Frequency of occurrence of phonemes in Hindi

A project funded by AIISH Research Fund (2017-2018)

Sanction No.: SH/CDN/ARFSP4/17-18 dated 26.12.17

Total Fund: Rs. 4,98,000/-

Project Report

Principal Investigator

Dr. N. Sreedevi

Professor in Speech Sciences

Department of Clinical Services

All India Institute of Speech and Hearing

Mysuru-06

Co-Investigator

Dr.Irfana. M

Lecturer in Speech Sciences

Netaji Subhash Chandra Bose Medical College

Jabalpur-03

Research Officer

Ms. Anu Rose Paulson

All India Institute of Speech and Hearing

Manasagangothri, Mysuru 570006

Acknowledgements

Our sincere gratitude to our former Director, Prof. S. R Savithri and our present Director, Prof. M. Pushpavathi, All India Institute of Speech and Hearing, Mysuru, for sanctioning and providing all the support and help in completion of the project.

Our warm and sincere thanks to the staff of AIISH library, particularly Mr. Nithesh David Kuruvila, for his timely help with the references. We also extend our gratefulness to all the participants, Ms Priyadarshini, V, Research Officer and other staff of DCS who helped in one way or the other in completion of this project report.

Dr. N. Sreedevi
Principal Investigator

Dr. Irfana. M
Co Investigator

Ms. Anu Rose Paulson
Research Officer

Table of Contents

LIST OF TABLES

List of Figures

CHAPTER I

INTRODUCTION

Communication is an indispensable part of our lives. It includes both spoken and written language as well as several non verbal cues which help us communicate effectively. Language enables us to exchange information efficiently through speech by delivering and receiving meaningful messages. All natural languages consist of a conventional system of spoken or written symbols with a definite number of sounds or phonemes. These sounds are arranged in a specific pattern and the study of categorical organization of these phonemes in a language is called phonology. A phoneme is "an underlying representation of speech sound" (Giegerich & Giegerich, 1992). It can have allophones which are realizations of the phoneme. Each language has its own phoneme inventory and phonological system. However, there exists in every language, a standardized form of the language and its variations, known as dialects. A standardized language is a dialect of the language which has been officially accepted in the community and is used for formal communication as in administrative matters, literature and economic life. It has a definite set of orthography, grammar and standards and is different from spoken form. The spoken form of the language may vary with respect to geographical regions or society.

Hindi, an Indo-Aryan language is an official language of India and the state language of various Northern states of India i.e. Madhya Pradesh, Delhi, Uttar Pradesh, Uttarakhand, Bihar, Rajasthan, Chhattisgarh, Haryana, Himachal Pradesh and Jharkhand. It is spoken by around three million people either as their first or second language (Kachru, 2006). About 50 dialects of Hindi are spoken by people across the country (Sinha, Jain & Agarwal, 2015). Hindi is also the lingua franca of countries such as Fiji (known as Fiji Hindi), Nepal, Bangladesh and Pakistan (Meena, 2015) and many non- Hindi speaking Indian states such as Arunachal Pradesh, Andaman and

Nicobar Islands, and major cities such as Mumbai, Hyderabad, Kolkata etc . It is also recognized as minority language in Mauritius, Surinam, Guyana, South Africa, and Trinidad and Tobago (Meena, 2015). Punjabi, Assamese, Bengali, Oriya Marathi and Gujarati are genetically related to Hindi (Kachru, 2006).

The standardized form of Hindi is known as Modern Standard Hindi. It has varieties of dialects such as Khari boli, Haryanvi, Bagheli etc which is spoken in various North Indian states. Modern standard Hindi, majorly based on the Khari boli dialect, is spoken widely by the urban population, taught in schools, used in newspapers, films, dramas and television news. There are 12 vowels and 38 consonants in the phonological system of Hindi. Two of these vowels, namely, [æ] and [ɒ], are borrowed from English and may not be observed distinctly in all the native speakers. All the vowels have their nasalized counterparts as well. Consonants [f, z, ʃ] despite being loan phonemes, are well established in Modern Standard Hindi (Ohala, 2004). Geminates always occur in the medial position. Although in orthography, these are present in final position of words, they are produced as singletons in formal speech (Ohala, 2004). However, there are regional varieties in spoken form of the language and the phoneme repertoire may vary accordingly. Speech sounds in Hindi has allophonic variations as well (Pandey, 2007). Moreover, Hindi has an alphasyllabic system also known as akshara systems, which use a combination of alphabetic and syllabic systems (Pandey, 2014). The phonemes of Hindi are provided in Appendix A.

Literature reports of various studies concerned with the frequency of occurrence of phonemes in several languages. Research in the area was present as early as the 1930s. Few of the initial studies in English language were carried out by Whitney (1874) and Dewey (1923). Bhagwat (1961), Ghatage and Madhav (1964) and Jayaram (1985) were few of the preliminary studies in Indian languages. Older studies considered written materials as source such as

newspapers, scripts of dramas, books and dictionaries (Whitney, 1874; Ramakrishna et al., 1957; Ghatage, 1994; Tamaoka & Makioka, 2004). Later, spoken materials such as interviews, lectures, radio announcements, telephone conversations etc. were utilized to determine the frequency count of phonemes (French, Carter and Koenig, 1930; Voelker, 1935; Guirao & Jurado, 1990; Sreedevi & Irfana, 2013). Several authors (Ferguson & Chowdhury, 1960; Thomas, 2005; Sandoval, Toledano, Torre, Garrote & Guirao, 2008) included both spoken and written sources for analysis. There are numerous written and spoken corpus (a text data or spoken data that is representative of the actual language under investigation) available in many languages. EMILLE- CIIL corpus available for Indian languages (Hindi, Marathi, Gujarati, Punjabi, Kashmiri, Urdu, Kannada, Tamil, Telugu, Malayalam, Oriya, Assamese, Konkani, Bengali, Sanskrit, Manipuri and Indian English) is one such example. It includes both written and spoken corpus.

These data have been used extensively to determine the frequently occurring phonemes, consonant cluster groups, syllables, syllable type frequency, word frequency (in different contexts, grammatical categories) and morphophonemic categories. The information gathered is useful in areas of speech language pathology, audiology, linguistics, and speech engineering. In the field of speech language pathology, this information can be employed to develop various speech materials for assessment and in selecting target sounds for treatment of articulation errors in individuals with communication disorders. The phonetically balanced word lists that audiologists use for assessing auditory processing disorders like staggered spondaic words (SSW), for checking speech identification scores (SIS), speech in noise test (SPIN) and speech recognition scores (SRT) in routine audiological evaluations are based on such phoneme frequency information and they are highly language specific. As reported by Sinha (2015), “syllables may be considered as the basic phonetic unit for processing of Indian languages”.

Therefore, analysis of the language should not be restricted only to phoneme level. Along with data on frequently occurring phonemes, information on syllable frequency is used considerably in developing text to speech systems, translation systems and synthesis systems. The data is used to teach a foreign language as well.

Need for the study

Different languages have different phoneme systems and dialects. A database of different dialects of the language with its phoneme frequency will help us expand the knowledge of a particular language. The phoneme statistics has widespread applications in various fields. In psychololinguistics, the frequency count of phonemes may play a significant role in specific areas of language processing. A child's phoneme acquisition pattern may be associated with frequency of different phonemes in the language. It may also give us an insight into the error patterns of brain-damaged individuals with phonological paraphasias (Robson, Pring, Marshall, & Chiat, 2003). Similarly, the information may further be used to determine if the speech errors in various speech and language disorders may be biased towards phonemes with higher frequency. The information on phoneme frequency is important in creating a phonetically/ phonemically balanced (PB) word list used for speech audiometry. A PB word list truly reflects the phoneme distribution in a language. The phonetic frequency may also be utilized effectively in speech therapy. Knowledge of the phoneme frequency has been beneficial in speech technology. This data is essential for the development of speech recognition systems, text to speech synthesis systems, translation systems, etc. Machine translation system can translate by putting words together that are statistically more likely to be in a certain structure. Words prepared based on the phonemes in that particular language and its frequency of occurrence. These technologies are widely used in the area of Augmentative and Alternative Communication

for the rehabilitation of individuals with communication disorders (Cerebral palsy, aphasia etc.). Research in the area of linguistics relies on frequency count of phonemes, morphemes and syllables. Such information can also be used to teach a foreign language.

India is a country with diverse languages. Hindi, the official language of India, is widely spoken by millions of people in various states of the country. Since, it is vastly spoken in different regions; there are several variations within the language. Earlier studies on phoneme frequency in Hindi for instance, Ghatage (1964) was from various written materials in the language. Moreover, these early reports may not be apt at present as there are numerous new words, modified and borrowed words used in day to day conversation. Also, there may be a difference in the frequency of occurrence of phonemes among written and spoken data. In addition, there is limited research available in spoken data in this language. Hindi, like any other language has its own distinct set of phonemes and syllables. Unlike foreign languages, it is an alphasyllabic language. Therefore, there is a need to create a database (spoken form) for Hindi. In light of these issues, present study was planned to obtain the frequency of occurrence of phonemes in conversational speech samples in Hindi.

Aim

To establish the frequency of occurrence of phonemes, consonant clusters, syllable types and word shapes in conversational Hindi

Objectives

1. To obtain a conversational database in Hindi and calculating the frequency of occurrence of each phoneme from it.

2. To find the order of frequency of occurrence of each place of articulation, manner of articulation and vowel type.
3. From the database obtained calculating the frequency of occurrence of phoneme position in words, syllabic types and word shapes.

Implications of the study

- Armed with a list of the most frequently occurring sounds and syllables, it will be possible to plan and develop effective therapy material for children as well as adults with speech and language disorders. An issue remaining in these children following the acquisition of age adequate language skills is the articulation errors. The target sounds for articulation therapy can be selected based on the frequently occurring sounds in the language.
- Information on frequency of occurrence of sounds and syllables of a language would also be helpful in developing stimulus material for determining speech discrimination scores and in preparation of central auditory processing disorders (CAPD) tests. It can be ensured that the PB word lists developed would have the same distribution of sounds and syllables as in the natural language.
- Sequence of acquisition of speech sounds in typically developing children and other aspects of phonological acquisition may be related to the frequency of occurrence of phonemes in a language (Leung & Law, 2004).
- Knowledge about the frequently occurring phonemes in a language can help understand the speech errors in various speech and language disorders. For instance, articulatory errors are more in individuals who misarticulates a frequently occurring phoneme in the language compared to an individual who misarticulates a sound that is less frequently

present in the language. Many a times, it is observed that individuals misarticulate sounds only in certain word positions. Therefore, information regarding phoneme position is also important. It may also help in analysis of neurolinguistic data in brain damaged individuals (Robson, Pring, Marshall & Chiat, 2003; Zangwill, 1975).

- The information may also be used to study certain aspects of language processing (Leung & Law, 2004).
- The result provides a database for the training of automatic speech recognition systems such as text-to-speech systems, synthesis systems etc. These systems are also used in the area of Augmentative and Alternative Communication.
- The findings have wide application in the field of linguistics as it can provide guidelines for learning a foreign language.

CHAPTER II

REVIEW OF LITERATURE

There are many variations among the phonological system of world languages. The distribution of phonemes, consonant clusters and the syllable inventory in a language can vary. Higher occurrence of these units indicates that they are most frequently produced by the speaker of that language and are thus more familiar with the articulation of these units. Research in this field was present as early as 1930s. Several studies have recorded data on the frequency count of phonemes, consonant clusters and syllables in various languages and dialects. Source of these studies were printed material such as newspapers, dictionaries, scripts of plays etc. or spoken materials, for instance, radio announcements, lectures, interviews, general conversations etc. Few researchers had shown interest in comparing and documenting both written and spoken data. The information documented has been used extensively in fields of linguistics, speech-language pathology and audiology, speech engineering and second language learning.

The literatures reviewed are discussed under the following headings:

I. Frequency of phonemes in non- Indian languages
II. Frequency of phonemes in Indian languages
III. Comparison across various Indian languages
IV. Comparison of frequency of phonemes across various languages of the world

I. Frequency of occurrence of phonemes in non- Indian languages

1. 1. Indo- European languages

1. 1. 1. English

One of the earliest and oldest studies on frequently occurring phonemes was cited by Gerber and Vertin, 1969. Whitney (1874) studied frequently occurring phonemes in American

English from ten literary classics. While phonemes /r/ /n/, /I/, /t/, /ə/, /d/, /s/ and /ɑ/ were frequently used, /ʒ/, /θ/, /j/ and /g/ were least present in the language. There was a good correlation between the reports of Dewey (1923) and Whitney (1874) since both the studies computed the frequently occurring phonemes from written material of English.

Dewey (1923) investigated the relative frequency of occurrence of English phonemes and syllables from various written and printed sources. Among the total phonemes /i/ (8.12%), /n/, (7.38%), /t/ (7.27%), /r/ (7.02%), /s/ (4.64%), /d/ (4.39%), /a/ (4.04%) and /l/ (3.82%) were the frequently present phonemes in the language. Vowels /ə/, /I/, /i/, /e/ and consonants /n/, /t/, /r/, /d/, /l/, /th/ was the most used phonemes in English language. The highly frequent phonemes and their number of occurrences in various syllable and word levels are provided in table 2.1 and 2.2. The most commonly observed syllables (occurring more than 100 times) were /thi/, /ev/, /in/, /and/, /i/, /ta/, /iŋ/ and /ər/.

Table 2.1

Most frequently occurring phonemes in various phoneme positions at syllable level in English (Dewey, 1923)

Initial position		**Medial position**		**Final position**	
Phonemes	**Occurrences**	**Phonemes**	**Occurrences**	**Phonemes**	**Occurrences**
/i/	9,943	/i/	10,010	/t/	14,520
/t/	9,620	/r/	9,370	/n/	14,200
/s/	7,850	/e/	8,740	/r/	12,530
/h/	6,755	/n/	8,460	/d/	11,240
/a/	6,109	/a/	7,945	/z/	10,220
/b/	6,090	/u/	7,040	/s/	7,000

Table 2.2

Most frequently occurring phonemes in various phoneme positions at word level in English (Dewey, 1923)

Initial position		Medial position		Final position	
Phonemes	**Occurrences**	**Phonemes**	**Occurrences**	**Phonemes**	**Occurrences**
/i/	6,730	/i/	18,750	/t/	11,770
/a/	6,690	/n/	15,650	/d/	9,810
/s/	5,575	/ə/	14,995	/z/	9,525
/w/	5,480	/r/	10,040	/n/	8,740
/t/	5,150	/t/	9,630	/r/	8,370
/θ/	5,120	/a/	8,071	/v/	5,315

French, Carter and Koenig (1930) transcribed colloquial telephone conversations of native American English adults and determined the relative occurrence of vowels, consonants and consonant clusters in the initial and final positions and type of phoneme structure in the language. A total of 80,000 words were obtained from conversations and the data was compared with the results of previous studies based on written matter. The vowel in the word ‘pin’ occurred most frequently for about 11.22%, followed by 7.9% and 6.40% for vowels in ‘pen’ and ‘pan’. The least frequently occurring vowel was in ‘poise’ (0.24%). Among the initial consonants, the per cent of occurrence ranged from 8.26% for -/w/- , followed by -/t/- (7.09%) to Z (0.47%). Consonant ‘r’ occurred as frequent as 13.87% as a final consonant followed by -/t/- (11.98%) and ‘the’ (0.06%) was the least frequent sound. In the initial consonant cluster, -/pr/- (1.27%) was the most frequently present cluster and -/kl/- (0.26%) the least frequent. Compound phoneme ‘no’ is present for 4.68% (most frequent) while -/rd/- occurred only for about 0.51%. A similar range of percentages was present while comparing written data and the information from the present study. However, few marked differences have been observed, for instance, unaccented vowel as in ‘about’, initial consonant -/y/- and the final consonant -/t/- are

predominant in conversations. CVC (33.5%) type of syllable is commonly found closely followed by CV (21.8%) and VC (20.3%) types.

