Statistical Analyses of Telugu Text Corpora

G.BharadwajaKumar[1], *KaviNarayanaMurthy*[1], *BBChaudhuri*[2]

1: Department of Computer and Information Sciences
University of Hyderabad, India
2: CVPR Unit, Indian Statistical Institute (Kolkata)
email: knmuh@yahoo.com,g_vijayabharadwaj@yahoo.com,bbc@isical.ac.in

Abstract

Corpora and corpus based studies are relatively recent and limited in Indian languages as compared to other major languages of the world. This paper is about statistical analyses of Telugu text corpora. Telugu is one of the major languages of India, ranking third position in terms of number of speakers in the country, spoken mainly in the southern state of Andhra Pradesh. Telugu belongs to the Dravidian family of languages, characterized by a rich system of morphology resulting in long and complex word forms. In this paper, we analyze a nearly 39 Million word text corpus of Telugu developed by us. This is perhaps for the first time that such a detailed statistical analysis of a large corpus is being carried out for any Dravidian language. We highlight the complexity of word forms and the effect of this on the statistical characterization of the Telugu language.

Key Words:- Statistical Analysis of Corpora, Corpus Linguistics, Telugu corpus, Zipf's law, Mandelbrot's law, Script Grammar

1 Introduction

Corpus based approaches to language have made significant contributions to linguistic research as also in education and language technology. A corpus is a large and representative collection of language material stored in a computer processable form[40]. Corpora provide realistic, interesting and insightful examples of language use for theory building and for verifying hypotheses[2, 25, 42, 37]. Insights obtained from analysis of corpora have led to fresh and better understanding of how language actually works[5, 6, 27, 41]. Corpora provide the basic language data from which lexical resources such as dictionaries, thesauri, word-nets, etc. can be generated[12]. Language technologies and applications such as Morphological Analyzers, Stemmers, Syntactic Parsers[24], Spell Checkers, Information Retrieval systems, Information Extraction systems, Automatic Text Summarization systems, Automatic Text Categorization systems, Machine Translation systems[17] etc. greatly benefit from language corpora. Development of large and representative corpora and annotating them with morphological, syntactic and semantic infor-

mation is therefore considered to be a priority area. Corpus based statistical approaches have emerged as promising alternatives to traditional linguistic approaches. Hybrid approaches that combine traditional linguistic approaches with corpus based statistical approaches have also become attractive.

While corpus based and statistical approaches to language have been well established elsewhere in the world, India is still lagging far behind[32, 35, 33, 9]. Corpus based studies in English date back to 1960s [23, 1, 19, 22, 26, 30, 14]. Corpus linguistics [29, 36, 20] has not yet become a major aspect of education and research in linguistics in India. Even plain text corpora available are inadequate and annotated corpora are hardly available in many languages. Shastri [39] and colleagues developed the *Kolhapur Corpus of Indian English (KCIE)* nearly three decades ago following the design of the Brown and LOB Corpus. It was only in the late eighties that the need for developing corpora in native Indian languages was felt. Small plain text corpora (about 3 Million words) were generated over the next ten years in a handful of major Indian languages through the initiatives of the TDIL (Technology Development in Indian Languages) group of the Department of Information Technology (then known as DoE, the Department of Electronics), Government of India. Even after development, it took many years for these corpora to be released for research. These corpora have not been carefully checked or proof read - there are errors. Prakash Rao et al [38] carried out basic statistical analysis of these corpora as also a few other small corpora of Hindi in the year 2002. Chaudhuri et al have developed and analyzed Bangla text corpora [8].

We see that there have been scattered attempts to develop and analyze small text corpora in Indian languages. Some work has been done in Bangla [10, 3, 28] and Hindi[43, 21] but many languages are yet to make a mark. As we shall see here, there is a need to develop large scale corpora, especially for Dravidian languages. In this paper we describe our efforts in developing a nearly 39 Million word corpus for Telugu and statistical analysis of this corpus and its subcorpora along various dimensions. We shall highlight the complexity of word forms and the effect of this on the statistical characterization of the Telugu language.

2 Preliminary Analysis of DoE-CIIL Corpora

Here we present a preliminary analysis of the DoE-CIIL corpus developed in part and distributed by CIIL (Central Institute of Indian Languages, Mysore). This includes corpora of 13 major Indian languages, each approximately 3 Million words in size. The languages included are Assamese, Bangla, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sindhi, Tamil, Telugu and Urdu. Of these, Kashmiri, Sindhi and Urdu are in Perso-Arabic scripts. The rest of the 10 languages are in various scripts, all of which are derived from the ancient braahmi script. All the braahmi based scripts are supported by ISCII, the Indian Script Code for Information Interchange, a 1991 BIS (Bureau of Indian Standards) standard[7]. All of our work is based on ISCII (UNICODE being nearly one-to-one equivalent to ISCII) and here we limit our analysis to the 10 languages whose corpora are available in ISCII encoding.

Note that these 10 languages cover only the Indo-Aryan and Dravidian families. Also see Prakash Rao et al [38] for a preliminary analysis of the DoE-CIIL corpora.

The figure 1 shows the results of a type-token growth rate analysis. Each distinct word form is a type and each occurrence of a type counts as a token. If we analyze the entire corpus in one go, we will get the total number of types, total number of tokens and the global type-token ratio. Instead, if we perform type-token analysis incrementally, by starting with a small randomly selected part of the corpus and iteratively adding more texts randomly, we get a type-token growth rate curve that shows how many new types will be found as the corpus size increases.

Tokenization is performed by a straight forward split across spaces. All punctuation marks and non-ISCII characters are treated as white space so as to get pure native language words. Note that by types we mean fully inflected word forms, not root forms or citation forms found in dictionaries. Also, saMdhi and compounding will have their effect and the tokens we get do not necessarily correspond to the linguistic definition of a word understood in semantic terms. There is no automatic way to extract words based on meaning. Wide coverage, high performance, robust morphological analyzers are not yet available in most languages under study and here we restrict our analyses to full words.

From the figure 1, it can be seen that Indo-Aryan languages show clear signs of saturation while Dravidian languages do not. Initially, almost every word seen will be a new word and hence the growth rate curve starts off as a 45 degree line (if the x and y axis are drawn to same scale). As we progress, many of the frequent words would have already occurred and we will start seeing more and more of repetitions of the same old words and less and less occurrences of new, hitherto unseen words. If the slope of growth rate curve reduces and approaches the horizontal, it indicates that most of the types in the language have already occurred and even if we build much larger corpora we are going to still see the same words used repeatedly and very few new words are likely to appear. It can be seen that Indo-Aryan languages have around 150000 to 200000 words. This is quite comparable to English. These are fairly small numbers and it is conceivable that the entire set of all types is simply stored in a list, thereby making it unnecessary to perform detailed morphological analysis or stemming for many of the simple applications such as spell checking. In the case of Dravidian languages, it is seen from figure 1 that there are no signs of saturation and we must expect to see many new, unseen word forms as we increase the size of the corpus further. Many possible words have not occurred even once in the corpus. The corpus is obviously inadequate for even simple applications.

