropy is a statistical parameter which measures the average information produced per each letter of a text in the language[1]. On the other hand, redundancy measures the amount of constraint imposed on a text in the language due to its statistical structure[2]. The high frequency of occurrence of the vowel /Λ/ as well as the strong tendency of the vowel /Λ/ to follow all consonants in the Kannada language illustrate this point. In other words, entropy measures how uncertain we are on the average of the outcome of the event while redundancy is an average measure of our confidence in the outcome. Redundancy is a complementary measure of information. Both these measures are applicable to any natural language, it being considered a sequence of symbols. Language, however, does not merely exist as a sequence of symbols; it exists mainly by being internalized as a set of rules governing the use of language symbols among its users. Accordingly, Shannon[3] devised a method of estimating these quantities which is more sensitive in the sense that the method takes account of the long-range statistics, that is, the influences extending over phrases and sentences. The method is based on the predictability of English: how well the succeeding letter in a text can be predicted when the preceding N letters are known. The method consisted of asking a number of native speakers of English to guess each letter in a text and then calculating the amount of information per symbol on the basis of the number of correct guesses one could make, the number of times one made an error in guessing correctly each letter, etc. By combining the results of some experiments in prediction as well as the results of theoretical analysis of some of the properties of ideal prediction, Shannon[3] estimated the upper and lower bounds for the entropy and redundancy in the English language. From such an analysis it appears that in ordinary English, the long-range statistical effects (up to 100 letters) reduce the entropy to something of the order of 1 bit per letter, with a corresponding redundancy of roughly 75%.

The purpose of the present study was to estimate the upper and lower bounds of entropy and redundancy in the Kannada language. The necessity for the study was both theoretical and practical; on the theoretical side, the interest was to have the corresponding information pertaining to Kannada language. On the practical side, the results of such a study would have possible applications in telegraphic transmission and more importantly in the study of speech disorders. The predominant occurrence of stuttering on the initial sound or syllable of a word, the misarticulation in children and the clue offered by the initial sound or syllable of a word for the perception of the whole word in the case of deaf and hard-of-hearing speakers, are the subject matters in speech pathology. They could be more profitably investigated with the information available on redundancy in the language. Redundancy pattern of a language is an important aspect of it and characterizes the internal constraints operating in the language. With increasing automation in language processing devices, statistical information of language and its operating constraints will be needed in increasing measures.

2. Prediction of Kannada text

The method proposed by Shannon[3] was employed in the present study for estimating entropy. The method is based on the fact that any one speaking a language possesses, implicitly, an enormous knowledge of the statistics of the language in question. Familiarity with the words, syntax, meaning, etc, enable him to complete a word or phrase in conversation. The method of collection of data was as follows: one hundred samples of Kannada text were selected at random from books and magazines, each 100 to 125 letters in length. The matter of the texts was of general interest. The subjects were required to guess each sample phoneme by phoneme. If the subject was wrong in his guess, he was told so and asked to guess again. This was continued until he found the correct phoneme. A typical result of this experiment is shown below. The first line is the original text (phoneme by phoneme) and the numbers in the second line indicate the number of guesses the subjects made to arrive at the correct phoneme. The Kannada text in the first lines is written in international phonetic alphabet[4].

(1) K ʌ n n ʌ ḍ ʌ   b^{h} ɔ δ e y u   K a * n ɔ ṭ ʌ k ʌ

(2) 9 1 4 1 1 1 1 / 2 1 1 1 1 1 1 / 7 1 2 1 1 1 1 1 1 /

(1) r ɔ j y ʌ d ʌ   ɔ ḍ. ʌ l i t ʌ   b^{h} ɔ δ e y ɔ g i d e.

(2) 1 1 1 1 1 2 1 / 7 4 1 1 1 1 1 / 2 1 1 1 3 1 1 1 1 1 1 /

(1) S u m ⊃ r u   m u : r u : v ∧ r.e   K o : ṭ i   J ∧ n ∧ r u
(2) 6 2 3 1 1 1 / 1 0 7 4 1 2 1 1 1 1 / 1 4 1 1 1 / 1 1 1 1 1 1 /
(1) i :   b ^{h} ⊃ δ e y ∧ n n u   m ⊃ t ∧ n ⊃ ɖ u t t· ⊃ r e.
(2) 3 1 / 2 1 1 1 1 1 1 1 1 / 2 1 1 1 1 1 1 1 3 1 1 1 1

The subjects made more guesses to arrive at the correct phoneme, as can be expected, more frequently at the beginning of words and syllables or in the middle of words where the line of thought has more possibility of branching out. The latter possibility is more pronounced in the Kannada language than in English because the syntactic and the semantic markers (word inflections) the words can take is more flexible in Kannada than in English.