Voelker (1935) utilized data of radio announcements and carried out a frequency count of phonemes in American English. The highest frequency count were for the phonemes /I/> /n/, /t/, /ə/, /r/, /d/ and /s/. Phonemes /ʒ/, /ʌ/, /θ/ and /ʃ/ were ranked the least (as cited in Gerber & Vertin, 1969).

Crystal (1981) reported the data for relative frequency of phonemes in English by Fry (1947). The frequently occurring phonemes are provided in table 2.3.

Table 2.3

Most frequently occurring phonemes in English (Crystal, 1981)

Vowels	**Percentage of occurrence**	**Consonants**	**Percentage of occurrence**
/ə/	10.74	/n/	7.58
/I/	8.33	/t/	6.42
/e/	2.97	/d/	5.14
/aI/	1.83	/s/	4.81
/ʌ/	1.75	/l/	3.66

Few of these commonly used phonemes are acquired early by children, for example, /t/, /d/, /n/, /I/ while few of the least frequent phonemes such as /ʒ/, /θ/, /ʧ/, /ʤ/, /ɔI/ turn up later in speech.

Mader (1954) collected interview samples from children of grade one, two, and three. The samples were typewritten and analyzed for frequency and position of occurrence of the consonant sounds in the recorded words. The information from the present study was compared with studies of Voelker (1934), Travis (1931) and French, Carter and Koenig (1930). [n], [t], [d], [r] and [s] were the most frequently occurring phonemes irrespective of position in which they occur and these made up 49% of the recorded utterances. These results were in close

resemblance to the study by Voelker and Travis. The disparity between Mader's study and the study by French, Carter and Koenig (1930) may be due to the difference in the material analyzed and age of participants. The least occurring phonemes were /ʧ, ʒ, ʤ, hw/. It was also observed that the sounds in English did not occur equally in the initial, medial or final position of words.

Tobias (1959) attempted to record the relative frequency of occurrence of sounds of American English using the data put forth by French, Carter and Koenig (1930). The word lists were retranscribed according to the General American pronunciations as given by Kenyon and Knott (1951). The first four most frequent phonemes and two of the least frequent phonemes are given in table 2.4.

Table 2.4

Highly frequent and least frequent phonemes in English (Tobias, 1959)

Most frequent phonemes	
Phoneme	Percentage of occurrence
[I]	9.22%
[t]	9.11%
[ə]	7.22%
[n]	6.43%
Least frequent phonemes	
[ɔI]	0.06%
[ʒ]	0.00%

Relative frequency of phonemes in General American English was determined by Hayden (1950). The data considered were recorded lectures. The frequently occurring phonemes in decreasing order from 9.96% to 3.09% are as follows: /ə/, /I/, /n/, /t/, /r/, /s/, /l/, /ɖ/, /d/ and /æ/. These phonemes made up almost 60% of the data. The least frequent phonemes with occurrence percentage less than 1% were /š, ŋ, ě, ǰ, θ, w̲, ž/.

Wang and Crawford (1960) performed a statistical analysis to clarify the disagreement between studies regarding frequency count of English consonants. The study compared data from ten different authors (Tranka, 1935; Fowler, 1957; Carroll, 1952; Hayden, 1950; Whitney, 1874; Dewey, 1923; Voelker, 1937; French, Carter and Koenig, 1930; Fry, 1947 and Tobias, 1959). The consonants were ranked from the data collected. Most of the studies have consonants /t, n, r, s, d, l/ as the most frequently present phonemes in English, irrespective of the style and dialectical variations and literary content. The study concluded that even with a sample size of as small as 6000 items (Whitney, 1874) the frequency data remained stable. Nevertheless, the type of corpus (dictionary/ running texts) and transcriptions does cause discrepancy among the studies.

Denes (1963) derived a variety of statistical information on spoken English (conversations and narratives from 'Phonetic Readers'). The top 5 frequently occurring vowels and consonants are provided in table 2.5. Vowels /ə, I, ai/ constitute 50% of the vowels in the data while among consonants /t, n, s, d, l, m/ make up half of the consonant occurrences. Most frequent initial and final vowels were /ə/, /i/, /ai/, /e/, /i:/. Front vowels and plosives were more common. Frequency of semi-vowels and liquids were comparatively less. Phonemes /ð/, /w/, /s/, /h/, /t/, /m/, /b/ were most frequently present in initial position of words whereas /t/, /n/, /s/, /d/, /l/ commonly occurred in the final position. The least occurring phonemes include the diphthongs /oə/, /oi/ and consonants /ʃ/, /θ/, /ʧ/, /dz/ and /ʒ/.

Table 2.5

Top five frequently occurring vowels and consonants in English (Denes, 1963)

Vowels	**Percentage occurrence**	**Consonants**	**Percentage occurrence**
/ə/	9.04	/t/	8.04
/i/	8.25	/n/	7.08

/ai/	2.84	/s/	5.08
/e/	2.81	/d/	4.17
/i:/	1.78	/l/	3.68

Considering stressed and unstressed syllables, vowels /ə/ and /i/ are more common in stressed syllables.

Delattre (1965) studied data taken from dramatization and narratives which was presumed to represent connected speech. It was carried out to see frequency count of the vowels and consonants in English. Among the total phonemes, /t/ accounted for 7.85% of the total occurrences followed by /ə/ (7.76%), /n/ (7.04%), /I/ (5.57%), /r/ (5.11%), /l/ (4.72%) and /s/ (4.59%). The least frequently used phonemes in the language were fricatives /θ, ʃ, ʤ, ʧ, ʒ/. Vowels /ɔ/, /I/, /æ/, /i/ and /ɑ/ and consonants /t/, /n/, /r/, /l/, /s/ and /d/ were most commonly used (as cited in Edwards, 2002). This study is in consonance with the data provided by Hanna, Hanna, Hodges and Rudorf (1966) and Dewey (1923).

Mines, Hanson and Shoup (1978) obtained 1,03,887 phonemes from conversational American English through interviews. Phonemes /a, n, t, i, s, r, l, d, Ɛ/ were the top 9 phonemes, which constituted for 47% of the data. Research conducted by Gerber and Vertin in 1969 compared the data from Whitney (1874), Dewey (1927), Voelker (1935), French, Carter and Koenig (1930) Tobias (1959) and Denes (1963) to compare validity of the methods used to analyze the frequently occurring phonemes. The source of these data varied from written literary classics, newspapers, text books and magazines to spoken telephone conversations, radio announcements and religious and scientific speeches. The materials transcribed included General American (Whitney, 1874) Eastern American (French, Carter and Koenig, 1930) and British English (Denes, 1963) pronunciations. The results revealed that there was high correlation among the six studies investigated and the methods used were valid. It was also interesting to

note that there was a good correlation between spoken languages regardless of the various dialects than spoken and written data of same dialect. To be more specific, spoken American English is similar to spoken British than written American.

1. 1. 2. French

Malecot (1974) analyzed the phonology of the dominant dialect of modern spoken French (Paris corpus). The computer-based analysis elicited data on occurrences of phonemes in word-initial and final positions, in consonant clusters in different positions, and number of consonant clusters created by word junctures within utterances. The results obtained are provided in table 2.6 and 2.7. The first five frequently occurring phonemes have been presented in decreasing rank order. Total nasal vowels (/ɑ̃, ɛ̃, ɔ̃, œ̃/) occur the most among vowels followed by front vowels (/i, e, ɛ, y, ə, œ/). Among consonants, fricatives (/f, s, ʃ, v, z, ʒ/) has 43,014 occurrences followed by stops (/p, t, k, b, d, g/) with 37,229 occurrences. Front consonants had relatively more occurrence than back consonants. Apicals had 51,498 occurrences which were high while considering the place of articulation, followed by labials (24,183) and uvular- r (14,265). The least frequent phonemes were /ɲ/ and /ŋ/.

Table 2.6

Frequently occurring phonemes in French (Malecot, 1974)

Phonemes	Occurrences
/a/, /ɑ/	16,216
/e/	16,051
/r/	14,265
/s/	12,293
/l/	11,179

Table 2.7

Frequently occurring phonemes in different phoneme positions in French (Malecot, 1974)

Initial position		Final position	
Phonemes	**Occurrences**	**Phonemes**	**Occurrences**
/s/	6,637	/ə/	8,334
/d/	6,485	/e/, /ɛ/	7,413
/p/	5,791	/r/	4,559
/a/, /ɑ/	5,318	/a/, /ɑ/	4,540
/k/	5,094	/i/	3,381

Considering different word positions, two-consonant clusters were the most in frequency followed by three-cluster consonants. Four and five-consonant clusters were present only in intervocalic position. In the initial position, there were 53 two-consonant clusters (5,835 occurrences). There were 6,591 occurrences of 162 different types of two- consonant clusters in the intervocalic position. In the final position, there were 39 different clusters of two-consonant clusters.

1. 1. 3. Spanish

Guirao and Jurado (1990) calculated phoneme frequency distribution of American Spanish in isolation, syllable, word positions and articulatory classes. The data was collected from modern plays in Spanish. Results revealed that both vowels and consonants were almost equally distributed, 48% and 52% respectively. Vowel /e/ was ranked first (15%) followed by /a/ > /o/ > /s/ > /n/ > /i/. The frequency distribution of vowels in Spanish was /e/, /a/, /o/ and that of consonants was /s/, /n/, /r/, /t/. The least frequent phonemes were /ɲ/ and /ɾ/ (0.2%). Among the vowels, /e/, /o/ and /a/ occur more frequently in final position and /u/ and /o/ occur mostly in initial position. The most frequent syllables in the language were /a/ > /ke/ > /no/ > /de/ > /se/ > /es/. These results are in consonance with studies by Guirao and Borzone (1972) and Thomas (2005). Relative frequency of phonemes present in the first 27 frequently occurring syllables is

given in table 2.8. Considering word position, stop /k/ occurred for 12.6% in initial position while vowels /e/ was present most frequently in final position (21.7%).

Table 2.8

Most frequently occurring phonemes in syllable initial and final positions in American Spanish (Guirao & Jurado, 1990)

Initial position	**Percentage of occurrence**	**Final position**	**Percentage of occurrence**
/t/	9.9	/a/	21.7
/s/	9.8	/e/	21.3
/k/	9.4	/o/	17.2
/d/	8.3	/s/	10.9
/n/	7.5	/i/	9.2

Percentage of occurrence in terms of articulatory classes is provided in table 2.9. Voiced consonants (29.1%) are more frequent than voiceless consonants (22.5%).

Table 2.9

Frequently occurring phonemes according to place and manner of articulation in American Spanish (Guirao & Jurado, 1990)

Place of articulation	**Percentage**	**Manner of articulation**	**Percentage**
Dentals (n, l, r, ɾ, d, s)	34.7	Unvoiced stops/ fricatives (p, t, k, s, f, x, ʧ)	11.5
Labials (b, p, m, f)	9.5	Nasals (m, n, ɲ)	10.5
Velars (x, k, g)	6.1	Liquids (l, r, ɾ)	9.7
Palatals (ɲ, ʒ, ʧ)	1.4	Voiced stops (b, d, g)	8.2
		Voiced fricative (ʒ)	0.7

Sandoval, Toledano, Torre, Garrote and Guirao (2008) created an inventory of spoken Castilian Spanish which included the frequently occurring phonemes and syllables from two important spontaneous spoken corpus of the language. The analyzed data from the spoken corpus was also compared with written corpus. Table 2.10 and 2.11 shows few of the frequently occurring phonemes and syllables (in descending order of frequency) in Castilian Spanish. It was

concluded that there are differences between the frequency of occurrence of phonemes and syllables in oral and written corpora which may be due to lexical variety.

Table 2.10

Five most frequently occurring vowels and consonants in spoken and written corpus of Castilian Spanish (Sandoval, Toledano, Torre, Garrote & Guirao, 2008)

Vowels	Spoken (%)	Vowels	Written (%)	Consonants	Spoken (%)	Consonants	Written (%)
E	15.12	A	12.89	S	8.11	S	7.33
A	12.27	E	12.74	N	7.05	ɾ	6.19
O	10.38	O	9.32	ɾ	5.12	l, d	5.46
I	7.22	I	7.59	t, l, k	4.52	T	4.31
U	3.14	U	3.04	D	4.36	K	3.80

Table 2.11

Five most frequently occurring syllables in spoken and written corpus of Castilian Spanish (Sandoval, Toledano, Torre, Garrote & Guirao, 2008)

Phoneme	**Spoken (%)**	**Phoneme**	**Written (%)**
.a.	4.94	.de.	4.49
.ke.	3.77	.a.	3.55
.de.	3.51	.la.	2.61
.es.	2.45	.ta.	1.77
.i.	2.34	.ke.	1.70

1. 1. 4. Dutch

Dutch phoneme and syllable frequencies were reported by Zuidema in 2009. The phonemes/syllables were analyzed from a large spoken corpus in Dutch language. Phoneme /@/ was most frequent. Next in rank were phonemes /t/, /n/, /d/, /r/, /s/ and /A/. Syllables type CV was the most common in occurrence. Syllable types CVC, VC, CVCC, CCV, CCVC were the next in frequency. CCCVCCCC and CCC type of syllables were the least frequent. Syllable /j@/

had the most frequent occurrence in the data followed by syllables /en/, /x@/, /d@/, /Ik/ and /t@/.

1. 1. 5. Romanian

Renwick (2011) determined the type frequency of vowels and consonants from a word list in Romanian language. Vowel /i/ accounts for 25% of the total vowel characters followed by vowels /a/ (19.9%) and /e/ (20.4%). The least frequently occurring vowels were /ə/ (5.3%) and /ɨ/ (1.7%). Most frequent consonants in Romanian are as follows: /r/ > /t/ > /n/ > /l/ > /c/ > /s/ and together these phonemes make up for 67.4% of the consonants present in the language. Vowels in Romanian language have grammatical functions and this correlated well with the frequency of occurrence of these phonemes. Interestingly enough, the top vowels /i, a, e/ act as morphological markers in many instances. However, there are a few exceptions as well.

1. 1. 6. Latin

Written data from Latin textbooks were analysed by Tambovtsev (2007) to establish the frequently occurring phonemes in antique language of Latin. Forelingual sounds (t, d, ʧ, s, z, n, l, r- 37%) were more in frequency followed by sonorants (m, n, l, r, j- 23%) and occlusive non-sonorants (p, b, t, d, k, g- 22%). This is in consonance with Zipf's data (Zipf & Rogers, 1939).