DoE-CIIL corpora are not entirely clean and our preliminary experiments with original and partly cleaned versions of some of these corpora have shown that while the general trends and the broad picture should be dependable, fine reading into the differences among language pairs may be a bit premature at this stage. For example, we cannot provide a statistically significant characterization of similarities and differences between Telugu and other Dravidian

languages until we build larger corpora in these languages. All that we can say for sure for now is that Dravidian languages have a significantly higher type-token ratio than Indo-Aryan languages. Also, larger corpora are needed for Dravidian languages to get a better picture. The rest of the paper will focus only on Telugu.

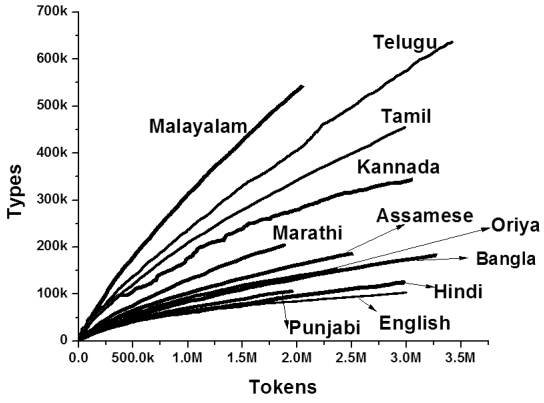

Figure 1: Type-Token Growth Rate Analysis of DoE-CIIL Corpora

The table 1 summarizes the results of type-token analysis. For the sake of comparison, a 3 Million word English Corpus derived from the British National Corpus (BNC) by random selection has been included.

Table 1: Type-Token Ratio

Language	Fam -ily	Tokens	Types	Type Token Ratio (%)
English	-	3000000	102690	3.42
Hindi	IA	2980968	124932	4.19
Punjabi	IA	1953762	105602	5.40
Bangla	IA	3270332	183131	5.60
Oriya	IA	2361191	152766	6.47
Assamese	IA	2502687	185616	7.42
Marathi	IA	1886143	204302	10.83
Kannada	Dr	3047568	343971	11.29
Tamil	Dr	2986273	454070	15.20
Telugu	Dr	3669322	636022	17.33
Malayalam	Dr	2046207	542239	26.50

3 The Nature of Telugu Language

A significant part of grammar that is handled by syntax in English (and other similar languages) is handled within morphology in Telugu (and other Dravidian languages). Phrases including several words (that is, tokens) in English would be mapped on to a single word in Telugu. Thus 'vaccaaDu' ((he) came), 'vastaaDaa' (will (he) come?), vaste (if (he/she/it/they/I/we/you) come), 'ragalagutaaDu' ((he) will be able to come), 'raaleekapooyaaDu' ((he) was unable to come), 'vaccinavaaDu' (the person (3P,sl) who came), 'raaDanukonnaavaa' (do you think he will not come?) are all single words (that is, tokens) in Telugu, written and spoken as atomic units. Verbs may include aspectual auxiliaries apart from tense and agreement. There are several types of non-finite forms too. A single verbal root can lead to formation of a few hundred thousand word forms. Nouns are also inflected for number and case. Derivation being very productive, even more forms become possible when we consider full word forms. Thus 'vaccinavaaDiki' (to the person (3P, sl) who came) is a noun in singular, dative case derived from the verb root 'vacc' (to come). External saMdhi (that is, conflation between two or more complete word forms) and compounding add

to the numbers. Naturally we will see very large number of types and the type-token ratio should be expected to be very high too. These are not simple concatenations or juxtapositions of complete words written without intervening spaces as is the convention in some languages of the world. These words are made up of several morphemes conjoined through complex morpho-phonemic processes. Telugu in particular, and Dravidian languages in general, are among the most complex languages in the world at the level of morphology, perhaps comparable only to Finnish and Turkish.

Modern Indian languages all have close ties with Sanskrit which is characterized by a rich system of inflectional morphology and a productive system of derivation, saMdhi and compounding. Yet, Dravidian morphology is significantly different and more complex than the morphology of Indo-Aryan languages. The focus of this paper is statistical analysis and detailed morphological characterization of Indian languages is beyond the scope of this paper.

Of course the root words will be much smaller in number compared to fully inflected word forms but morphological analysis, lemmatization, or even stemming is a challenging task and implemented systems available today are far from adequate to get a clear picture in terms of roots. We therefore limit ourselves to the exploration at full-word level in this paper.

We have seen that a 3 Million word corpus is hardly sufficient for a Dravidian language like Telugu. Given this scenario, it is interesting to see if saturation can be seen in the type-token growth rate with a larger corpus. We therefore developed a corpus of over 8 Million words for Telugu. Even in this corpus, (along with the DoE-CIIL corpus of Telugu adding up to about 12 Million words,) no signs of saturation are seen. See figure 2. Nearly 2100000 types have been obtained and even if we keep all these 2100000 words in a dictionary, we should still expect to see many new unknown words in any new Telugu text. We therefore chose to continue to develop larger corpora. More details of these corpora are given in the next section.

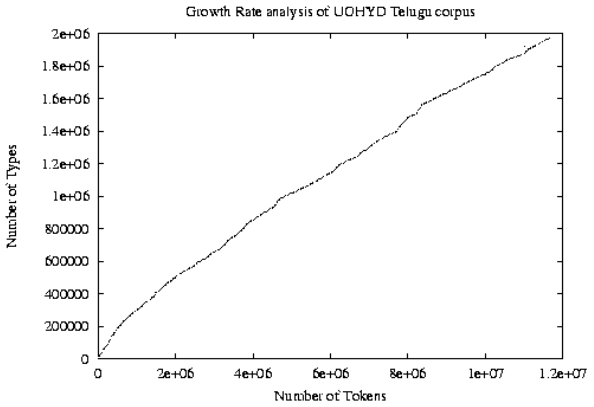

Figure 2: Type-Token Growth Rate Analysis of a 12 Million word Telugu Corpus

4 The LERC-UoH Telugu Corpus

The Telugu corpus developed at the Language Engineering Research Centre (LERC), Department of Computer and Information Sciences, University of Hyderabad, India, hereafter referred to as LERC-UoH corpus, adds up to nearly 39 Million words, perhaps one of the largest corpora for any Indian language today. This corpus includes 3 major sub-corpora

known as the CLC, NP1 and NP2 corpora (apart from the DoE-CIIL corpus). The CLC corpus includes 221 full books carefully selected by a panel of experts to include a wide variety of Telugu writings including a variety of genres, types and styles - modern and ancient, prose and poetry. This corpus has been checked and validated by a two-stage proof-reading process. The CLC corpus includes over 8 Million words. The NP1 corpus has been developed from the iinaaDu newspaper, one of the widely read newspapers in this region. iinaaDu runs a school of journalism of its own and its editors and sub-editors are well trained. The NP1 corpus is about 26 Million words in size, spread across nearly 9,400 files. This corpus was created by downloading selected articles from the on-line version of the newspaper and converting to standard ISCII [7] notation using tools developed here. The NP2 corpus was created similarly from the aaMdhra-prabha newspaper, another popular newspaper of the region. The NP2 corpus is smaller - about 1.3 Million words in size. All these corpora are ISCII encoded and are seen to be reasonably clean, although the NP1 and NP2 corpora have not been fully manually checked. UNICODE versions can be easily obtained since UNICODE for Indian languages has been designed with ISCII as a basis and ISCII and UNICODE have nearly one-to-one correspondence. Together with the DoE-CIIL corpus, we thus have a nearly 39 Million word corpus for Telugu.