Out of 105 symbols above* the subject guessed right on the first guess 82 times, on the second guess nine times, on the third and fourth guesses four times each, and only six times required more than five guesses to get the correct phoneme. It may be said that results of this order are typical of prediction by a subject who is a native speaker of the language and possesses ordinary knowledge of the language. It is believed that texts in scientific work and poetry generally lead to somewhat poorer scores.

Five subjects whose primary language was Kannada predicted the phonemes in the text. In this study, spaces were considered as an additional letter in the text thus making a 51-letter alphabet in Kannada. One hundred samples were obtained in which the subjects had 0,1,2,3,... 14 preceding letters. As an aid to prediction, the subjects made use of various statistical tables[5]— letter tables, digram and trigram tables, a table of the frequencies of the initial syllable of words, and a list of 1000 most familiar words in Kannada[6]. The results of this test together with the results of a similar test in which 99 letters were known to the subjects are summarised in Table I.

The columns in Table I correspond to the number of preceding letters known to the subject plus one; the rows correspond to the number of guesses. The entry in column N at row S is the number of times the subject guessed the right letter at the Sth guess when (N–1) letters were known. Thus, entry 44 in column 5, row 3 means that out of the 1000 guesses, the subjects obtained the correct letter 44 times on the third guess when four preceding letters were known. The entries in the first two columns of Table I were not obtained from the experimental procedure outlined above but were calculated directly from the known letter and digram frequencies in Kannada[5]. Thus, with no known letters, the most probable letter in Kannada is the vowel /∧/ (probability 0.168); the next guess if the first guess is wrong should be a space (probability 0.12). These probabilities are the sequences with which the right guess would occur at the first, second trials with best prediction. Similarly, a simple calculation from digram tables gives the entries in column 2 when the subjects use the table to best advantage.

* Even the word space was considered here as an alphabetical symbol and hence, Kannada was considered to have a 51-letter alphabet in all our calculations.

Table I
Prediction scores of the subjects when (N−1) letters are known (N = 1,2,3...15...100)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	100
1	168	193	310	584	780	762	762	766	762	800	772	848	768	806	808	962
2	120	127	114	146	96	94	86	86	90	64	88	54	108	62	62	20
3	66	87	62	62	44	40	40	38	32	22	30	24	28	38	14	—
4	62	70	70	76	30	22	28	30	28	26	24	24	16	20	18	8
5	53	62	38	14	6	16	14	10	16	8	6	8	12	6	6	—
6	49	43	62	18	10	6	12	8	8	12	18	4	10	10	8	2
7	49	37	64	22	8	8	26	8	20	12	8	6	14	16	8	4
8	43	37	26	8	4	4	2	6	2	10	4	4	6	8	4	—
9	38	35	66	10	2	10	6	14	8	6	10	2	10	6	6	4
10	33	30	46	16	4	8	4	4	4	4	4	2	—	—	2	—
11	29	27	28	10	6	4	8	8	2	6	4	4	10	4	4	—
12	28	27	24	6	2	—	4	2	2	6	2	2	—	4	6	—
13	27	23	16	2	—	4	2	8	6	4	6	2	—	6	4	—
14	23	23	14	6	—	4	2	2	2	4	4	2	4	4	4	—
15	22	22	14	2	2	4	2	—	2	6	—	6	2	2	—	—
16	19	18	12	4	—	2	—	—	4	2	6	—	4	4	2	—
17	17	17	4	.2	—	4	—	4	2.	2	2	—	2	—	2	—
18	14	16	10	4	2	2	—	2	—	—	—	4	2	—	—	—
19	13	15	8	6	2	2	—	—	4	2	2	—	2	—	—	—
20	11	11	4	—	2	—	—	2	—	2	—	—	—	2	—	—
21	11	10	2	2	—	—	—	—	2	2	—	—	—	—	—	—
22	10	10	2	—	—	2	—	—	2	—	—	—	—	2	—	—
23	9	8	—	—	—	—	—	—	—	—	—	—	—	—	—	2
24	9	7	2	—	—	—	—	2	—	—	2	—	—	—	—	—
25	8	7	—	—	—	—	2	—	—	—	—	—	2	—	—	—
26	7	5	2						2							
27	7											2				
28	6	3														
29	6	1				2										
30	5	1														
31	5	1									2					
32	4	1									2	2				
33	4	1									2					
34	4	1														
35	3	0.5														
36	3															
37	2															
38	1															

It must be said here that as the frequency tables in columns 1 and 2 were determined from long samples of Kannada text, they are subject to less sampling errors than other columns. Also, the subjects were instructed to guess the most probable next letter, the second most probable letter, etc., for each possible N-gram of text.