1.1. 7. Russian

Smirnova and Chistikov (2011) used a computer aided tool for processing large Russian text and spoken corpora (classics, playwrights and transcripts of interviews and public lectures) in order to determine the frequency of occurrence of phonemes, syllables etc. Results revealed phonemes /a, i, t, j, o, n, s/ were most frequently present in Russian. The nine least frequent phonemes were /c, k', p', sc, b', z', g', f', h'/.

1. 2. Sino- Tibetian languages

1. 2. 1. Hong Kong Cantonese

Leung and Law (2004) determined the frequency of occurrence of onset, coda, rimes and tones in a widely spoken dialect of Hong Kong Cantonese. The data was taken from Hong Kong Cantonese adult language corpus. The level tones were more frequent (34.65%) followed by falling tones (29.09%) and rising tones (27.19%). In the onset inventory, 54.09% were coronal consonants, 20.78% were velars while the least occurring were glottal sounds (10.29%). Unaspirated sounds occurred more frequently than their aspirated counterparts. Considering nasal consonants, [m] was most frequent as a rime and least frequent as a coda. Open syllables were twice as frequent as closed syllables.

1. 3. Altaic languages

1.3. 1. Japanese

The frequency of occurrence of Japanese phonemes in various phonological contexts (word initial, intervocalic, word final) is reported in literature. In the word initial position and intervocalic position, phonemes /k/, /š/ and /s/ were most common whereas /n/ was the only phoneme present in the word final position (as cited in Broeder & Murre, 1999).

Frequency of occurrence of Japanese phonemes, morae and syllables were studied by Tamaoka and Makioka (2004). A lexical corpus created from Asahi newspaper by Amano and Kondo (2000) was used for the study. The results yielded were interesting. The vowels /a/, /i/, /u/ and /o/ had similar frequency counts while vowel /e/ was the least frequent. Long vowel /aʀ/ was more common. Consonants /k/ (17.24%), /t/ (15.53%) and /r/ (13.11%) were used much more frequently than bilabials /p/ and /b/. Special sound /ɴ/, /ʀ/ and /ǫ/ appeared frequently in the language. Most commonly observed syllable was with the special sound /ɴ/, /k/+V (vowel) +/ɴ/ (20%). Few of the frequent bi-mora in Japanese was /ka/+/i/, /te/+/i/, /ko/+ /u/, /se/+ /i/.

1. 4. Paleo- Siberian languages

1. 4. 1. Korean

Shin (2010) investigated the occurrence of Korean phonemes through analysis of items in Korean dictionary. Among the phonemes, /a/ had highest frequency of occurrence while /ɯi/ had the least number of occurrences. Considering only the consonants, highest ranking was for /k/ followed by /n/, /ŋ/, /l/ and /tɕ/. There were 1,283 syllable types present in the data and among these, most frequent types (82 types) accounted for 50% of the data. The high occurring syllables were mostly of V and CV type and the lowest was for GV (Glide-Vowel) type. Syllable /ha/ had highest occurrence followed by /li, tɕi, i, ki, sɑ, tɕʌk/ and so on.

Shin in 2008 analyzed frequency of occurrence of Korean phonemes and syllables in spontaneous speech of adults. An overall of 47% of vowels and 53% consonants were present in the utterances. Phoneme /a/ was ranked highest and /n, k, i, l, ʌ/ were next in rank. Highest frequency was assigned to CV type (62.2%) and lowest for VC (3.9%) type. Most of the commonly occurring syllables were part of grammatical words and morphemes.

1. 5. Kra- Dai languages

1. 5. 1. Thai

Munthuli, Tantibundhit, Onsuwan, Kosawat and Wutiwiwatchai (2015) performed a statistical analysis of large scale written and spoken Thai corpora, to determine the frequency of occurrence of vowels, consonants (in initial and final positions), lexical tone and syllable type. The study also compared the statistics across the corpora. The results revealed discrepancies in frequencies of initial and final consonants and vowels across the corpora. Nevertheless, few of the top phonemes have been indicated. /s, t^h, k, k^h, l, ʔ/ are few of the most frequently present initial phonemes and /ŋ, n, j, t/ as the four most common final consonants. Among the vowels, /a,

a:, i:, ɔ:/ are commonly present in the database. On an interesting note, there was a stable agreement in the frequency of occurrence of lexical tones and syllable types. Mid tones occur more frequently followed by low tone. Syllable type CVC has the highest frequency closely followed by CVVC.

1. 6. Uralic languages

1. 6. 1. Hungarian

Vowel count in Hungarian language was carried out by Lotz in 1952. Written texts were considered for the same. Vowel /ɔ/ had the highest occurrence (22.48%) followed by /æ/ (15.85%), /o/ (11.10%), /E/ (10.79%) and /ɑ̈ / (10.62%).

Tarnóczy (1961) compares frequency of occurrence of phonemes in Hungarian language by Nemes (1934), Mikes (1935, 1936, 1937), Tarnóczy (1952, 1954), Tolnai (1906) and Ve´rtes (1953). Slight variations were observed among the various commonly occurring phonemes among various authors. The frequently occurring phonemes reported by different authors in Hungarian are given in table 2.12.

Table 2.12

Comparison of frequently occurring phonemes reported by various authors in Hungarian (Tarnóczy, 1961)

Phonemes	Nemes	Mikes	Tarnóczy	Tolnai	Ve´rtes	Average
ɛ	11.7	11.21	10.8	10.67	10.28	10.93
ȧ	9.9	9.37	9.9	9.61	10.71	9.90
t	8.0	7.70	7.15	8.04	7.62	7.70
l	6.1	6.27	5.9	5.95	4.98	5.84
n	5.6	5.81	5.75	5.49	5.67	5.66
k	4.8	5.73	5.55	5.16	5.29	5.30
o-o:	5.0	4.61	5.3	5.28	5.31	5.08
i	4.5	4.91	4.5	4.46	5.30	4.73
r	3.95	4.03	4.4	4.19	4.50	4.21
m	4.35	3.69	4.1	4.17	4.3	4.12
ʃ	3.35	4.41	3.8	3.77	3.66	3.80

1. 7. Niger- Congo languages

1. 7. 1. Setswana language

A set of conversations in Setswana language were examined by Palai and O'Hanlon (2004). 53.9% of the sounds were present in the initial position and 41.8% in the medial position while only 4.1% were present in final position. In the data, /l/ was the most frequent phoneme. Other phonemes next in rank were /n/, /x/, /r/, /b/, /m/, /k/, /ʧ/ and /s/. Consonants /n/, /x/ and /b/ /l/, /r/, /m/ occurred the most in initial position and medial positions respectively. Except /ŋ/ (4.07%) no other sounds occurred in the final position.

1. 7. 2. Nko script

Nko script was created in order to write Bambara language which is one of the Manding languages spoken in Mali. Text books and dictionary written in Nko script was analysed by Rovenchak (2011) to determine the phoneme, syllable and tone frequencies. Consonants /l/, /k/, /d/, /n/ and /m/ were used more frequently while /p/, /g/, /z/, /rr/ and /v/ were the least frequent with percentage of occurrence less than 1%. Consonant /k/ was common in word initial position while there were no words starting with the phoneme /rr/. Among the vowels, /a/ was observed to be the most frequent in initial position of words. Vowels with high and low tones were predominant in the data. Long nasal vowels were rare. A high frequency of occurrence was observed for CVCV structure with same vowels.

1. 8. Austronesian languages

1. 8. 1. Ilocano language

Sagon (2006) developed phonetically balanced word list in Ilocano language (spoken in Philippines) based on frequency of occurrence of phonemes and syllable structures in the language. The frequencies of phonemes and syllables types were determined from written

articles in Ilocano. The most frequently occurring phonemes and syllable structures are provided in table 2.13 and 2.14 respectively. As in spoken English (Denes, 1963) /t/ and /n/ were the two most highly ranked phonemes in Ilocano language. Sounds /v/, /z/, /j:/ and /f/ occur very rarely or has no occurrence at all in the language.

Table 2.13

Frequently occurring vowels and consonants in Ilocano language (Sagon, 2006)

Vowels	**Frequency of occurrence**	**Consonants**	**Frequency of occurrence**
/a/	0.2235	/t/	0.0834
/i/	0.1241	/n/	0.0723
/u/	0.0367	/k/	0.0488
/o/, /e/	0.0341	/d/	0.0362
/ay/	0.0193	/s/	0.0361
/aw/	0.0056	/g/	0.0352

Table 2.14

Frequently occurring syllable types in Ilocano language (Sagon, 2006)

Syllable type	**Frequency of occurrence**
CVCVC	0.2447
CVCV	0.1809
CVCCV	0.1330
VCV	0.1011
VCVC	0.0851
CVCCVC	0.0798

II. Frequency of occurrence of phonemes in Indian languages

Being a multilingual country, there is a need to obtain information on frequency of occurrence of phonemes in various languages to understand them better. Study by Bhagwat (1961) in Marathi and Ghatage (1964) in Hindi, Jayaram (1985) in Kannada and Ghatage (1994)

in Malayalam were few of the initial researches carried out in Indian languages. Earlier studies considered only written materials to determine the phonemic and morphemic frequencies in different languages. However, there were studies which were conducted in the recent past for instance, Sreedevi, Smitha and Vikas (2012) in Kannada, Sreedevi and Irfana (2013) in Malayalam which used spoken samples such as narrations and conversations as the source material.

2. 1. Indo- European languages

2. 1. 1. Hindi

Study by Ghatage (1964) was one of the first studies to explore the phonemic and morphemic frequencies in Hindi language. Written materials were considered for the study. Khan (1990) reported vowel /a/ to be highly occurring in among the vowels. Phoneme /k/ was the most frequently occurring consonant in written Hindi. Considering the place of articulation, dentals (16.32%) had a higher occurrence followed by velars (12.59%) and labials (9.65%). Phonemes /k, h, s, m, p, n, ɲ, b, d, w/ and /r, n, k, ʈ, s, j, h, l, m/ had the highest occurrence in initial and final positions. Velars had higher mean percentage of occurrence in the initial position while stops were predominant in final position. The study also concluded that the commonly occurring syllable types in Hindi were CV followed by CVC.

350K and 50K Hindi text corpora were analyzed by Chourasia, Samudravijaya, Ingle and Chandwani (2007). The phoneme data were collected in order to train speech recognition systems. Results revealed vowels /a/ and /A/ to be most frequent in both the corpora. In the 350K corpora, few of the frequently occurring phonemes were /a/> /A/> /e/> /I/> /r/> /k/> /y/. Phonemes /a/> /A/> /r/> /e/> /k/> /i/> /I/> /n/ were the common ones in 50K corpus. It was also observed that /dh/ was the rarest sound followed by /jh/.

Various frequently occurring biphone (a sequence of two phonemes) and triphone (a sequence of three phonemes) pairs were determined. Phoneme pair /ar/, /ra/ and /an/ occurred the most in the corpora. Phoneme triplets (h,E,sil), (k, a, r), (a, r, a) and (p, a, r) were most frequently occurring in 350K and 50K corpora.

De (1973) analyzed the frequency of Hindi phonemes and syllable structures for development of speech audiometry material. The percentage of vowels and consonants are provided in table 2.15.

Table 2.15

Most frequently occurring vowels and consonants in Hindi (De, 1973)

Vowels	**Percentage occurrence**	**Consonants**	**Percentage occurrence**
/a:/	24.63%	/b/, /s/	7.8%
/a/	23.27%	/p/	7.5%
/e/	11.65%	/m/	5.5%
/i:/	8.27%	/k/	5.4%
/i/	7.86%	/g/, /ʧ/, /t̪/, /b^{h}/, /r/	4.0%

The results also revealed that nasalized /o/ and /a/ were the most frequent among the nasalized forms of vowels. CVC syllable structure which was the most frequently occurring, had 45% of occurrence. CV and VC syllable structures had 30% and 20% occurrence respectively. Rest of the syllable structures accounted for only 5% of occurrence. In Hindi, syllable structures CV and C_1VC_2 have the highest frequency of occurrence (Sinha, 2015).

Shailaja, Manjula and Praveen (2011) studied phonotactics in Hindi speaking typically developing children and children with phonological impairment. Results revealed that syllable shapes CV, CVC, VC, V, CCV, CVCC and VCC were evident in their speech. Syllable shape CV was predominant in both the groups followed by CVC type. Word shapes CV and CVC had higher occurrence among the monosyllables and CVCV had a higher percentage of occurrence

among the disyllables. CV, CV, CV was highly present among the trisyllables. The VC structure was seen much less in Hindi speaking children. Malviya, Mishra and Tiwary (2016) carried out an analysis of Hindi phonemes from large Hindi EMILLE corpus. Phonemes /a:, k, r, e, i, n, i:, t̪, s/ were most commonly present in the corpus.

2. 1. 2. Maithili

Frequently occurring consonants in colloquial Maithili (CM) was investigated by Yadav in 1976. The data was borrowed from dictionary entries by Jha (1952). Consonant /b/ occurred more frequently followed by /k/, /s/, /p/ and /m/. Aspirated phonemes occurred relatively less compared to their unaspirated counterparts in CM. Among resonant [+h] segments, [r^h] occurred more number of times. [l^h] occurred the least (only 19 entries) among all the phonemes.

2. 1. 3. Urdu

Ghazali (2002) determined the frequency of various syllable templates in Urdu. The words selected for frequency count were taken from an Urdu dictionary. CVV (37%) syllable structure was the most common followed by CVC (21.8%), CVVC (16.5%) and CV (16%) structures. The syllables were grouped as light syllables, heavy syllables and super heavy syllables. Light syllables are those which contain only one element in its rhyme while heavy syllables have more than one element in its rhyme. The heavy syllables (CVC, CVV, VC, VV) occurred for 62.4% of the data, super heavy syllables (CVCC, CVVC, CVVCC, VCC, VVC) for 21.4% and light syllables (CV, V) for only 16.2% of the data.

2. 1. 4. Marathi

Marathi portion of the written corpus (EMILLE/CIIL) was used by Berkson and Nelson (2017) to compute the frequently occurring phonemes in this language. The authors determined both token and type frequencies of the same. However, the results suggested that there were only

slight variations between type and token frequencies. On the whole, frequency count is more for plain (unaspirated) sounds (93%) than aspirated sounds. The token frequency for these sound classes is mentioned in table 2.16. Among the consonants, alveo- palatals and dentals were the most followed by labials. Retroflexes and glottal [h] had comparatively the least frequency count.

Table 2.16

Frequency of plain and aspirated obstruents and sonorants in Marathi (Berkson & Nelson, 2017)

	Plain	**Aspirated**	**Total**
Voiceless obstruent	37%	2.8%	40%
Voiced obstruent	10%	3.6%	14%
Sonorants	45%	0.9%	46%

Low central vowel [a] occurs 46% in the data, which was the most frequently occurring vowel. This was followed by [i] (23%) and [e] (16%). It was also inferred that back rounded vowels occurred much less when compared to other vowels.

CV bigram frequencies were also reported by the authors. All the consonants occur relatively more frequently with vowel [a]. It also occurs frequently with vowels [i] and [e] but the frequency of occurrence is much less than that of its occurrence with [a].