A corpus should be constructed in keeping with the principles of corpus linguistics [29, 20, 11, 4]. It must be 'large' and 'representative'. A balanced corpus, however, does not mean nearly equal amounts of material from various genres, types and styles. In fact that would reflect a highly lop-sided view of reality. Telugu, like most other Indian languages, has mainly remained a literally language for ages. Application of language in areas other than literature is a relatively recent phenomenon. Even to this day, a vast majority of higher education, research, commerce, business, law, etc. are done in English and Telugu and other Indian languages play only a secondary role in India. Naturally we find very little of scientific and technical writings compared to literature. Instead of forcing an artificial balance in a naturally unbalanced world, we have chosen to include more material from literature. Further, newspapers cover a wider variety of topics and styles including sports, science and technology, politics, economics and business, cinema etc. No corpus should be put to use for a given application without a careful analysis of its nature and contents. While there can be no guarantee that our corpus is good enough for any given use or application, we feel that the corpus is good enough for some kinds of applications we have in mind and sub-sets can also be carefully selected for specific research goals or applications. Our aim shall be to strive to build larger, more balanced and more representative corpora.

In the following sections we give details of various statistical analyses we have carried out on this LERC-UoH Telugu corpus (including the DoE-CIIL corpus of Telugu).

5 Type-Token Analysis

Type-Token growth rate analysis has been carried out separately for each sub-corpus as also for the entire corpus. The figure 3 shows

the results. It can be seen that over the entire corpus of about 39 Million tokens, 3318717 types have been obtained and still the curve shows no signs of bending down. We should therefore expect many more types in the language. The overall type-token ratio is 8.51 %, much larger than for Indo-Aryan languages (which averages to 5.812 %) and English (3.42 %). See also [16, 15]. Analysis of the sub-corpora is also quite revealing. Table 2 shows the results. The CLC corpus, including as it does a wide variety of large literary works, shows the steepest growth rate. The newspaper corpora are much better behaved - newspapers tend to use a somewhat restricted vocabulary and the writing style is more consistent and constrained. We rarely find poetry in a newspaper. The DoE-CIIL corpus lies somewhere between these two but by itself it is hardly sufficient in terms of size.

It can be seen that, by the very nature of on-line news resources, some news articles which have a longer validity period tend to remain in some form, condensed though they may be, for several days. This results in a content-wise repetition and (although there are no duplicate files in the corpus) this repetition tends to bring down the type-token ratio. This is why the type-token growth rate curve for the NP1 corpus is so much lower than for other sub-corpora. While the type-token ratio (in percentage) for other sub-corpora have ranged between 12 and 19, the NP1 corpus shows a ratio of just over 6, bringing down the average ratio too. NP2 corpus, which is also from a Telugu daily, has a type-token ratio of 12.2 %. More detailed analysis is required to ascertain the exact reasons for the very low value for the NP1 corpus. If we look at the sub-corpora excluding NP1, we get an average type-token ratio of 16.06 % for Telugu.

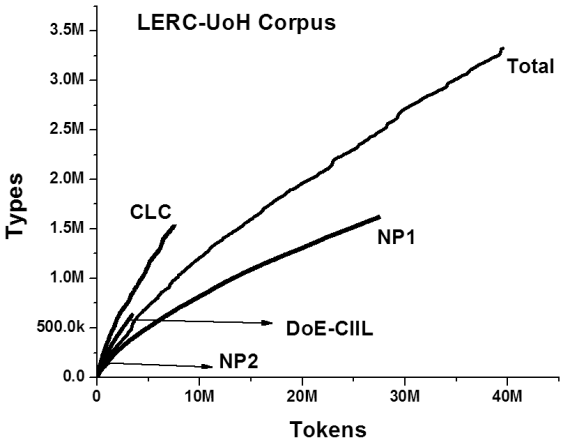

Figure 3: Type-Token Growth Rate Analysis of LERC-UoH Telugu Corpora

It is interesting to see how the type-token ratio varies as corpus size increases. See Figure 4. The ratio starts off high and quickly climbs down to about 0.2 by the time we have seen about a Million words. That is, at this stage, we can expect to see a new word once in every five words. This ratio further reduces gradually to about 0.1, meaning that every tenth word will be a new, hitherto unseen word. Interestingly, this ratio remains more or less at this level even when the corpus has grown to nearly 39 Million words, showing only a very small down trend. We should thus expect a new word roughly every ten words. This correlates with the fact that we have obtained 3318717 types from a corpus of 38960974 words. In other families of languages which show clear saturation, the type-token ratio would asymptotically approach

zero, meaning most of the types in the language have already been seen and new words continue to appear rarely. Telugu is different.

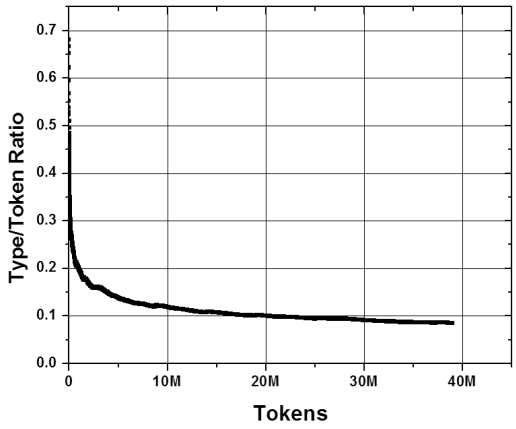

Figure 4: Type/Token Ratio for LERC-UoH Corpus

Table 2: Type-Token Ratio

Name	Tokens	Types	Type-Token Ratio (%)
CLC	8206738	1530584	18.65
NP1	26860087	1615991	6.02
NP2	1300583	158700	12.20
CIIL	3669322	636022	17.33
TOTAL	38960974	3318717	8.51

5.1 Repetition Analysis

It is interesting to study at what rate new types appear. The very first token we get is obviously a new type. The second token may be either a new type or a repetition of the type already seen and in all probability it will be a new type. We may find that the first few tokens are all new types but soon some of the most frequent types will start re-appearing. Towards the end of the corpus, we will find that most of the tokens are repetitions and only once in a while a new type appears. Thus a study of repetition intervals can be quite revealing, both in the beginning part and towards the end. Figures 5 and 6 show the repetition intervals - intervals after which a repetition occurs, as we see more and more of the corpus. The first few words are all new words and in one specific run (taking corpus files in random order), the first few repetitions occur after 5 to 8 words. This interval soon reduces to somewhere between 2 and 3. That is, after 2 to 3 new words, an already seen word re-appears. The repetition interval gradually reduces to about 0.1 and stabilizes there, indicating that roughly 9 out of 10 words are already seen words and new words appear once in about 10 words. If this interval were to reduce to nearly zero, that would have indicated saturation, where almost all the words we will be seeing are already seen words and very few new words are expected.