It can be seen from Table I that the prediction gradually improves, apart from some sampling fluctuations, with increasing knowledge of the past as indicated by the large number of correct first guesses and the small number of high rank guesses.

Table II
Upper- and lower-bound entropy for N-gram text in Kannada

Letters known	Upper bound	Lower bound
1	4.379	3.425
2	3.947	3.126
3	3.557	2.586
4	2.230	1.298
5	1.325	0.546
6	1.513	0.736
7	1.478	0.723
8	1.498	0.727
9	1.524	0.744
10	1.394	0.668
11	1.495	0.722
12	1.082	0.476
13	1.436	0.738
14	1.318	0.618
15	1.084	0.477
100	0.263	0.097

3. Entropy bounds from prediction scores

The upper and lower experimental bounds for the entropy of the 51-letter alphabet of Kannada was calculated by the method given by Shannon (cf. eqn. 17[3]). The prediction frequencies are arranged in decreasing order of magnitude in Table I. The data have not been smoothed and hence some of the sampling fluctuations have remained as such. The upper and lower bounds, thus calculated, are given in Table II. The values in Table II are plotted against N in fig. 1.

It is evident that there are considerable sampling errors in these figures, particularly in the lower-bound entropy figures (Table II). However, it must be mentioned that the lower bound has been proved with only the 'ideal' predictor and that the frequencies used here are from human prediction.

If there are N-possible letters in the alphabet and they occur with equal probabilities, then with each letter we get $H = \log_2 N$ bits of information. When $N = 26$, as in the case of English alphabet, we get 4.71 bits* of information with each letter. Similarly, when $N = 50$, as in the case of Kannada alphabet, we get 5.64 bits of information with each letter (or 5.67 bits when $N = 51$ including space).

However, the different letters of the alphabet do not occur with equal probabilities or even independently of the particular sequence of letters preceding them. For example, as

* The figure for the entropy in the case of English refers to the entropy per letter and not to that of speech sound as in the case of Kannada language. The word space has been counted equal to a letter in English and a speech sound in the Kannada language.