2. 1. 5. Bengali

A frequency count of Bengali phonemes was provided by Ferguson and Chowdhury in 1960. The phoneme count was carried out from written as well as spoken material. The first five frequently occurring phonemes are provided in table. The top 10 phonemes in this study were /e, o, a, r, n, k, i, l, b, t/. These results were in rough agreement with a study by Chatterji (as cited in Ferguson & Chowdhury, 1960) wherein the 10 most frequent phonemes were /a, e, o, r, i, ɔ, n, k, b, l/. The frequently occurring consonants /r, n, k, l, b/ were commonly present in stems and in

certain high frequency inflexional suffixes. The most common initial clusters present were stop+ liquid or /s/ + stop. V, CV, VC, CVC, VV, CVV were the typically occurring syllable structures.

2. 1. 6. Gujarati

Frequently occurring phonemes and syllables in Gujarati were reported by Pandit (1965). Phoneme /a/ was ranked highest followed by /a:/, /h/, /n/, /r/, /i/, /e/ and /k/. Retroflexes and vowel /ɔ/ were comparatively less frequent in the language. The syllables with maximum occurrence were /ne/, /va:/, /che/, /mã:/, na:/, /ni:/ and /a:/.

It was reported by Patel (2004) that, in Gujarati language, CV syllables had the highest frequency of occurrence. However, VC type of syllables was less used in the language. Also, VCC syllable structure was rarely found.

2. 1. 7. Punjabi

Singh and Lehal (2010) analyzed a large Punjabi corpus to determine the syllable frequencies (in initial, medial and final positions) for the preparation of Punjabi speech database for TTS system. Since syllables are the basic unit for a syllabic language like Punjabi, the authors suggested that syllables be used for the TTS system. The authors concluded that the occurrences of both nasal and non-nasal syllables are predominant at the initial position than the medial and final position. Furthermore, syllables of CV type (both nasal and non-nasal) occur more in maximum words in the corpora while VCC type syllables (both nasal and non-nasal) occur least in frequency.

2. 1. 8. Oriya

Kelkar (1994) studied phonemic and morphemic frequencies in Oriya. The sources were similar to that of Ghatage's (1994) study. The results indicated that vowel /ə/ was the most occurring followed by /a/ and /I/. /r/, /k/ and /t/ were the most found consonant phonemes.

2. 2. Dravidian languages

2. 2. 1. Kannada

Nair and Ramachandran in 1958 studied the frequently occurring phonemes in Kannada following analysis of written materials representing language in daily use such as newspapers, journals, books etc. for the purpose of generating a telegraphic code. Nearly a fifth of the total phonemes accounted for the vowel /a/ which was the most frequently occurring phoneme whereas nasals /ŋ/ and /ɲ/ (not present at all in the sample considered) and aspirated consonants such as /g^h/, /$ʈ^h$/ etc. were the least frequently occurring phonemes. Vowels /a, i, u/ and consonants /t, d, n, g, r, v, k/ are among the most frequently occurring phonemes. Except for the long vowel /a:/, all other long vowels and aspirated consonants were infrequent when compared to short vowels and unaspirated consonants.

Frequency of phonemes in colloquial and literary style in Kannada was investigated by Nayaka (1967). Sentences which were part of everyday conversations in colloquial style were considered for analysis of colloquial Kannada. Literary style was considered from reading of passages (by the author) from newspapers and literary texts.

Colloquial style used more number of consonants than vowels. On the other hand, literary style used 8% more vowels than consonants. Short vowels had more occurrences than long vowels in both styles. Consonant /n/ occurred more frequently than /m/ and aspirated consonants were less frequent than their unaspirated counterparts. Also, voiced sounds were more common

when compared to corresponding unvoiced sounds. The most commonly occurring phonemes in both styles of Kannada are provided in table 2.17.

Table 2.17

Most frequently occurring phonemes in literary and colloquial style of Kannada (Nayaka, 1967)

Literary style		**Colloquial style**	
Phonemes	**Percentage of occurrence**	**Phonemes**	**Percentage of occurrence**
/a/	20.62	/a/	12.82
/i/	7.80	/a:/	6.83
/n/	6.35	/n/	6.28
/u/	5.55	/t/	5.92
/a:/	5.19	/i/	5.85
/d/	4.88	/r/	5.09
/r/	4.80	/d/	4.70

Ranganatha (1982) explored the frequency of phonemes and syllables from written sources such as periodicals, newspapers, fiction and non- fiction books in Kannada language. Vowel /a/ was observed to be highly occurring phoneme. Phonemes next in rank were /i/>/ u/> /r/> /d/> /a:/> /e/> /n/> /t/> /k/. Aspirated phonemes, consonants /f/ and /g/ and diphthong /au/ occurred least in frequency. Syllables /da/, /ra/, /va/, /du/, /ga/, /ya/, /na/, /a/ were few of the highly occurring syllables in the language.

Jayaram (1985) computed the phoneme and syllable distribution in Kannada using written materials such as newspapers and magazines. Results were similar to study by Nair and Ramachandran (1958). In general, the short vowels had more occurrences compared to their longer counterparts (except /a:/). Similarly, unaspirated consonants were present more frequently than the aspirated counterparts. This is in consonance with Zipf's principle of least effort (1949). Vowel /a/ was the most frequent (19.04%) followed by /i/> /u/> /a:/. Consonant /n/ was the most frequent phoneme (6.99%). Next in rank were /r/ (6.04%) > /d/ (5.54%) > /t/ (3.78%) > /l/ (3.34%) > /v/ (3.15%) > /k/ (2.52%). The most frequent syllable types were CV (72.78%) and

CVC (16.01%) whereas the least frequent were VC, CCVC and other types. Syllables /da/, /ra/, /va/, /a/, /na/ and /ga/ were highly present while syllable /sva/ was least present in Kannada. The results also suggested that various phonemes and syllables had different frequencies in word and sentence initial positions (table 2.18 & table 2.19).

Table 2.18

Frequently occurring phonemes in Kannada (Jayaram, 1985)

Sentence initial position		**Word initial position**	
Sounds	**Frequency (%)**	**Sounds**	**Frequency (%)**
/a/	12.60	/m/	10.64
/i/	9.22	/k/	9.98
/a:/	8.57	/b/	8.06
/n/	7.47	/s/	7.98
/k/	6.39	/h/	7.48

Table 2.19

Top five frequently occurring syllables and their percentage of occurrence in Kannada (Jayaram, 1985)

Syllables	**Frequency (%)**	**Syllables**	**Frequency (%)**
/a/	10.98	/a/	6.98
/a:/	8.92	/a:/	4.07
/i/	7.13	/ma/	3.74
/i:/	5.65	/ka/	3.33
/na/	2.54	/sa/	2.99

A study by Rupela and Manjula (2006) reported syllable shape CV to be prominent in children as young as 3 years. Similar results have been obtained by Priya and Manjula (2016). The most common word shape was CV followed by VC and CVC. Disyllables were prominent in Kannada.

More recently, a study was conducted by Sreedevi, Smitha and Vikas (2012) in Kannada language using conversation samples. Results revealed that the mean frequency of consonants were more than that of vowels. Among the vowels, /a/ had the highest percentage of occurrence of 14.6% while /ə/ had only 0.56%. Consonant /n/ has the highest frequency of occurrence of about 7.87% followed by /r/ (5.43%) and /l/ (5.14%). The least frequently occurring phonemes in conversational Kannada were the aspirated sounds such as /k^h/, /t^h/ etc, fricatives /s/, /ʃ/ and affricates /c/ and /ɟ/. The frequently occurring phonemes in descending order is as follows: /a/, /n/, /I/, /e/ /r/, /a:/, /d/, /l/, /u/, /k/. These sounds account for 70.2% of the total phonemes in the data.

Frequency of occurrence of phonemes was computed by Manjula, Geetha, Sharath and Antony (2015) to develop a phonemically balanced word list for audiological evaluation. The corpus selected included both written and spoken data with a total of 15000 phoneme occurrences. The most frequently occurring phoneme was vowel /a/ followed by vowel /i/. Among consonants, /n/, /r/, /t/ and /l/ were the top frequently occurring phonemes.

2. 2. 2. Malayalam

Ghatage (1994) determined the phonemic frequencies in Malayalam from several written sources. The descending order of frequently occurring phonemes is as follows: /a/> /i/> /u/> /m/> /a:/> /n/> /k/. Aspirated consonants, phoneme /f/ and diphthong /au/ were few of the least occurring sounds in the language. Syllables /a/, /ra/, /va/, /ga/, /vi/, /ru/ and /ku/ were highly occurring in the language.

Frequency of occurrence of phonemes in Calicut, Ernakulam and Thiruvananthapuram dialects of Malayalam was investigated by Sreedevi and Irfana (2013) from conversation samples. In general, the percentage of consonants was higher than vowels. On the whole, the

frequency of occurrence of phonemes in Malayalam was as follows: /a, i, k, ə, a:, ʈ, t/. Vowel /a/ was the most frequently occurring phoneme in all the three dialects.

Percentage of vowels was higher in Thiruvananthapuram dialect followed by Ernakulam and Calicut dialects. Vowels and consonants present in each of the dialects of Malayalam are provided in table 2.20.

Table 2.20

Top frequently occurring vowels and consonants in Malayalam (Sreedevi & Irfana, 2013)

Dialects of Malayalam	Vowels	Consonants
Calicut	/a, i, ə, a:, e/	/k, n, ʈ, t, l/
Ernakulam	/a, i, ə, a:/	/k, ɳ, ʈ, p, t, l, m/
Thiruvananthapuram	/a, i, ə, e, a:/	/k, n, ʈ, t, m/
Most commonly occurring phonemes in the three dialects	/a, i, ə/	/k, n, ʈ/

Considering the place of articulation, velar /k/ had the highest occurrence followed by alveolar nasal /n/ in all the three dialects. Stops had the highest ranking followed by nasals in all three dialects based on manner of articulation. Diphthongs and aspirated consonants had reduced occurrence in Malayalam.

2. 2. 3. Tamil

Tamil phonology was investigated by Vasanthakumari in 1989. The frequency of occurrence of vowels and consonants were almost equal as observed. Maximum occurrence was for vowel /a/ (25%) and vowels /i/ and /u/ have same occurrence percentage of 8%. Based on tongue height, low vowels had highest occurrence and based on tongue advancement, central vowel had the highest rank. Among the consonants, /k/ was ranked first followed by /p/, /m/, /t/, /ʈ/, /l/ and /ɭ/. Phonemes /g/, /d/, /f/, /ɖ/, /n/ and /ɳ/ were the least occurring in Tamil. Labials and alveolars were more in frequency when classified according to place of articulation. Stops were more while trills and flaps were the least in number in the language.

2. 2. 4. Telugu

Kumar, Murthy and Chaudhuri (2007) conducted statistical analysis of Telugu text corpora. Results revealed consonants occurred more frequently in word initial (86%) and medial position (57%) than word final position. Vowels occurred for about 88% of the time in word final position. The most frequent phonemes, bigrams (a pair of consecutive written units) and trigrams (three consecutive written units) in each word position are provided in table 2.21.

Table 2.21

Most frequently occurring vowels, consonants, bigrams and trigrams in Telugu (Kumar, Murthy & Chaudhuri, 2007)

Word positions	Vowels	Consonants	Bigrams	Trigrams
Word initial	a, a:, i, e, u	p, v, k, s, n	pr, vi, ni, sm, ka	nir, pri, pra:, ka:r, tel
Word medial	a, i, u, e:, o:	m, n, r, l, k	a:m, rm, a:n, a:r, ra:	imc, unn, nna:, uku, tun
Word Final	u, i, a:, o:, e:	m, n, l, r, k	ni, lu, nu, ru, lo:	a:ru, ulu, mdi, iki, uku

Neethipriya (2007) conducted a study on phonotactics in Telugu speaking children. Syllable shape CV was predominant in children. Word shapes CV, VC and CVC were commonly present in the data.

Another study by Kalyani and Sunitha (2009) investigated phoneme frequency in Telugu to build a dictation system from a Telugu text corpus. Approximately 48% of the text was vowels while consonants covered 52 %. Among the vowels, 51% were open vowels /a, a:/ and 41 % were half closed front vowels /e, e:/. Half closed back vowels /o, o:/ had least occurrence of 6%. Alveolars (46%) were the most common among consonants, followed by bilabials (20%). Glottal sounds occurred the least in frequency (0.82%).

Kumar and Mohanty (2012) utilized data of frequently occurring phonemes by Rao and Thenarasu (2007) in order to develop Telugu speech audiometry material. The phoneme count was derived from Telugu corpora available in Centre for Applied Linguistics and Translation Studies Language Technology Lab, University of Hyderabad, India. The first 10 frequently occurring consonants were as follows: /n/, /r/, /l/, /k/, /w/, /t̪/, /p/, /m/, /d̪/ and /y/.

III. Comparison across various Indian languages

Similarly, studies have also been conducted in various Indian languages such as Hindi, Kannada, Malayalam, Tamil, Telugu etc. Ramakrishna, Nair, Chiplunkar, Atal and Rajaraman (1957) used materials from books, journals, newspapers etc. to compute the relative frequency of occurrence of phonemes in Malayalam, Telugu, Tamil, Marathi and Hindi in order to develop a common telegraphic code. Most frequently occurring vowels and consonants in these languages is provided in table 2.22 and table 2.23. A striking similarity was observed between Kannada and other Dravidian languages such as Malayalam, Tamil and Telugu (cited in Ramakrishna, Nair, Chiplunkar, Atal, Ramachandran & Subramanan, 1962).

Table 2.22

Relative frequency of occurrence of phonemes in Malayalam, Telugu and Tamil (Ramakrishna, Nair, Chiplunkar, Atal and Rajaraman, 1957)

Malayalam		**Tamil**		**Telugu**	
Vowels	**Percentage**	**Vowels**	**Percentage**	**Vowels**	**Percentage**
/a/	13.88	/a/	15.15	/a/	16.70
/i/	8.01	/i/	8.04	/i/	7.23
/u/	7.03	/u/	7.96	/u/	7.16
Consonants	**Percentage**	**Consonants**	**Percentage**	**Consonants**	**Percentage**
/n/	7.55	/k/	9.19	/n/	6.35
/k/	5.90	/t/	6.92	/r/	4.80
/t/	4.90	/n/	5.18	/k/	4.07

/j/	4.05	/m/	4.59	/m/, /t/	3.43

Table 2.23

Relative frequency of occurrence of phonemes in Hindi and Marathi (Ramakrishna, Nair, Chiplunkar, Atal and Rajaraman, 1957)

Marathi				**Hindi**			
Vowels	**Percent -age**	**Conson- ants**	**Percent -age**	**Vowels**	**Percen -tage**	**Conson- ants**	**Percen -tage**
/a/	19.05	/r/	4.84	/a/	20.37	/k/	5.96
/a:/	13.40	/t/	4.74	/a:/	8.41	/r/	4.76
/e/	4.19	/j/	3.63	/e/	5.76	/h/	4.25
		/n/	3.50			/n/	4.02

Ramaswami (1999) explored the common characteristics in several Indian languages such as Hindi, Malayalam, Marathi, Kannada, Tamil etc. Short vowels were more common than long vowels in most languages. Nasalized vowels were less frequent than oral vowels. Most languages have higher number of front vowels and unrounded vowels. Low central vowels were observed to be very common. Oral vowels [a], [i], [e], [o] and [u] occur in almost all the languages. The most frequently occurring phonemes in most of the languages are listed in table 2.24.