6 Coverage Analysis

Coverage analysis deals with the examination of how much of a corpus can be covered by a given set of types. We perform a type-token analysis and prepare a list of types sorted in decreasing order of frequency of occurrence. By

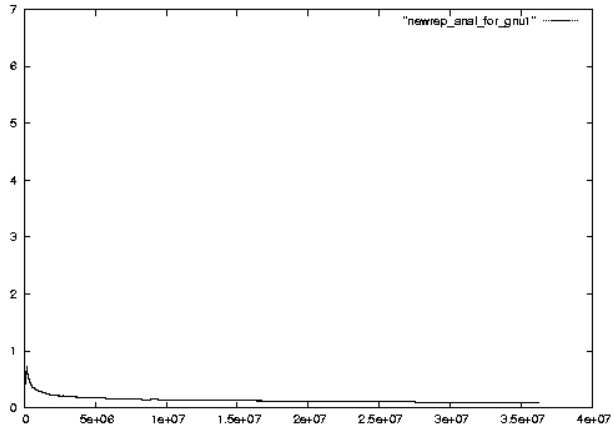

Figure 5: Analysis of Repetition of words in Corpus

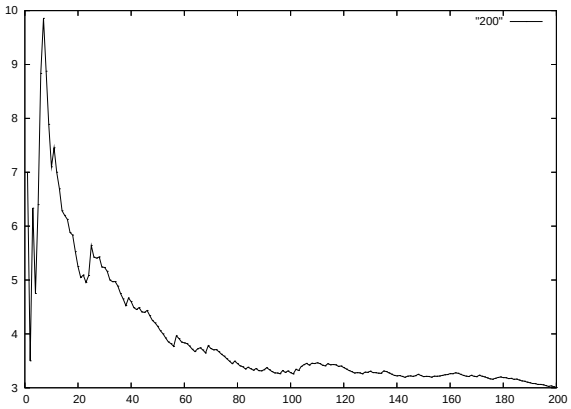

Figure 6: Repetition Analysis: Initial Region

thresholding on this list, we can select the most frequent n words in the language, for any given value of n. We then explore what percentage of words in a corpus are found in the list so selected. Here we perform self-coverage analysis - coverage analysis on the same corpus from which the words are extracted. (It would be instructive to perform coverage analysis on other corpora as and when they become available.)

From the figures 7 and 8 and also from tables 3 and 4, we can see that about 3700 most frequent words are sufficient to give about 50% coverage of the corpus. 60% coverage can be obtained by just the first 9000 words or so. 95% coverage requires 1.37 Million types, far higher than for English (For example, the most frequent 20,000 words give a coverage of about 95% on the British National Corpus which totals to nearly 100 Million words). This being a self-coverage analysis 100% coverage can be obtained by using all the words in the word list.

A dictionary based spell checking system, which proposes to maintain a list of full words, would require about 1.4 Million words to be stored even to get a reasonable coverage of about 95% (which can be lower for new, unseen texts). Compare this with Indo-Aryan languages where we can obtain higher coverage with about 100000 to 150000 words. Dravidian languages in general (and Telugu in particular) are an order of magnitude more complex than Indo-Aryan languages.

7 Zipf's Law

In his book entitled *Human Behaviour and the Principle of Least Effort*[45], Zipf argued for a unifying principle called *the Principle of*

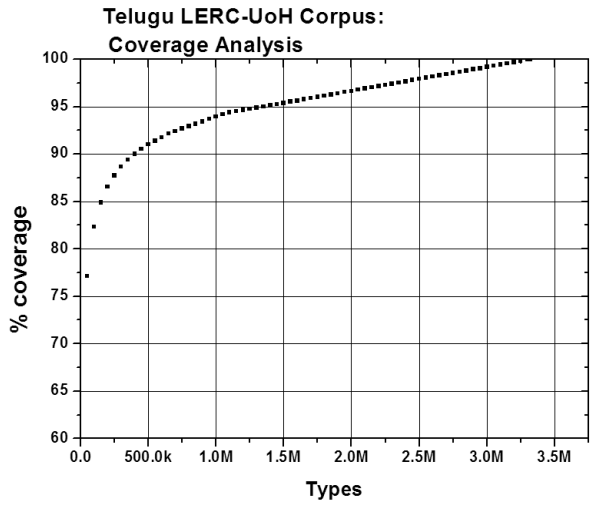

Figure 7: Self Coverage Analysis of LERC-UoH Telugu Corpus

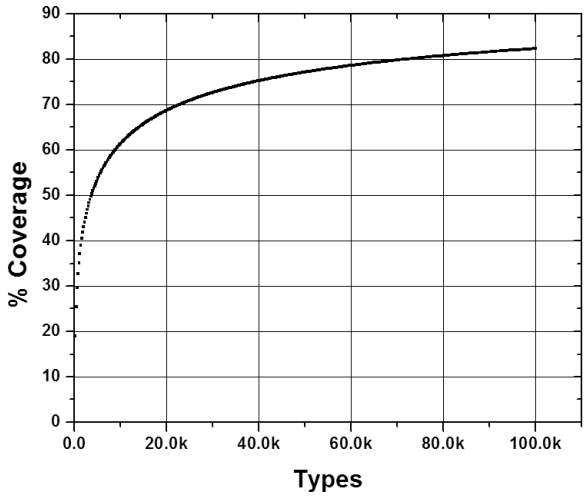

Figure 8: Self Coverage Analysis of most frequent 100000 types of LERC-UoH Telugu Corpus

Least Effort that he claimed underlies all of human cognition. Both speakers and listeners tend to minimize the rate of work they need to do immediately as also in the final analysis. One of the well known fall-outs of this theory is the well-known Zipf's law. Zipf's law states that the frequency of occurrence of words is inversely proportional to their rank (when ordered in terms of their frequency of occurrences):

$$f \propto \frac{1}{r}$$

That is, there is a constant k such that $f * r = k$.

There are a few very common words, a middling number of medium frequency words and a large number of low frequency words. The speaker's effort is minimized by having a small vocabulary of common words and the listener's effort is reduced by having a large vocabulary of individually rarer words leading to reduced ambiguities. The maximally economical compromise between these competing needs is argued to be the kind of reciprocal relationship between frequency and rank. Zipf's law predicts that a plot of frequency versus rank on a log-log scale should be a straight line with a slope of -1. In practice, while this general trend is seen in the middle portion, the most frequent words forming the top part of the graph and the least frequent words forming the tail show poor agreement with the law. Yet, Zipf's predictions are significant in the sense that they highlight the importance of hyperbolic distributions.