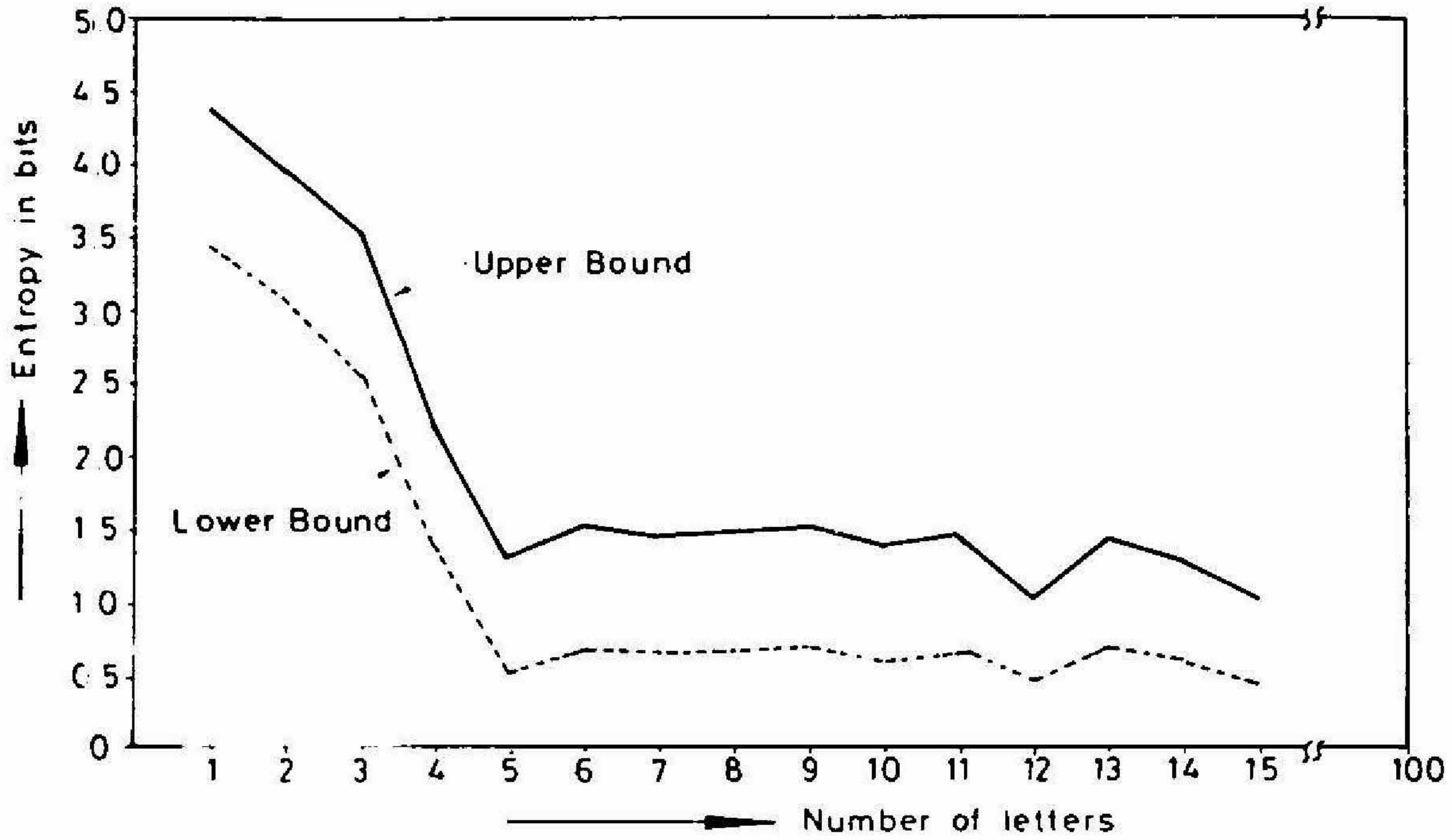

FIG. 1. Upper- and lower-experimental bounds for entropy of 51-letter Kannada.

said in an earlier section, the vowel /Λ/ which occurs most frequently in the Kannada language accounts for roughly 19% of the total number of letters in a long text followed by letter /i/ with 7.50% and /n/ with 6.99%[5]. If we use the actual frequencies with which letters occur in the English language, we find that the information per letter is 4.03, and not 4.71. bits which would have been the case if all letters were equally probable. Actually, the letters in any language do not occur independent of each other. A word is a cohesive group of letters with strong internal statistical influences and consequently, also, the N-grams within the words. Shannon[3] has shown that the information per letter calculated on the basis of digrams and trigrams in English language turns out to be progressively small, being 3.31 and 3.10 bits per letter. Further, considering the probabilities of longer sequences, the amount of information per symbol comes down to 1 bit per letter. In other words, it can be shown that the entropy per letter continues to decrease as 'N' becomes larger and larger.

The maximum entropy possible for a 26-letter alphabet is 4.71 bits per letter when all letters are equally probable and independent of each other. The ratio of the actual entropy obtained in a given set of messages to its maximum possible entropy is called the relative entropy and one minus the relative entropy is called the redundancy. In English, one finds a redundancy of 15% on the basis of letter frequencies alone, about 30% on the basis of digram frequencies and as much as 72% or more for longer sequences[7]. Similarly, in Kannada, one finds a redundancy of 22% on the basis of letter frequencies, 30% on the basis of digram frequencies and as much as 95% or more for longer sequences. The more severe the constraint imposed by grammar, syntax, etc., in a language, the more the redundancy.

Acknowledgement

I am thankful to Dr K. Richard, formerly of IISc, Bangalore who was of immense help to me in the mathematical calculations that appear in this paper.

References

1. Shannon, C. E. A mathematical theory of communication, *Bell System Tech. J.*, 1948, **27**, 379–423.

2. Shannon, C. E. A mathematical theory of communication, *Bell System Tech. J.*, 1948, **27**, 623–656.

3. Shannon, C. E. Prediction and entropy of printed English, *Bell System Tech. J.*, 1951, **30**, 50–64.

4. Gleason, H. A. *An introduction to descriptive linguistics*, 1966, Holt, Rinehart and Winston, U.S.A.

5. Jayaram, M. Sound and syllable distribution in Kannada and their application to speech and hearing, *J. All India Inst. Speech Hearing*, 1986, **17**, 79–88.

6. Rathna, N. *1000 most frequent words in Kannada*, 1979, personal communication

7. Ramakrishna, B. S., Nair, K. K., Chiplunkar, V. N., Atal, B. S., Ramachandran, V. and Subramanian, R. *Some aspects of relative efficiencies of Indian languages*, 1962, Indian Institute of Science, Bangalore.