Table 2.24

Most commonly occurring phonemes in different Indian languages in terms of sound classes (Ramaswami, 1999)

Class of sounds	**Phonemes**
Stops (unaspirated)	[p], [t], [ʈ], [k]
Affricates (unaspirated)	[c], [ɟ]
Fricatives	[s], [š], [h]
Nasals	[m], [ɳ], [n], [ŋ]
Lateral	[l]
Trill	[r]
Semi vowels	[w], [j]

Bharathi, Prakash, Rajeev and Bendre (2002) statistically analyzed syllable frequency of CIIL corpora of 10 languages, namely, Hindi, Punjabi, Marathi, Oriya, Assamese, Bengali, Telugu, Tamil, Malayalam and Kannada. Percentage of the five most frequently occurring syllables was extracted. The results for Hindi, Punjabi and Telugu are provided in table 2.25. The data also revealed that total number of syllables is higher for South Indian languages like Malayalam, Tamil etc, than North Indian languages (Hindi, Punjabi). For instance, 7050 syllables are present in Malayalam whereas Punjab has only about 2035 syllables.

Table 2.25

Percentage of occurrence of five most frequently occurring syllables in Hindi, Telugu and Punjabi (Bharathi, Prakash, Rajeev & Bendre, 2002)

Hindi	**Percentage**	**Telugu**	**Percentage**	**Punjabi**	**Percentage**
/ra/	5.27	/na/	2.71	/ra/	5.13
/ka/	3.60	/la/	2.59	/sa/	3.24
/na/	2.84	/ni/	2.34	/ka/	3.06
/sa/	2.80	/ka/	2.20	/na/	2.58
/pa/	2.17	/va/	1.89	/a/*	2.40

*Note: Unisyllable has been considered as a distinct syllable in the corpus

A syllable level analysis was carried out by Prakash, Prakash and Murthy (2016) using continuous speech in six languages namely, Bengali, Hindi, Marathi, Kannada, Tamil and Telugu in order to develop a TTS system. Results suggested that among syllables with two phones, CV form was the most common structure in all languages. This was followed by syllables with three phones CVC. Syllable containing four or five phones were present rarely and occurred due to the presence of loan words from English.

IV. Comparison of frequency of phonemes across various languages of the world

Yegerlehner and Voegelin (1957) determined frequently present phonemes in nine different languages, namely, Maori, Hidatsa, Winnebago, Shawnee, Choctaw, Havasuoai, Navaho, Chontal and Tarascan. Texts in these languages were analysed for the same. Vowel /a/ was first in rank in Maori, Hidatsa, Choctaw, Tarascan and Navaho while /i/ was ranked highest in Havasuoai and Shawnee. Vowel /e/ and /u/ were most common in Winnebago and Chontal respectively. The list of consonants in decreasing order of frequency is given in table 2.26.

Table 2.26

Frequently occurring phonemes in Maori, Hidatsa, Winnebago, Shawnee, Choctaw, Havasuoai, Navaho, Chontal and Tarascan (Yegerlehner and Voegelin, 1957)

Languages	Frequently occurring phonemes
Maori	/t/, /k/, /h/, /r/
Hidatsa	/n/, /h/, /k/, /ʔ/
Winnebago	/ æ /, /g/ /n/, /r/
Shawnee	/w/, /k/, /l/, /t/
Choctaw	/t/, /h/, /l/, k/
Havasuoai	/k/, /á/, /y/, /n/, /l/
Navaho	/ʔ/, /d/, /x/, /i/, /t/
Chontal	/n/, /á/, /h/, / ʔ/, /t/
Tarascan	/n/, /k/, /s/, /r/

Sigurd (1968) reported the frequency of occurrence of phonemes in five languages, namely, Samoan (Bridgeman, 1961), Kaiwa (Bridgeman, 1961), American English (Roberts, 1965), Bengali (Ferguson & Chowdhury, 1960) and Swedish (Fant & Ritcher, 1958). The most frequent phonemes in each language and their percentage of occurrence are given in table 2.27 and 2.28.

Table 2.27

Frequently occurring phonemes in Samoan and Kaiwa (Sigurd, 1968)

Samoan		Kaiwa	
Phonemes	**Frequency (%)**	**Phonemes**	**Frequency (%)**
/a/	25.66	/a/	17.67
/e/	10.10	/e/	11.27
/i/, /l/	8.90	/o/	9.13
/u/	8.62	/ɨ/	7.08
/o/	7.84	/r/	6.50
/ʔ/	6.32	/i/	6.33
/t/	5.88	/b/	5.08

Table 2.28

Frequently occurring phonemes in American English, Bengali and Swedish (Sigurd, 1968)

American English		Bengali		Swedish	
Phonemes	**Frequency (%)**	**Phonemes**	**Frequency (%)**	**Phonemes**	**Frequency (%)**
/ə/	11.82	/e/	12.36	/n/	8.2
/i/	9.29	/o/	10.27	/a/	7.8
/t/	6.95	/a/	8.16	/t/	6.7
/y/	6.77	/r/	7.74	/r/	6.2
/r/	6.58	/n/	6.97	/s/	5.3
/n/	6.29	/k/	6.54	/d/	5.1
/e/	4.74	/i/	5.44	/e/, /ə/	4.9

According to Kiparsky (1979), CV type of syllables were the most common whereas VC type was least present most languages of the world.

Frequently occurring syllable types in English and Italian were reported by Black and Singh (1968). CVC syllable type was the most common in English (33.5%). VC (20.3%) and CV (21.8%) syllables were next in rank. CCVC type was the least frequent syllable structure in the language. CV syllables accounted for 70% of the total syllable types in Italian. Other syllable structures were relatively less common.

A study by Daur (1983) reported the incidence of frequently occurring syllable types in English and Spanish. Closed type syllables highly occurred in English unlike Spanish which has more of open type syllables. 34% and 30% of the syllable types were accounted for by CV and CVC syllables. In Spanish 58% of the total syllables was CV type.

Thomas (2005) as part of his study prepared a corpus from spoken and written material for five languages namely, Cantonese, Mandarin, Italian, German and English. The author obtained a list of frequently occurring phonemes in these languages. The list of commonly present vowels, diphthongs and consonants for the five languages are given in table 2.29, 2.30, 2.31, 2.32 and 2.33.

Table 2.29

Frequently occurring phonemes in Cantonese (Thomas, 2005)

Vowels	**Frequency (%)**	**Diphthongs**	**Frequency (%)**	**Consonants**	**Frequency (%)**
/i/	18.18	/ɐi/	27.39	/k/	15.40
/ɔ/	17.72	/ei/	20.45	/ts/	10.78
/ɐ/	15.39	/ou/	19.32	/j/, /l/	10.60
/a/	13.86	/ɐu/	14.93	/h/	10.25
/ɛ/	12.00	/ai/	5.49	/t/	8.93

Table 2.30

Frequently occurring phonemes in Mandarin (Thomas, 2005)

Vowels	**Frequency (%)**	**Diphthongs**	**Frequency (%)**	**Consonants**	**Frequency (%)**
/i/	29.54	/au/	36.13	/j/	12.99
/a/	23.88	/ai/	27.19	/w/, /t/	9.50
/e/	21.67	/ou/	12.91	/ş/	6.76
/i/	29.54	/au/	36.13	/j/	12.99
/u/	11.50	/ei/	9.80	/tɕ/	6.65

Table 2.31

Frequently occurring phonemes in German (Thomas, 2005)

Vowels	Frequency (%)	Diphthongs	Frequency (%)	Consonants	Frequency (%)
/ɛ/	15.77	/ai/	61.66	/n/	16.74
/ɐ/	13.86	/au/	27.65	/t/	11.28
/a/	11.09	/oy/	10.68	/R/	11.15
/I/	10.54			/d/	8.99
/e:/	9.54			/s/	6.17

Table 2.32

Frequently occurring phonemes in American English (Thomas, 2005)

Vowels	Frequency (%)	Diphthongs	Frequency (%)	Consonants	Frequency (%)
/ə/	22.98	/aI/	34.52	/n/	11.56
/I/	15.09	/eI/	27.46	/t/	11.53
/æ/	10.75	/əʊ/	24.09	/s/	8.28
/i/	9.85	/aʊ/	11.90	/d/	7.94
/a/	7.85			/ɹ/	6.78

Table 2.33

Frequently occurring phonemes in Italian (Thomas, 2005)

Vowels	Frequency (%)	Diphthongs	Frequency (%)
/i/	26.75	/n/	13.89
/ɛ/	26.20	/r/	13.14
/a/	21.86	/t/	13.02
/ɔ/	19.43	/l/	11.02
		/d/	9.19

Additionally, the phonemes were grouped according to their place and manner of articulation and these were compared across languages which yielded interesting results. It was observed that each language concentrated their articulation in certain regions (eg. alveolar) and

had one or two major manner of articulation. Most of the phonemes in Cantonese language are concentrated in the back of the throat (palatal, velar, glottal). This language has more of plosives, laterals and fricatives. Mandarin concentrates most of its articulation in the palatal region with three major places of articulation- alveolar, palatal, alveolar- palatal. Bilabials are also frequent in this language. On the other hand, more than 70% of the sounds have alveolar region as most frequent place of articulation. Most sounds are produced at the front region of the mouth. For instance, the language has a large concentration of nasals, laterals and trills. As German and English have Germanic origin, there are many similarities observed. Fricatives and nasals are almost equal in frequency in both languages and both languages have more sounds concentrated in the front region. However, German has a high concentration of back sound- uvular trills while inter-dental fricatives and approximants (/j/ and /w/) are most common in English.

Shin, Kiaer and Cha (2012) stated the decreasing order of most frequently occurring Korean syllables: /ha/> /li/> /tɕi/ >/ i/> /ki/ > /sɑ/>/ tɕʌk/. This is in consonance with study by Shin (2010). In spontaneous speech, syllable /kɯ/ was ranked first followed by /kɑ/, /i/, /nɑ/, /te/ and /nuIn/. Among the frequently occurring syllables, /tɕi/ and /i/ were present among the top ten frequently occurring phonemes in both speech and dictionary. Few syllables like /kɯ/, /lʌ/ and /ni/ were present more commonly in speech while /si/, /sɑ/, /to/ and /tɕɑ/ were present frequently in dictionary. Syllables /tɛ/, /i/, /lɑ/ and /nɛ/ had similar distributions in both dictionary and speech.

While considering the overall percentage of syllables (irrespective of the source), CVC type syllables account for 60.6% of all syllable types and CV type comprised of 9.9% of the possible syllable types. On the contrary, in dictionary, CV type syllables were the most frequent (42.8%). CVC type accounted for 36.1% of the total occurrences. On the other hand, GV type syllables accounted only for 2.1% of the total syllables.

The phoneme frequency of English (both speech and dictionary) was also reported in the study. In dictionary, vowels /i/, /ə/, /æ/, /e/ and diphthongs /eI/ and /aI/ were most commonly used. Conversely, /ə/, /I/, /e/, and /ʌ/ were most frequently present in speech. Front vowels and unrounded vowels were used more in both in dictionary and speech. The most frequent consonants in dictionary were /t/, /s/, /n/, /l/, /r/, k/ and /d/ but /n/, /t/, /d/, /s/, /l/ and /ð/ were the frequent sounds in speech. Fricatives and stops were more commonly used. Nasals and approximants had equal number of occurrences. Laterals and affricates had the least percentage of occurrences. In terms of place of articulation, alveolar and palate-alveolar sounds had highest occurrence followed by labials and velars. Other places of articulation accounted for only about 35% of the data.

Comparing both the languages, velars were commonly used in Korean than English. However, alveolars were more used in English. Affricates and nasals which were less frequent in English were observed to be more frequent in Korean. Therefore, it can be concluded that frequency of occurrence of phonemes has been widely studied across various languages. It is also noted that the phoneme and syllable frequency varied among languages, dialects and also across written and spoken sources. Hence, the current study aimed at determining the frequency of occurrence of phonemes in spoken Hindi.

CHAPTER III

METHOD

The study aimed to obtain the frequency of occurrence of phonemes and syllables in conversational speech of Hindi from major Hindi speaking regions, such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand.

The objectives of the study were:

1. To obtain a conversational database of Hindi language
2. To calculate the frequency of occurrence of phonemes in Hindi from the database obtained
3. To determine the order of frequency of occurrence of phonemes with respect to each place and manner of articulation and vowel type
4. To obtain the frequency of occurrence of syllable types, word shapes, consonant clusters and phoneme position in the word

Participants: Participants were native speakers of Hindi language in the age range of 20-to-70 years. All had a minimum education of 12th standard and studied in Hindi or English medium. The participants were exposed to Hindi and used the language in daily conversation most of the time. The data was collected from individuals native to major Hindi speaking urban/semi urban regions of India such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. Participants included students, house wives, office goers or retired employees. It was a group discussion where all the participants were involved in natural conversation and had equal opportunities to take turns. There were a total of 18 audio recordings of natural conversation in Hindi. The duration of each recording session was about 20 minutes. A minimum of 4-5 participants were included in each group. Out of the total of 91 participants, 33 were

males, and 58 were females. The participants were assessed informally for any speech, language and hearing disorders. Details of the participants in each recording session are provided in table 3.1.

Table 3.1

Number of participants in each recording session

Recording sessions	Males	Females	Total participants
R1	1	5	6
R2	2	4	6
R3	3	2	5
R4	1	5	6
R5	3	3	6
R6	1	4	5
R7	4	0	4
R8	0	6	6
R9	1	5	6
R10	4	0	4
R11	4	2	6
R12	1	3	4
R13	0	4	4
R14	2	3	5
R15	1	3	4
R16	2	3	5
R17	0	4	4
R18	3	2	5
Total no. of participants	33	58	91

Instrumentation: The conversations were recorded using Olympus (LS 100) digital recorder. Toshiba (Satellite C665) laptop, Philips (Shl3095) headphones were used for transcription. Analysis was carried out with the help of Systematic Analysis of Language Transcripts (SALT- Clinical Demo Version 2012.4.5) software. The program enables researchers in eliciting, transcribing, and analyzing language samples from one or more individuals. It provides clinicians and researchers a platform to transcribe samples of

everyday speech such as conversations and narrations into a common format and generate reports containing various measures such as syntax, semantics, discourse, fluency, error cateegories, frequency of words, morphemes etc.

Procedure: Initially, the participant details were collected, such as age, education and general medical history. The participants were asked to sit in a circle. The digital recorder was placed in the center, equidistant from each of the participant. The participants were encouraged to speak freely and naturally on any subject of their common interest for about 20 minutes. No specific topics were provided to prevent the high occurrence of certain phonemes. The participants were encouraged to speak as naturally as possible only in Hindi and avoid words from other languages. However, they were allowed to use frequently used loan words from English.