Table 3: Self Coverage Analysis of LERC-UoH Corpus

Number of Types	%Coverage
10	3.67
100	13.79
1000	35.04
10000	61.26
25000	70.78
50000	77.02
100000	82.24

Table 4: Self Coverage Analysis of LERC-UoH Corpus

%Coverage	Approx. No. of Types
50	3710
60	8930
70	23070
80	73000
85	155000
90	405000
95	1375000
96	1760000
97	2150000
98	2540000
99	2930000
100	3318717

Although many phenomena are appropriately modelled by the well-known Gaussian or Normal distribution, there are important phenomena in the world that are hyperbolic in nature. Zipf's law is a good example.

Figures 9 and 10 show plots of the frequency of words against their rank on a log-log scale for the LERC-UoH Telugu corpus. These figures show that the Zipf's law plots for Telugu are very similar to well known plots obtained for English and other languages. The least square regression fit line for the entire data shows a slope of -0.96 (figure 9) while if we take only the middle portion that is more or less linear, we get a slope of -1.12 (figure 10). In order to check the validity of the general trend with respect to corpus size, we have also plotted the curves for parts of the LERC-UoH Telugu corpus. It can be seen that the slope is low for smaller corpora and as the corpus size increases, the plots stabilize and conform better to Zipf's predictions. See figure 11.

Mandelbrot has studied these laws extensively. He has also suggested the following empirical distribution to obtain a closer fit:

$$f = P(r + \rho)^{-B}$$

Or

$$logf = logP - Blog(r + \rho)$$

where P, B and ρ are parameters to be experimentally determined. After extensive experimentation, we have obtained a good fit as shown in figure 12 for P = 15, B = 1.0, and $\rho = 20$.

8 Script Grammar and Akshara-s

Indian languages are written in a number of different scripts. A few are perso-arabic while all the others are derived from the ancient braahmi script. The braahmi based scripts are phonetic in nature - there is nearly one-to-one

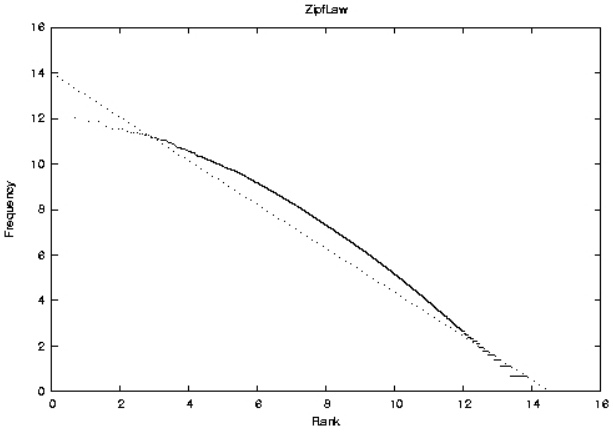

Figure 9: Zipf's Law: Dotted line is a least squares fit on whole data

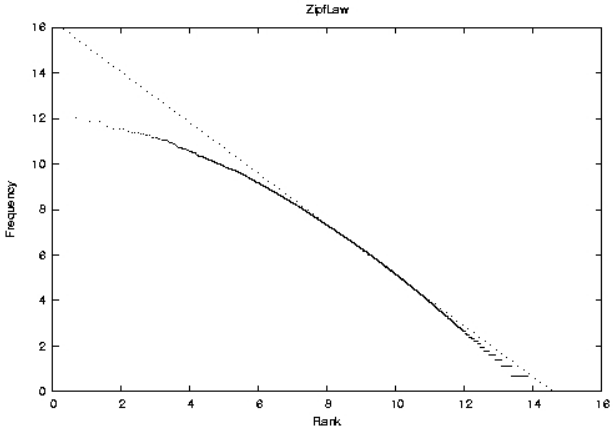

Figure 10: Zipf's Law: Dotted line is a least squares fit on middle portion only

correspondence between the written and spoken forms of language. These scripts are primarily syllabic in nature. The units of writing are akshara-s, which are in turn made up of symbols corresponding to basic sounds (phonemes) in the language. Akshara-s are basically C*V combinations where C denotes a consonant sound and V a vowel sound. With about 35 consonants, 15 vowels and 2, 3 and even 4 consonant clusters being quite common, the number of possible akshara-s is very large. Although akshara-s are really atomic units of writing in these languages, it would not be practicable to consider akshara-s as atomic units for the purpose of character encoding in computers - there would then be simply too many codes in the code space. Interestingly, these Indian scripts feature a script grammar, a grammar that specifies all valid combinations. Figure 13 gives the script grammar for braahmi based Indian scripts, adapted from the ISCII standard [7] and depicted as a finite state machine. This simple grammar accepts all valid akshara-s and rejects all invalid akshara-s, irrespective of whether these combinations occur frequently or not, irrespective of whether a combination has appeared even once in the corpus or not.

A grammar at the level of scripts is a unique characteristic of Indian languages and scripts. Telugu is normally written in the Telugu script. The script grammar given here is applicable to Telugu script also. As can be seen from this script grammar, a distinction is made between

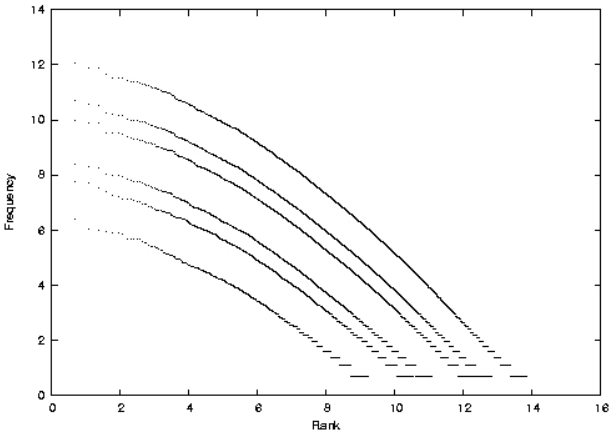

Figure 11: Zipf's Law: Effect of Corpus Size

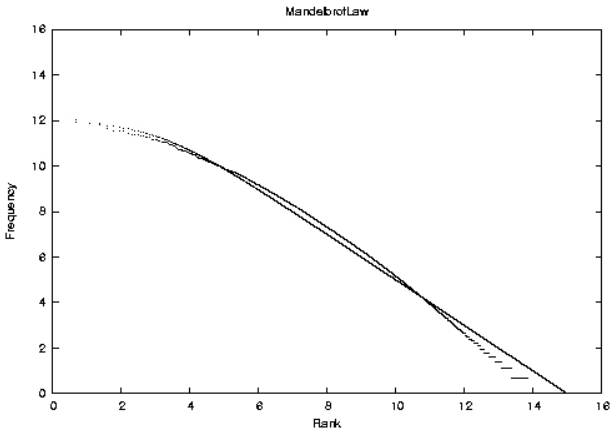

Figure 12: Mandelbrot's Law

independent vowels and vowel sounds required to pronounce consonants. The latter are called vowel-maatra-s. Pure consonants rarely occur in Telugu and most other languages and combination with the 'a' vowel sound is the most common. Hence consonants are taken to have an implicit 'a' vowel maatra and addition of any other vowel maatra replaces this default vowel. Also, an explicit halaMt symbol is needed to remove the implicit 'a' maatra while forming consonant clusters. If and when we need to depict an independent pure consonant, the convention is to use two halaMt-s in sequence. There are alternative ways of writing the script grammar but the one given here has become the standard as given in the ISCII standard [7].

We give below coverage analysis of types in terms of akshara-s. Although nearly 20000 akshara-s have appeared in the corpus, more than 95% of all words are made up of only about 5000 akshara-s. See Figure 14. The infrequent akshara-s account mostly for loan words (for example, English or Urdu words written in Telugu script), proper names, spelling errors etc. Applications such as word processors, DTP systems, font-character mapping systems and rendering engines must be designed to handle at least these 5000 most frequent akshara-s correctly.

8.1 Character Level Analysis

Akshara-s are appropriate units of writing, there is really no such thing as 'alphabet' or

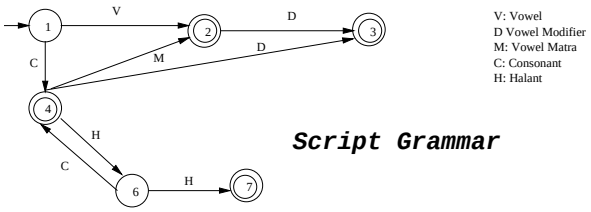

Figure 13: Script Grammar

'character' in Indian languages. If at all, we have to take the building blocks of akshara-s, namely vowels, consonants, vowel maatra-s, vowel modifiers and halaMt as characters. In English, characters are represented internally as bytes and byte-level analyses may be acceptable for such languages. It has been shown with quantitative evidence that akshara-s form appropriate units in Indian languages, not bytes [34] Yet, it would be instructive to study the distribution of sub-akshara units such as vowels and consonants. Table 5 shows the distribution of vowels and consonants as also their bigrams and trigrams in the word-initial, word-medial and word-final positions. Here we have counted vowel-maatra-s as vowels and the vowel modifiers (namely the anusvaara, visarga and ardhaanusvaara) as consonants. The halaMt is of course ignored. Also note that two or more vowels never occur together in Telugu (and other Dravidian languages) - a consonant invariably gets inserted between. Pure vowels never occur in word medial and word final positions - only vowel maatra-s are allowed here. Telugu words do not end in consonants. Consonant ending words in the corpus are mostly proper names and loan words from English.

The global vowel-to-consonant ratio is 0.80.

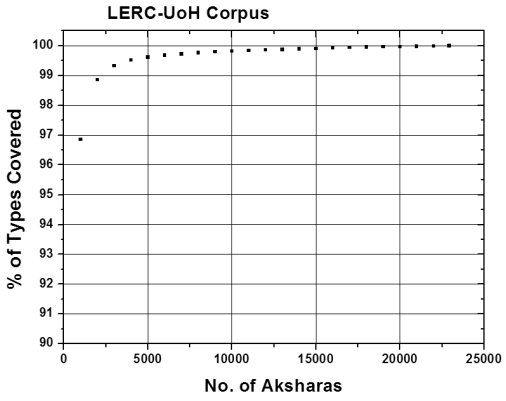

Figure 14: Coverage Analysis in terms of Akshara-s

Some languages of the world have only CV syllables and consonant clusters are not allowed. Consonant clusters are a regular phenomenon in Telugu. We also see that word initially, consonants occur more frequently than vowels in Telugu. Thus the global vowel-to-consonant ratio shows the effect of CCV, CCCV and larger consonant clusters in Telugu. Vowel-to-consonant ratio in word medial position also corroborates with this observation. Telugu words are complex.

9 Word Length Analysis

Table 6 shows the word length distribution for various languages in the DoE-CIIL corpora as also for complete LERC-UoH Telugu corpus in

Table 5: Vowel-Consonant Statistics

Word Initial	
V-s	14.22 %
Freq. V-s	a,aa,e,i,u
C-s	85.78 %
Freq. C-s	p,v,k,s,n
Freq. Bigrams	p-r,v-i,n-i,s-m,k-a
Freq. Trigrams	n-i-r,p-r-i,p-r-aa k-aa-r,t-e-l
Word Medial	
V-s	43.28 %
Freq. V-s	a,u,i,ee,oo
C-s	56.72 %
Freq. C-s	m,n,r,l,k
Freq. Bigrams	aa-m,r-m,aa-n,aa-r,r-aa
Freq. Trigrams	i-m-c,u-n-n,n-n-aa u-k-u,t-u-n
Word Final	
V-s	87.37 %
Freq. V-s	u,i,aa,oo,ee
C-s	12.63 %
Freq. C-s	m,n,l,r,k
Freq. Bigrams	n-i,l-u,n-u,r-u,l-oo
Freq. Trigrams	aa-r-u,u-l-u,m-d-i i-k-i,u-k-u

terms of akshara-s.

We can see that Telugu words tend to be long and complex. Figures 15 and 16 show the distribution of word lengths expressed in terms of akshara-s as also in terms of bytes. It can be seen that the average word length is higher than for English. The skew shows that long words are also quite common.

In terms of bytes (that is, 'characters'), the mean word length for Telugu from the LERC-UoH corpus is 11.61 and the standard deviation is 4.41. See also [18]. In contrast, English words have a mean length of 8.18 with a standard deviation of 3.12 (based on a 3 Million word English Corpus derived from the British National Corpus (BNC) by random selection). See also [13, 31].

Table 6: Word Length Analysis in Akshara-s

Language	Fam-ily	Min	Max	Mean	Std. Dev.
Hindi	IA	1	32	3.77	1.33
Punjabi	IA	1	12	3.32	1.09
Bangla	IA	1	18	4.25	1.30
Oriya	IA	1	15	4.28	1.29
Assamese	IA	1	19	4.26	1.35
Marathi	IA	1	30	4.38	1.41
Kannada	Dr	1	17	5.29	1.80
Tamil	Dr	1	16	5.34	1.63
Telugu	Dr	1	25	4.99	1.62
Malayalam	Dr	1	20	6.21	2.23
Telugu-LERC-UoH	Dr	1	39	5.64	2.06

9.1 Word Length Variation with Frequency

Words that occur frequently tend to be small words. It is therefore interesting to explore the relation between word frequency and word length. Figure 17 shows the scatter diagram of word length measured in akshara-s as a function of logarithm of word frequency. Word length is averaged over all words of a given frequency. It can be seen that the least frequent words are larger and word length shows a gradual decrease as we move towards more frequent words. High frequency words show a greater spread in terms of word length. Yet we can see a trend - words tend to become smaller and smaller as we move