Loan words were considered as in the present day conversations of urban and semi-urban population, they are highly prevalent due to improved education levels and exposure to English. For instance, loan words such as *holiday*, *friends* etc are used more commonly than their native equivalents. Using conversational data is more appropriate to arrive at the frequency of occurrence of phonemes in more natural settings as reported in the literature. This is important as SLPs need to choose more frequent phonemes of the language for speech sound correction as it would enhance speech intelligibility in a short span of time. It is also recommended for development of articulation test and drill material. Also Audiologists require the frequent phonemes in the conversational samples for preparing several test materials to assess individuals with hearing impairment. Eg PB word List. The outcome of the current study is intended for use by SLPs and Audiologists.

It was a group discussion where all the participants were involved in natural conversation and had equal opportunities to take turns. These words are provided in appendix B. Topic of conversation in each recording session is given in table 3.2.

Table 3.2

Topic of conversation in each recording session

Recording sessions	Topic of conversation
Recording 1	General talk about home, hobbies, daily activities
Recording 2	General conversation about self, daily routine, food etc.
Recording 3	Conversation about home and surroundings
Recording 4	Daily activities, discussion about classes, programs in college etc.
Recording 5	Daily activities, classroom routine, class discussions, places to visit etc.
Recording 6	Talk about self, home, daily routines
Recording 7	About self, daily activities, rituals, politics, about work
Recording 8	About school, college, exams, classes
Recording 9	About self, common diseases, common events
Recording 10	About places to visit, history of famous temples
Recording 11	About classes, college, examinations, a social event
Recording 12	Education system, politics
Recording 13	About self, general conversation
Recording 14	About travel, places visited, experience in new college
Recording 15	About native place, general conversation about schools, weather, children
Recording 16	General conversation
Recording 17	Conversation on places to visit, festivals
Recording 18	General conversation on self and daily routine, schooling of children

Using Conversational data is more appropriate to arrive at the frequency of occurrence of phonemes in more natural settings as reported in the literature. This is important as SLPs need to choose more frequent phonemes of the language for speech sound correction as it would enhance speech intelligibility in a short span of time. It is also recommended for development of articulation test and drill material. Also Audiologists require the frequent

phonemes in the language for preparing several test materials to assess individuals with hearing impairment. Eg PB word List. The study outcome is intended for use by SLPs and Audiologists.

Data Analysis: International Phonetic Alphabet (IPA) transcription put forth by Ohala (1994) for Hindi was used to transcribe the conversation samples (Appendix A). Exclamatory remarks and repetitive words were excluded from the sample before analysis. However, commonly used English loan words were considered for analysis.

SALT clinical demo version 2012.4.5 (Miller & Iglesias, 2008) software was utilized to analyze the raw data to determine the frequency count. The software can also be used to compare an individual's language sample with a reference database of language measures. For the present study, a database of Hindi phonemes was prepared, and it consisted of all the phonemes (Ohala, 1994), consonant clusters (Ohala, 1983; Kachru, 2006), syllable types (Shailaja, Manjula & Praveen, 2011) and word shapes (Shailaja, Manjula & Praveen, 2011) present in Hindi language.

The conversation samples were analyzed using the following steps:

1. The conversation sample was transcribed using the database of Hindi phonemes (Appendix A), consonant clusters, syllable types and word shapes were prepared initially (Appendix C).
2. A set of codes were prepared for each of the units (phonemes, consonant clusters, syllable types and word shapes) in the database. Few examples for the same are provided in table 3.3. These codes were utilized to determine the most frequently

occurring phonemes among the total phonemes in Hindi, in various phoneme positions (initial, medial, final) and most commonly present cluster groups, syllable types and word shapes. The complete set of codes is given in appendix C.

Table 3.3

Examples for codes for phonemes, consonant clusters, syllable types and word shapes

Phonemes/ consonant clusters/ syllable types/ word shapes from database	**Codes**
/k/	K
/k^h/	Kh
/ʈ/	T
/ʈ̪/	Th
/m/ (initial position)	m: I
/m/ (medial position)	m: M
/m/ (final position)	m: F
/pr/	Pr
/ll/	Ll
CV	CV
VCCV	VCCV
Monosyllables	CV, CVC, VC, V

3. The transcribed data were coded accordingly.

C 3: p:I uu DD:M i au r:F ch:I aa y:M a CVVCV VC CVVCV.

C 4: p:I uu r:M i CVVCV.

C 2: p:I uu DD:M i au r:F ch:I aa y:F CVVCV VC CVVC.

C 5: s:I a b:M j:M i d:I e n:M i ch:I a h:M i y:M e thh:I a n:I a y:I aa r:F CVCCV CVCV CVCVCV CV CV CVVC.

C 4: s:I a b:M j:M i bh:I i n:I a h:M im d:I e th:M e CVCCV CV CVCV CVCV.

C 1: au r:F uu p:M a r:F s:I e m:I u jhh:M e s:I a m:M a jhh:F m:I em n:I a h:M i aa th:M a k:I i p:I uu DD:M i k:I e s:I aa th:F s:I a b:M j:M i l:I a g:M n:M i ch:I a h:M I y:M e n:I a VC VVCVC CV CVCV CVCVC CV CVCV VVCV CV CVVCV CV CVVC CVCCV CVCCV CVCVCV CV.

C 2: v:I o a ch:M a r:F d:I e th:M e h:I e CV VCVC CVCV CV.

C 1: au r:F z:I a r:M uu r:M i dh:I o DD:M i h:I e y:I aa r:F m:I a th:M l:M a b:F p:I uu DD:M i ch:I aa y:F y:I a p:I uu DD:M i au r:F ch:I aa y:F h:I a r:F k:I o i kh:I a y:M e au r:F ei s:M a th:I o k:I o i n:I a h:M im k:I i v:I a h:M am l:I o g:F f:I r:M ii m:I em r:I a h:F r:I a h:M a h:I e y:I aa p:I ei n:I a h:M im k:I a r:F r:I a h:M a VC CVCVCV CVCV CV CVC CVCCVC CVCV CVC CV CVCV VC CVC CV CVV CVCV VC VCV CV CVV CVCV CV CVCV CVC CCV CV CVC CVCV CV CV CV CVCV CVC CVCV.

Figure 3.1. Sample of transcribed and coded file

In figure 3.1, ‘C 1’, ‘C 2’, ‘C 3’, ‘C 4’, ‘C 5’ represents individuals involved in the recording session and ‘p:I’, ‘DD:M’, ‘y:F’ stands for ‘p’ in initial postion, ‘DD’ in medial position and ‘y’ in final position of the word ‘puuDDi’. Codes CV, VVCV. CVVCV etc represents the various syllable and word shapes.

4. The coded files were then loaded to SALT software.

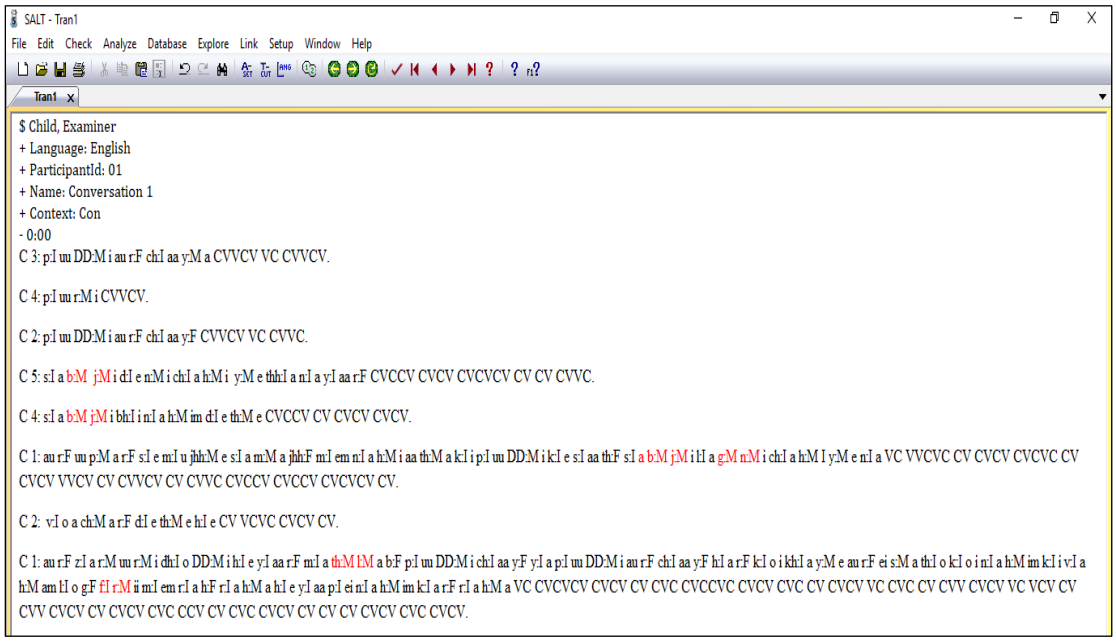

Figure 3.2. Sample of transcribed file loaded in SALT software

5. The set of codes prepared earlier (step 2) were incorporated into an editable standard wordlist in the software. The codes were entered separately to determine the following:
    a. Frequently occurring phonemes from the total phonemes in Hindi
    b. Frequently occurring vowels and consonants
    c. Frequently occurring phonemes in various phoneme position
    d. Frequently occurring consonant clusters
    e. Frequently occurring syllable types

f. Frequently occurring word shapes

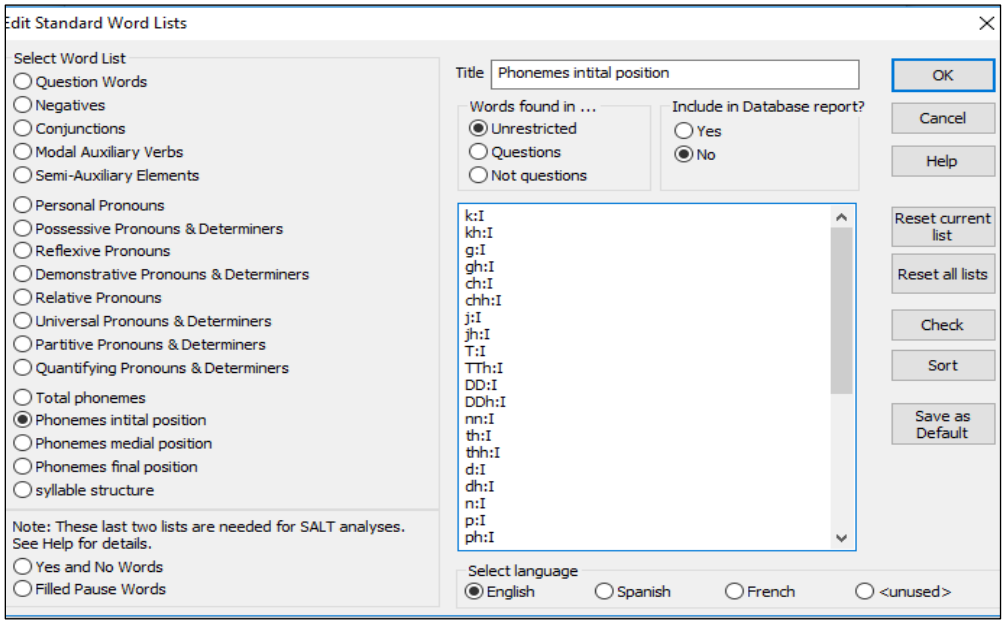

Figure 3.3. Sample for codes for initial phoneme position in the editable standard word list

6. The software then compared the data with the set of codes and provided information on the frequency count of phonemes, cluster groups, and syllable types and word shapes.

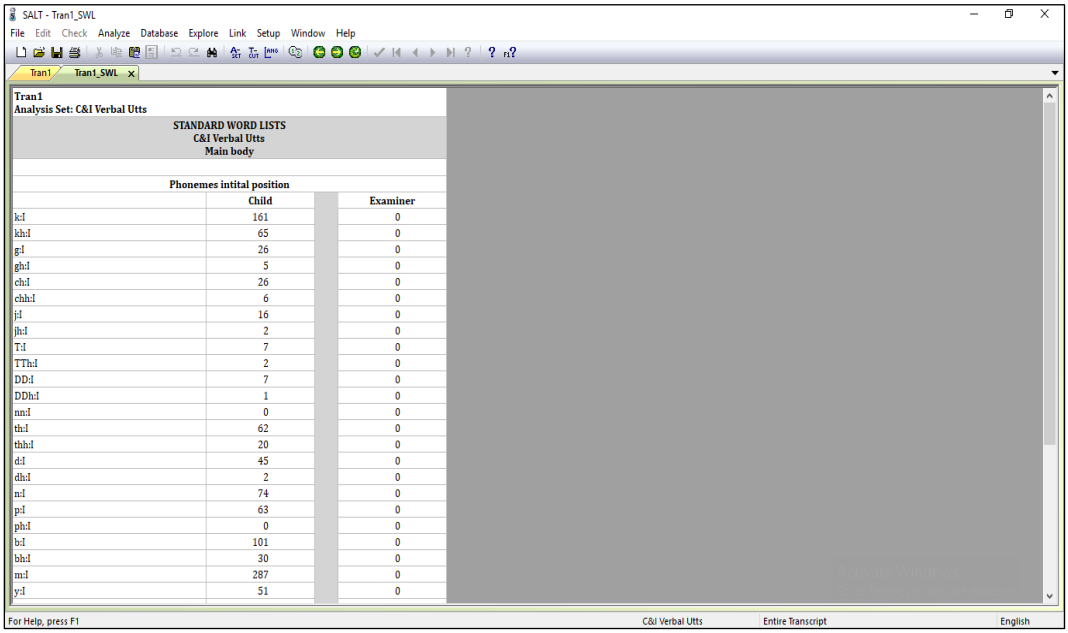

Figure 3.4. Sample output for initial phoneme position from SALT analysis

7. The information of frequency count of phonemes were compiled in an excel sheet for further analysis of frequently occurring consonants with reference to their place and manner of articulation and voicing features.

Inter-judge reliability: Three post-graduate speech-language pathologists including the researcher, trained in the transcription of speech samples served as judges. The phonetic codes to be used and the procedure of transcription were explained to the judges. 10% sample of each recording was subjected to inter-judge reliability measures. The samples were played to the judges individually. The transcribed samples were statistically analyzed for frequently occurring phonemes. The Cronbach alpha reliability index was 0.87 which indicated a good consistency among the three judges.

Intra-judge reliability: 10% of each recording sample was randomly selected and transcribed and reanalyzed by the researcher. Cronbach alpha index obtained was 0.90 indicating a good inter judge reliability.

Statistical analysis

The data was subjected to statistical analysis. Descriptive statistics was employed to determine the mean percentage of occurrence of various phonemes, phoneme positions, consonant clusters, place and manner of articulation, syllable types and word shapes. Friedman test was carried out to determine statistical significance across various parameters. Subsequently, Wilcoxon signed rank test was performed to determine pair wise significance across the parameters under consideration.