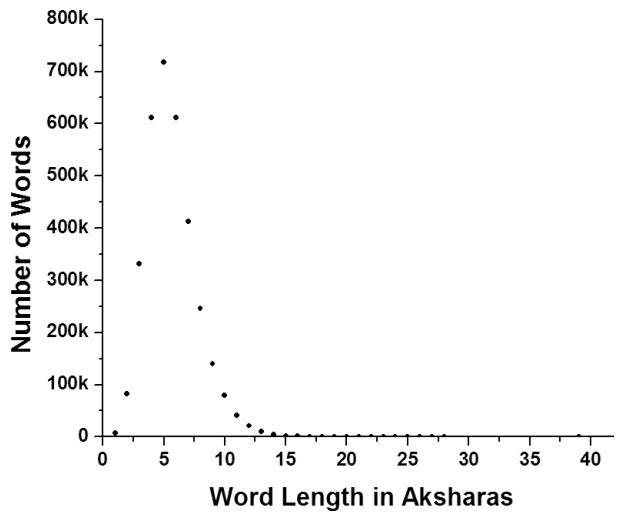

Figure 15: Word Length Analysis in terms of Akshara-s

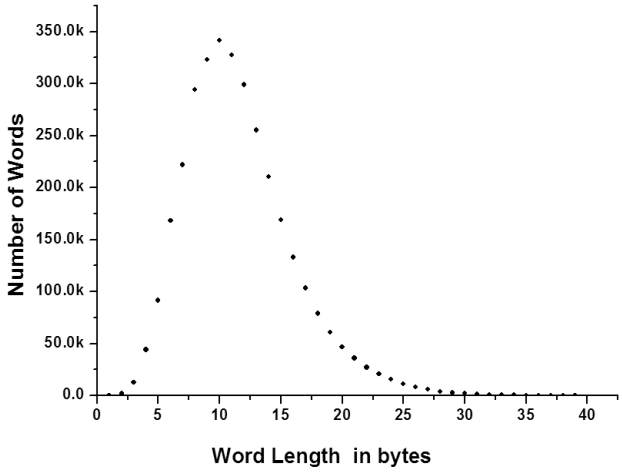

Figure 16: Word Length Analysis in terms of Bytes

towards the most frequent words. The streaks that we see are due to clustering effect due to averaging.

10 Entropy, Perplexity

Entropy is a measure of information content. Entropy is related to probability, redundancy and uncertainty and is thus invaluable in language analysis. The more we know about something, the lower the entropy will be because we are less surprised by the outcome of a trial. Entropy can be interpreted as the minimum number of bits required to encode a given piece of information. Entropy can be calculated using the formula

$$H(X) = \sum_{x=1}^{N} -p(x) \log_2 p(x)$$

where N is the number of Word Types in the language.

H-maximum will be obtained when the probabilities of all the words in the corpus are same.

$H_{max} = \log_2 N$

$H_{Relative} = \frac{H_{actual}}{H_{max}}$

$Redundancy = \frac{H_{max} - H_{actual}}{H_{max}}$

Perplexity is useful for evaluating language models. A perplexity of k means that you are as surprised on average as you would have been

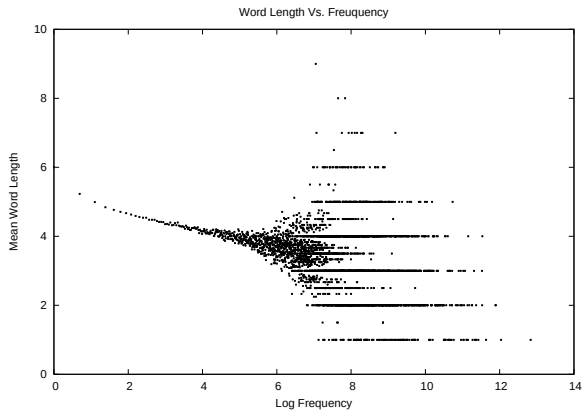

Figure 17: Word Length in relation to Word Frequency

if you had had to guess between k equiprobable choices at each step. Perplexity of the given model can be evaluated by

$$P(x) = 2^{H(x)}$$

where H(x) is the entropy of the given model.

The values of the Entropy and Perplexity for the LERC-UoH corpus are shown in table 7.

Table 7: Entropy Analysis of LERC-UoH Telugu Corpus

Entropy	15.6412
Relative Entropy	0.722
Redundancy	0.278
Perplexity	51105.821

11 Sentence Level Analysis

A preliminary analysis of sentence lengths (in words) has been performed. Full stops, exclamation marks, question marks and semi-colon are treated as sentence boundary markers. Full stop can also serve as decimal point in numbers. Simple heuristics have been used to distinguish these as also for handling abbreviations and acronyms. The distribution of sentence lengths (figure 18) shows that Telugu sentences are often quite short. The average length of a sentence in Telugu is 10.09 words, which is much smaller than the average sentence length for English[44], as can be expected (Average length of sentences in the British National Corpus is about 23 words). The distribution is skewed and the mode is 8.17.

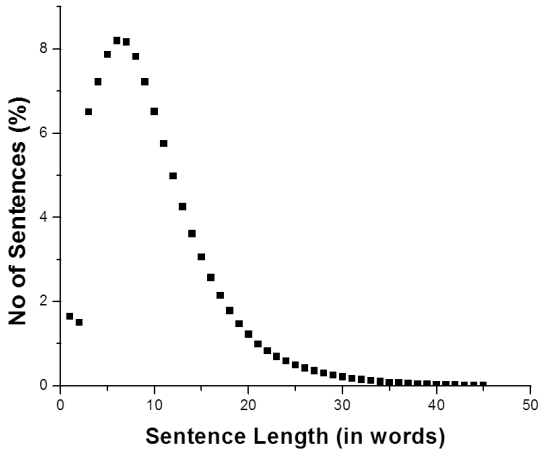

Figure 18: Sentence Length Distribution

12 Word Bigram Analysis

The LERC-UoH corpus is fairly large and only a preliminary analysis of word level bigrams has been conducted. Analysis is done in parts and bigrams with low frequency are pruned. More than 100000 bigrams with frequency of occurrence greater than 10 have been found. A quick look at these bigrams reveals some interesting observations. We can see many acronyms (ex. eM-pi (M.P.), si-eM (C.M.), es-ai (S.I.), ee-pi (A.P.)), abbreviations (ex. yai-es (Y.S. for Y.S. Rajashekhara Reddy)), phrases and multi-word expressions (mukhyamaMtri-caMdrabaabunaayuDu (Chief Minister Chandra Babu Naidu), (keeMdra-prabhutvaM (Central Government), maajii-maMtri (Ex-Minister), jillaa-pariSattu (District Council), telugudees'aM-paarTi (Telugudesam Party)), compound verbs (phiryaadu-cees'aaru ((he) complained), vyaktaM-cees'aaru ((he) expressed), viDudala-cees'aaru ((he) released), haami-icchaaru ((he) promised)), skeletal two word sentences (vileekalurutoo-maaTlaaDaaru (Talked/discussed with journalists)) etc. It would be interesting to carry out more detailed explorations using pattern matching techniques.

13 Conclusions

In this paper we have described a variety of statistical analyses of a fairly large text corpus of Telugu. It is perhaps for the first time that such a detailed statistical analysis has been carried out on a large corpus for a Dravidian language. The paper shows that Dravidian languages are more complex than Indo-Aryan languages and English at the level of words. In particular, Telugu shows a very rich system of morphology leading to long and complex words. These analyses point to issues relating to technology development as also to detailed linguistic analysis necessary for a complete understanding of the language. Larger corpora are needed in Dravidian languages for meaningful analysis and technology development. Corpus linguistics and quantitative studies are both interesting and useful and thus call for greater attention at this point of time.

References

[1] J. Aarts and W. Meijs. *Corpus Linguistics: Recent development in the use of Computer corpora in Enlgish Language Research*. Rodopi, Amsterdam, Atlanta, 1984.

[2] M. Barlow. Corpora for theory and practice. *International journal of Corpus linguistics*, 1(1):1–38, 1996.

[3] N. Bhattacharya. *Some statistical studies of the Bangla language*. PhD thesis, Indian Statistical Institute, Kolkata, 1965.

[4] D. Biber. Representativeness in corpus design. *Literary and Linguistic computing*, 8(4):243–257, 1993.

[5] D. Biber. Investigating language use through corpus-based analyses of association patterns. *International journal of Corpus linguistics*, 1(2):171–198, 1996.

[6] D. Biber, S.Conrad, and R. Reppen. *Corpus Linguistics : Investigating language strcture and use*. Cambridge Unviersity Press,Cambridge, 1998.

[7] BIS - Beareau of Indian Standards. Indian script code for information interchange

- ISCII. In *IS 13194: 1991*, New Delhi, India, 1991.

[8] B B Chaudhuri and S Ghosh. A statistical study of bangla corpus. In *Proceedings of International Conference on Computational Linguistics, Speech and Document Processing*, pages 32–37, 1998.

[9] Niladri Sekhar Dash. Corpus linguistics in india: Present scenario and future direction. *Indian Linguistics*, 64(1-4):85–113, 2003.

[10] Niladri Sekhar Dash and Bidyut Baran Chaudhuri. Bangla script: A structural study. *Linguistics Today*, 2(1):1–28, 1998.

[11] Niladri Sekhar Dash and Bidyut Baran Chaudhuri. The process of designing a multidisciplinary monolingual sample corpus. *International Journal of Corpus Linguistics*, 5(2):179–197, 2000.

[12] Niladri Sekhar Dash and Bidyut Baran Chaudhuri. Relevance of corpus in language research and application. *International Journal of Dravidian Linguistics*, 32(2):101–122, 2002.

[13] W. P. Elderton. A few statistics on the length of english words. *Journal of Royal Statistics*, Series: A(CXII):436–445, 1949.

[14] R. Garside, G. Leech, and G. Sampson. *The computational analysis of English: A corpus based approach*. Longman, London, 1987.

[15] I. J. Good. Distribution of word frequencies. *Nature*, 179: 595, 1957.

[16] K. Hofland and S. Johansson. *Word Frequencies in British and American English*. Norway Computing Center for the Humanities, 1982.

[17] Hla Hla Htay, G. Bharadwaja Kumar, and Kavi Narayana Murthy. Constructing english-myanmar parallel corpora. In *Proceedings of Fourth International Conference on Computer Applications*, pages 231–238, Yangon, 2006.

[18] B.D. Jayaram and M.N. Vidya. Word length distribution in indian languages. *Glottometrics*, 12, 2006.

[19] S. Johansson and A. B. Stenstrom. *English computer corpora: Selected papers and research guide*. Mouton de Gruyter, Berlin, 1991.

[20] G. Kennedy. *An introduction to Corpus Linguistics*. Addison-Wesley, 1998.

[21] I. Khan, S. K. Gupta, and S. H. S. Rizvi. Statistics of printed hindi text charecters: priliminary results. *Journal of IETE*, 37(3):268–275, 1991.

[22] G. Knowles, B. J. Williams, and L. Taylor. *A corpus of formal British English speech: The Lancaster/IBM spoken English Corpus*. Longman, London, 1997.

[23] Henry Kucera and W. Nelson Francis. *Computational Analysis of present-day American English*. Brown University Press, 1967.

[24] G. Bharadwaja Kumar and Kavi Narayana Murthy. Ucsg shallow parser. In *Proceedings of CICLING 2006*, pages 156–167, Mexico city, Mexico, 2006. LNCS 3878.

[25] I. Lancashire, C. Percy, and C. Mayer. *Synchronic Corpus linguistics*. Rodopi, Amsterdam, Atlanta, 1996.

[26] M. Ljung. *Corpus-based studies in English*. Rodopi, Amsterdam, Atlanta, 1997.

[27] C. Mair and M. Hundt. *Corpus Linguistics and Linguistics theory.* Rodopi, Amsterdam, Atlanta, 2000.

[28] B. P. mallik. *Sheslekha: Linguistic statistical analysis.* Bangla Academy, Kolkata, 2000.

[29] T. McEnery and A. Wilson. *Corpus Linguistics.* Edinbourgh University Press: Edinbourgh, 1996.

[30] A. C. F. Meyer. *English Corpus Linguistics.* Cambridge Unviersity Press,Cambridge, 2002.

[31] G. A. Miller, E. B. Newman, and E. A. Friedman. Length-frequency statistics for written english. *Information and Control*, 1:370–389, 1958.

[32] B. K. Murthy and W. R. Deshpande. Language technology in india: Past, present and the future. In *Proceedings of the SAARC conference on extending the use of multilingual and multimedia information technology*, Pune, India, 1998.

[33] Kavi Narayana Murthy. *Natural Language Processing - an Information Access Perspective.* Ess Ess Publications, New Delhi, India, 2006.

[34] Kavi Narayana Murthy and G. Bharadwaja Kumar. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(1):57–80, 2006.

[35] Atul Negi, Kavi Narayana Murthy, and Chakravarthy Bhagvati. Foundational issues of document engineering in indian scripts and a case study in telugu. *Vivek*, 16(2):2–7, 2006.

[36] M. P. Oakes. *Statistics for Corpus Linguistics.* Edinbourgh University Press: Edinbourgh, 1998.

[37] N. Oostdijk and P. Hann. *Corpus based research into language.* Rodopi, Amsterdam, Atlanta, 1994.

[38] K. Praksh Rao, Akshar Bharati, Rajeev Sangal, and SM Bendre. Basic statistical analysis of corpus and cross comparison among corpora. In *Proceedings of the Recent advances in Natual Language Processing (ICON-2002)*, pages 121–129, 2002.

[39] S. V. Shastri. The kolhapur corpus of indian english and work done on its basis so far. *International Computer Archieve of Modern English (ICAME) Journal*, 2:15–26, 1988.

[40] J. Sinclair. *Corpus, Concordance, Collocation.* Oxford University Press, Oxford, 1991.

[41] M. Stubbs. *Texts and Corpus analysis.* Oxford: Blackwell publishers, 1996.

[42] W. Teubert. Corpus linguistics: A partisan view. *International journal of Corpus linguistics*, 4(1):1–16, 2000.

[43] J. N. Tripathi. A statistical analysis of devanagari (hindi) text charecters. *Journal of IETE*, 17(1):25–27, 1971.

[44] C. B. Williams. A note on the statistical analysis of sentence length as a criterion of literary style. *Biometrika*, 31:356–361, 1991.

[45] G. K. Zipf. *Human Behaviour and the Principle of Least effort: An introduction to Human Ecology.* Addison-Wesley, 1949.