CHAPTER IV

RESULTS AND DISCUSSION

The study aimed at obtaining the frequency of occurrence of phonemes, consonant clusters and syllable types in conversational speech of Hindi from major Hindi speaking regions, such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. The objectives of the study were as follows:

I. To obtain a conversational database of Hindi language
II. To calculate the frequency of occurrence of phonemes in Hindi from the database
III. To determine the order of frequency of occurrence of phonemes with respect to each place and manner of articulation and vowel type
IV. To obtain the frequency of occurrence of syllable types, word shapes, consonant clusters and phoneme position in the word

The current investigation involved a total of 91 participants in the age range of 20-to-70 years who were native speakers of Hindi language. Data was collected from various states of major Hindi speaking belt, namely, Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. A total of 18 audio recordings of natural conversation in Hindi were considered. Each recording session had minimum of 4-5 participants. The recorded samples were transcribed using International Phonetic Alphabet (IPA) given by Ohala (1994) and the data was analyzed using SALT software (version 2012.4.5). The results are discussed under the following sub sections.

1. Frequently occurring phonemes in conversational Hindi
2. Frequently occurring vowels and consonants in Hindi
3. Frequently occurring phonemes in initial, medial and final word positions in Hindi
4. Frequently occurring consonant clusters in Hindi

5. Frequently occurring syllable types in Hindi
6. Frequently occurring word shapes in Hindi

1. Frequently occurring phonemes in conversational Hindi

Figure 4.1 depicts the total number of phonemes recorded in each of the recording sessions. The number of phonemes recorded varied from 6,000 to 12,000 phonemes. The first recording session (R1) elicited the maximum number of phonemes of 12,216 and the least number of phonemes of 6,128 were recorded in the sixth session (R6). A grand total of 1, 48,862 phonemes were present from 18 conversational recordings.

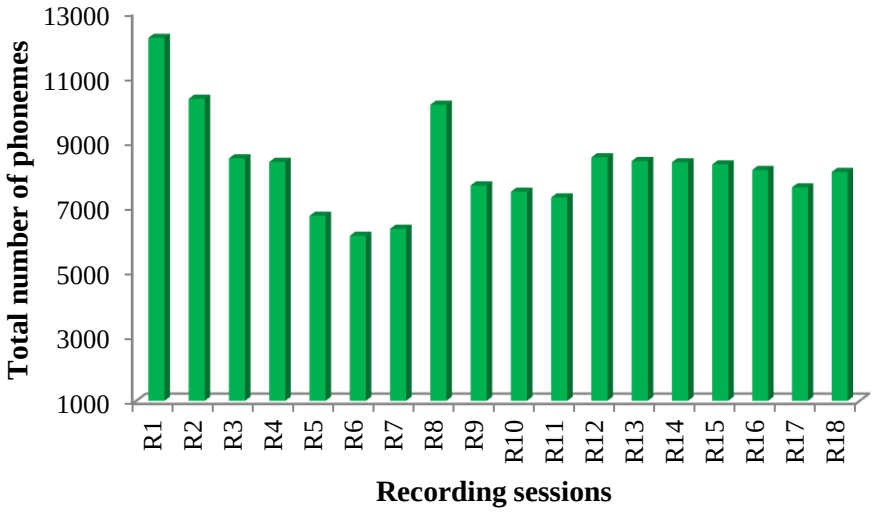

Figure 4.1. Total number of phonemes present in each recording session

Total corpus (1, 48,862 phonemes) of 18 recording sessions included consonants, vowels and diphthongs. Based on descriptive statistics, the mean percentage of occurrence of consonants was higher than vowels and diphthongs. Mean occurrence of consonants and vowels were 54.42% and 44.50% respectively (Fig. 4.2). Diphthongs had least occurrence of 1.08%. The results of the present study is in consonance with earlier studies by Denes (1959), Delattre (1965) in spoken English, Thomas (2005) in American English, spoken Cantonese, Mandarin and Italian

and Renwick (2011) in written English. Similar results have been obtained in several Indian languages such as written Hindi (Ghatage, 1964), written and spoken Gujarati (Pandit, 1965), Kannada (Ranganatha, 1982; Jayaram, 1985; Sreedevi et al, 2012), Tamil (Vasanthakumari, 1989), Telugu (Kumar & Mohanty, 2012) and spoken Malayalam (Sreedevi & Irfana, 2013).

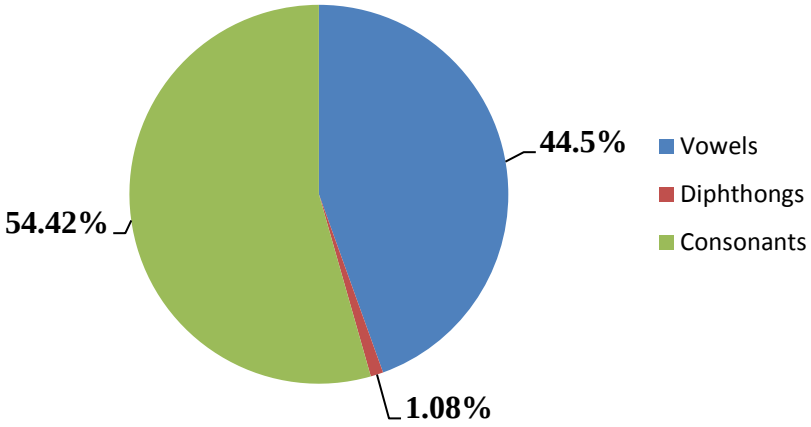

Figure 4.2. Mean occurrence of vowels, diphthongs and consonants

The mean percentage of occurrence of the phonemes in descending order is provided in table 4.1. Vowel /i/ had the highest occurrence of 19.43% in conversational Hindi. Phonemes /n/ (15.99%), /a/ (9.17%), /e/ (6.12%), /f/ (5.22%), /h/ (3.99%) and /k/ (3.80%) were the next frequent phonemes observed in the data. However, the results are contradictory to the earlier studies in written Hindi. According to Chourasia Samudravijaya, Ingle and Chandwani (2007), phonemes /a/> /A/> /e/> /I/> /r/> /k/> /y/ frequently occurred in Hindi. Similarly, phonemes /a:/, /a/, /e/, /i:/, /i/, /b/, /s/, /p/, /m/ and /k/ were few of the highly occurring phonemes in a study by De (1973). Aspirated sounds such as /ʈʰ/, /ʧʰ/, /gʰ/, /bʰ/, /ɟʰ/ and /ɖʰ/ had the least frequency of occurrence in Hindi. This is in agreement with studies in all Indian languages. The aspiration feature is much less used in colloquial language.

Table 4.1

Mean percentage of occurrence of phonemes in spoken Hindi in descending order

Phonemes	Mean %	Phonemes	Mean %	Phonemes	Mean %
/i/	19.43	/u/	1.14	/ũ/	0.17
/n/	15.99	/b/	1.11	/z/	0.16
/a/	9.17	/ai/	0.90	/æ/	0.15
/e/	6.12	/ɻ/	0.87	/t/	0.13
/f/	5.22	/v/	0.87	/ã:/	0.13
/h/	3.99	/ʈ/	0.64	/c^h/	0.12
/k/	3.80	/ã/	0.64	/$ʈ^h$/	0.08
/m/	3.11	/ɖ/	0.56	/ɳ/	0.06
/r/	3.02	/p^h/	0.44	/ŋ/	0.04
/o/	2.94	/k^h/	0.36	/$ʈ͡^h$/	0.04
/s/	2.33	/i:/	0.31	/g^h/	0.04
/p/	2.29	/ə/	0.30	/b^h/	0.03
/a:/	2.11	/ũ:/	0.30	/õ/	0.03
/l/	2.01	/d^h/	0.27	/ĩ:/	0.03
/d/	1.71	/u:/	0.22	/$ɻ^h$/	0.02
/j/	1.52	/i:/	0.22	/ʂ/	0.02
/ʈ͡/	1.37	/au/	0.18	/w/	0.01
/g/	1.23	/ẽ/	0.18	/$ɖ^h$/	0.01
/c/	1.16	/ʃ/	0.18		

2. Frequently occurring vowels, diphthongs and consonants in Hindi

2.1. Vowels and diphthongs

Hindi has nasal and non- nasal vowels. Non- nasal vowels (96.22%) occur more frequently than nasal vowels (3.78%). The same is depicted in figure 4.3. Higher occurrence of non- nasal vowels than nasal vowels have been reported by Ramaswami (1999). In the present study, 90.31% of the total vowels were short vowels while only 5.91% constituted the long vowels (Fig. 4.3). Similar observations have been observed in various Indian languages (Sreedevi, Smitha & Vikas, 2012; Sreedevi & Irfana, 2013).

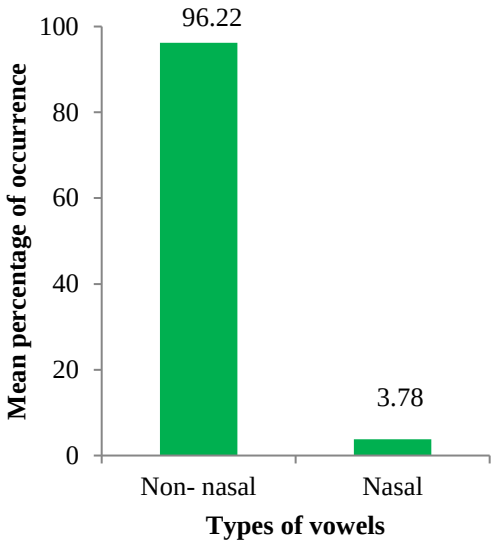

Figure 4.3. Mean percentage of occurrence of nasal and non-nasal vowels

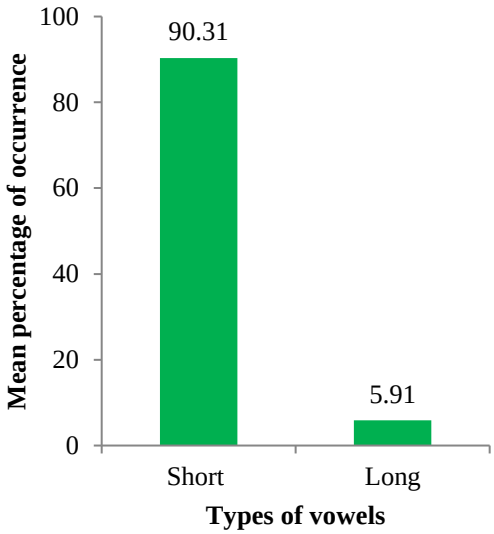

Figure 4.4. Mean percentage of occurrence of short and long vowels

In general, front- high vowel /i/ was predominant with a mean of 46.11% followed by vowels /a/ (20.46%), /e/ (13.64%) and /o/ (6.56%). However, according to reports by Ghatage and Madhav (1964), Chourasia, Samudravijaya, Ingle and Chandwani (2007), De (1973), Khan (1990) and Malviya, Mishra and Tiwary (2016), vowel /a/ was the most predominant vowel in Hindi. Similarly, other Indian languages such as Marathi (Berkson & Nelson, 2017), Gujarati (Pandit, 1965), Kannada (Nayaka, 1967; Ranganatha, 1982; Jayaram, 1985; Sreedevi, Smitha & Vikas, 2012), Malayalam (Ghatage, 1994; Sreedevi & Irfana, 2013), Tamil (Vasanthakumari, 1989) and Telugu (Kumar, Murthy & Chaudhuri, 2007; Kalyani & Suitha, 2009) had vowel /a/ as the highly occurring vowel. However, vowel /i/ had a high occurrence in English (Dewey, 1923; Voelker, 1935; Tobias, 1959).

In the present study, among the nasal vowels, vowel /ã/ (1.42%) had relatively higher occurrence. It is partially in consonance with a study by De (1973) which reported vowels /õ/ and /ã/ as the most frequent nasal vowels in Hindi. Considering the diphthongs, /ai/ had the highest mean percentage of 82.98%. According to the present study, diphthongs are highly reduced in conversational Hindi. Similar observations have been made by Sreedevi and Irfana (2013) in Malayalam, Ranganatha (1982) in Kannada and Renwick (2011) in Romanian. Table 4.2 depicts the mean percentage of occurrence of vowels and diphthongs in spoken Hindi.

Table 4.2

Mean percentage of occurrence of vowels and diphthongs in Hindi

Vowels	**Mean %**	**Vowels**	**Mean %**	**Diphthongs**	**Mean %**
/a/	20.46	/ã/	1.42	/ai/	82.98
/a:/	4.72	/ã:/	0.28	/au/	17.02
/i/	46.11	/ĩ/	0.48		
/i:/	0.70	/ĩ:/	0.07		
/u/	2.53	/ẽ/	0.67		
/u:/	0.50	/ũ/	0.41		
/e/	13.64	/ũ:/	0.37		
/o/	6.56	/õ/	0.07		
/æ/	0.33				
/ə/	0.68				

High vowels (/i/, /i:/, /u/ and /u:/) occurred more frequently with a percentage of occurrence of 51.8% compared to mid (/e/, /o/, /æ/) and low (/a/, /a:/) vowels (Fig. 4.5). Front vowels (/i/, /i:/, /e/, /æ/) had the highest occurrence (63.16%) as observed from the data. The central (/a/, /a:/, /ə/) and back (/u/, /u:/, /o/) vowels had 26.86% and 9.97% of occurrence respectively. Front vowels were also highly present in English (Denes, 1959; Thomas, 2005), Cantonese, Mandarin, German, Italian (2005) and Telugu (Kalyani & Sunitha, 2009). Figures 4.5 and 4.6 illustrate the mean percentage of occurrence of various vowel categories. On application of Friedman test, a significant occurrence was noted across various vowel categories (table 4.3). Hence, a pair wise

comparison of high, mid and low vowels and front, central and back vowels were carried out using Wilcoxon signed rank test. Analysis of table 4.4 and 4.5 revealed a significant difference across each of the pairs.

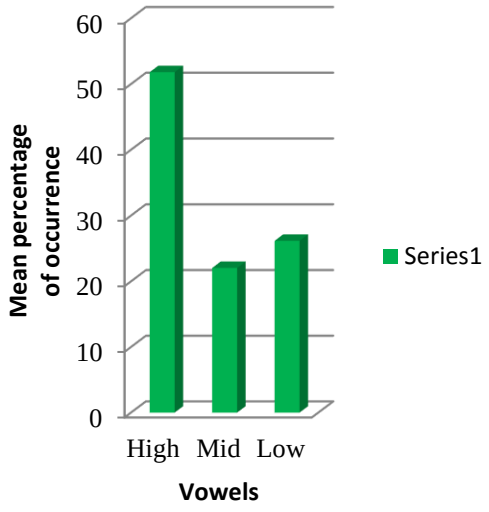

Figure 4.5. Mean percentage of occurrence of high, mid and low vowels

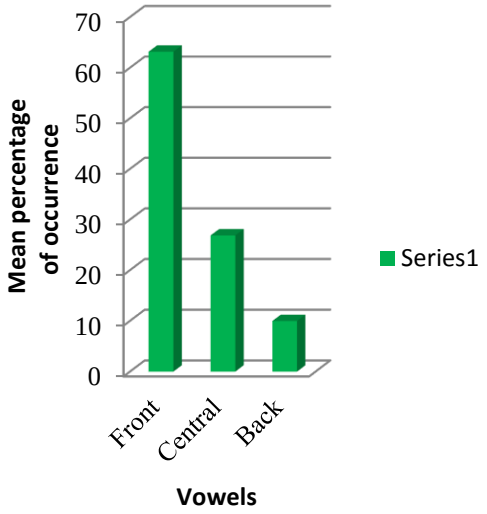

Figure 4.6. Mean percentage of occurrence front, central and back vowels

Table 4.3

Comparison of mean scores across various types of vowels using Friedman test

Types of vowels	χ^2 (4)	p value*
High vowels Mid vowels Low vowels	34.11	0.00
Front vowels Central vowels Back vowels	36.00	0.00

Note. *p<0.05

Table 4.4

Pair wise comparison of mean percentage occurrence across high, mid and low vowels using Wilcoxon signed rank test

Types of vowels	|Z|	p value
Mid- high	3.724	0.000*
Low- high	3.724	0.000*
Low- mid	3.680	0.000*

Note. *p<0.05

Table 4.5

Pair wise comparison of mean percentage occurrence across front, central and back vowels using Wilcoxon signed rank test

Types of vowels	|Z|	p value
Front- central	3.724	0.000*
Back- front	3.724	0.000*
Back- central	3.724	0.000*

Note. *p<0.05

2.2. Consonants

The first five frequent consonants in Hindi in descending order were as follows: /n/ (29.14%) > /f/ (9.52%) > /h/ (7.27%) > /k/ (6.93%) > /m/ (5.66%) > /r/ (5.50%). In general aspirated consonants had the least occurrence. Table 4.6 shows the mean percentage of occurrence of consonants in Hindi. Kannada also had phoneme /n/ as the most frequently occurring consonant in the language (Nayaka, 1967; Sreedevi, Smitha and Vikas, 2012). It was also the second most frequent phoneme in Malayalam (Sreedevi & Irfana, 2013). Among the non- Indian languages, English (Voelker, 1935; Mader, 1954; Crystal, 1981; Thomas, 2005), Swedish (Sigurd, 1968), German (Thomas, 2005) and Italian (Thomas, 2005) had a high occurrence of nasal phoneme /n/. However, several other studies in Hindi reported consonant /k/ to be the most frequently occurring phoneme (Khan, 1990; Malviya, Mishra & Tiwary, 2016).

Table 4.6

Mean percentage of occurrence of consonants in Hindi

Consonants	Mean%	Consonants	Mean%
/k/	6.93	/ʈ/	2.50
/$k^{h/}$	0.65	/$ʈ^{h}$/	0.07
/g/	2.24	/d/	3.12
/g^{h}/	0.06	/d^{h}/	0.51
/c/	2.10	/n/	29.14
/c^{h}/	0.22	/j/	2.77
/ɟ/	1.59	/r/	5.50
/$ɟ^{h}$/	0.04	/l/	3.66
/ʈ/	1.16	/v/	1.59
/$ʈ^{h}$/	0.14	/ʃ/	0.33
/ɖ/	1.03	/ʂ/	0.04
/$ɖ^{h}$/	0.02	/s/	4.25
/ɳ/	0.11	/h/	7.27
/p/	4.18	/t/	0.24
/p^{h}/	0.81	/z/	0.29
/b/	2.03	/w/	0.03
/b^{h}/	0.06	/f/	9.52
/m/	5.66	/ŋ/	0.08

Study by Ghatage (1964) reported phoneme /k/ as the most frequently occurring consonant in written Hindi which is in agreement with the present finding. Considering the place of articulation, dentals (16.32%) had a higher occurrence followed by velars (12.59%) and labials (9.65%). Phonemes /k, h, s, m, p, n, ɟ, b, d, w/ and /r, n, k, ʈ, s, j, h, l, m/ had the highest occurrence in initial and final positions. Velars had higher mean percentage of occurrence in the initial position while stops were predominant in final position. The study also concluded that the commonly occurring syllable types in Hindi were CV followed by CVC.

2.2.1. Manner of articulation

Nasals (34.67%) were more predominant followed by stops (25.3%) and fricatives (24.25%). Among nasals, /n/ had the highest occurrence followed by /m/ (5.66%) while /ŋ/ (0.08%) had least occurrence. Voiceless unaspirated stops had an occurrence of 15.02%. Voiced unaspirated stops had relatively less occurrence of 8.42%. Similar results have been obtained in other languages such as Malayalam, Kannada, Tamil, Telugu (Ramakrshna, Nair, Chiplunkar, Atal and Rajaraman, 1957) and Cantonese and Mandarin (Thomas, 2005). Consonant /f/ had highest mean percentage (9.52%) followed by consonant /h/ (7.27%), among the fricatives. Consonant /ş/ was least present among the fricatives in the corpus. Considering approximants, liquid /l/ occurred with a mean of 3.66%. Hindi has only a single trill /r/, which was present for 5.50% in the corpus. Affricates had the least occurrence of 3.93%. On the contrary, stops had higher occurrence in many Indian languages (Jayaram, 1985; Kalyani & Sunitha, 2009) and non-Indian languages (Denes, 1957; Guirao & Jurado, 1990; Thomas, 2005).

Figure 4.7 depicts the mean percentage of occurrence of consonants based on manner of articulation. The results of Friedman test revealed a significant occurrence of various manner of articulation (table 4.7). Further, Wilcoxon signed rank test was employed to establish pair wise comparison of various categories stops, nasals, fricatives affricates, approximants and trills (table 4.8). All the pairs were observed to have a significant difference except fricatives and stops. The mean occurrence of fricatives (24.25%) and stops (25.3%) were nearly similar.

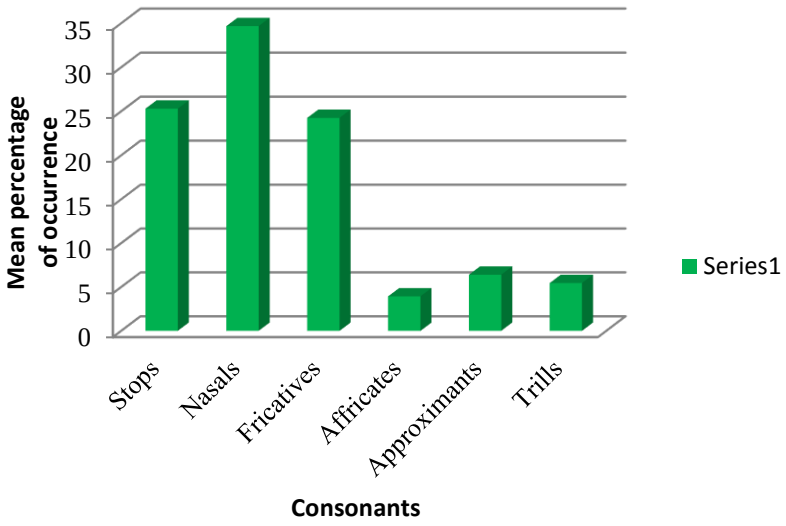

Figure 4.7. Mean percentage of occurrence of consonants based on manner of articulation

As data did not follow assumptions of normality, hence Nonparametric statistics are applied.

Table 4.7

Comparison of mean percentage of occurrence across various manners and place of articulation using Friedman test

Manner and Place	χ^2 (4)	p value
Place of articulation	118.56	0.000*
Manner of articulation	84.75	0.000*

Note. *p<0.05

Table 4.8

Pair wise comparison of mean percentage occurrence across stops, nasals, fricatives affricates, approximants and trills using Wilcoxon signed rank test

Manner of articulation	|Z|	p value
Nasals- stops	3.724	0.000*
Fricatives- stops	1.786	0.074
Affricates- stops	3.724	0.000*
Approximants- stops	3.724	0.000*
Trills- stops	3.724	0.000*
Fricatives- nasals	3.724	0.000*

Affricates- nasals	3.724	0.000*
Approximants- nasals	3.724	0.000*
Trills- nasals	3.724	0.000*
Affricates- fricatives	3.724	0.000*
Approximants- fricatives	3.724	0.000*
Trills- fricatives	3.724	0.000*
Approximants- affricates	3.724	0.000*
Trills- affricates	3.246	0.001*
Trills- approximants	3.243	0.001*

Note. *p<0.05

2.2.2. Place of articulation

In general, alveolars had the highest percentage of occurrence (42.79%) while retroflexes occurred in least frequency (2.5%). More than 50% of the alveolars were accounted for by consonant /n/. In general, bilabials and labiodentals had almost equal percentage of occurrence. Voiced dental /d/, voiceless bilabial /p/ and voiceless fricative /f/ were most frequent among dentals, bilabials and labiodentals respectively. Among the velars, /k/ had the highest occurrence. It is also among the top five phonemes. Unlike Malayalam (Sreedevi & Irfana, 2013), Telugu (Kalyani & Sunitha, 2009) and Marathi (Berkson & Nelson, 2017), Hindi had relatively higher occurrence of glottal sound /h/. Palatal approximant /j/ and voiceless stop /ʈ/ had highest mean percentage in the palatals and retroflexes respectively. Similar to Hindi, Telugu (Kalyani & Sunitha, 2009; Kumar & Mahanty, 2012), Cantonese, Mandarin, Italian, German and American English (Thomas, 2005) had higher occurrence of alveolars. Both Kannada (Sreedevi, Smitha & Vikas, 2012) and Malayalam (Sreedevi & Irfana, 2013) had higher occurrences of dentals. However, the present study was contradictory to the study by Khan (1990). The later study reported dentals (16.32%) to be the most frequently occurring phonemes followed by velars (12.59%) and labials (9.65%).

Figure 4.8 depicts the mean percentage of occurrence of consonants based on place of articulation. Statistical significance was determined using Friedman test (table 4.7). As there was significance observed across bilabials, labiodentals, dentals, alveolars, palatals, retroflexes, velars and glottal categories, Wilcoxon signed rank test was applied to establish pair wise significance. The pairs palatals- dentals, bilabials- labiodentals and glottal- palatals did not have a significant difference which indicates these categories had a similar percentage of occurrence in Hindi.

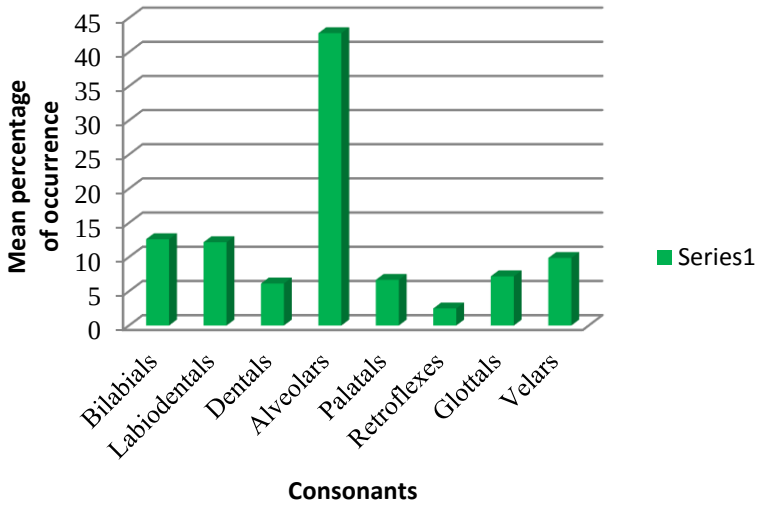

Figure 4.8. Mean percentage of occurrence of consonants based on place of articulation

Table 4.9

Pair wise comparison of mean percentage occurrence across bilabials, labiodentals, dentals, alveolars, palatals, retroflexes, velars and glottal using Wilcoxon signed rank test

Place of articulation	**|Z|**	**p value**
Alveolars- dentals	3.724	0.000*
Labiodentals- dentals	3.724	0.000*
Retroflexes- dentals	3.724	0.000*
Velars- dentals	3.724	0.000*
Palatals- dentals	2.345	0.019
Bilabials- dentals	3.724	0.000*

Glottal- dentals	3.333	0.001*
Labiodentals- alveolars	3.724	0.000*
Retroflexes- alveolars	3.724	0.000*
Velars- alveolars	3.724	0.000*
Palatals- alveolars	3.724	0.000*
Bilabials- alveolars	3.724	0.000*
Glottal- alveolars	3.724	0.000*
Retroflexes- labiodentals	3.724	0.000*
Velars- labiodentals	3.636	0.000*
Palatals- labiodentals	3.724	0.000*
Bilabials- labiodentals	1.024	0.306
Velars- retroflexes	3.724	0.000*
Palatals- retroflexes	3.724	0.000*
Bilabials- retroflexes	3.724	0.000*
Glottal- retroflexes	3.724	0.000*
Palatals- velars	3.724	0.000*
Bilabials- velars	3.574	0.000*
Glottal- velars	3.686	0.000*
Bilabials- palatals	3.724	0.000*
Glottal- palatals	1.847	0.065
Glottal- bilabials	3.274	0.000*

Note. *p<0.05

3. Frequently occurring phonemes in initial, medial and final word positions in Hindi

Consonants /n/ (35.41%), /h/ (9.52%), /f/ (8.88%), /k/ (8.24%), /d/ (3.56%) and /s/ (3.52%) were the most frequently present in initial word position (table 4.10) whereas /f/ (12.29%), /n/ (12.09%), /m/ (8.83%), /r/ (7.52%), /h/ (6.87%) and /k/ (6.35%) commonly occurred in final position (table 4.11). Consonant /n/ occurred with highest percentage frequency of 28.57% in the final position as well (table 4.12). This was followed by consonants /r/ (17.50%), /m/ (7.35%), /d/ (5.67%), /s/ (4.88%), /l/ (4.81%), /k/ (4.49%) and /p/ (4.49%).

Study by Khan (1990) in written Hindi reported phonemes /k, h, s, m, p, n, ɟ, b, d, w/ and /r, n, k, t̪, s, j, h, l, m/ as the highest occurring phonemes in initial and final positions. Velars had higher occurrence in the initial position while stops were predominant in final position.

Table 4.10

Mean percentage of occurrence of consonants in initial word position in Hindi

Consonants	Mean%	Consonants	Mean%
/k/	8.24	/ʈ/	3.24
/$k^{h/}$	0.79	/$ʈ^{h}$/	0.13
/g/	1.03	/d/	3.56
/g^{h}/	0.11	/d^{h}/	0.82
/c/	1.68	/n/	35.41
/c^{h}/	0.28	/j/	1.16
/ɟ/	1.86	/r/	1.97
/$ɟ^{h}$/	0.04	/l/	2.81
/ʈ/	0.50	/v/	2.34
/$ʈ^{h}$/	0.14	/ʃ/	0.24
/ɖ/	0.44	/ʂ/	0.01
/$ɖ^{h}$/	0.01	/s/	3.52
/ɳ/	0.01	/h/	9.52
/p/	2.99	/t/	0.10
/p^{h}/	1.44	/z/	0.22
/b/	2.48	/w/	0.01
/b^{h}/	0.05	/f/	8.88
/m/	3.92	/ŋ/	0.00

Table 4.11

Mean percentage of occurrence of consonants in medial word position in Hindi

Consonants	Mean%	Consonants	Mean%
/k/	6.35	/ʈ/	1.98
/$k^{h/}$	0.61	/$ʈ^{h}$/	0.00
/g/	4.44	/d/	1.55
/g^{h}/	0.00	/d^{h}/	0.21
/c/	2.72	/n/	12.09
/c^{h}/	0.22	/j/	6.81
/ɟ/	1.19	/r/	7.52
/$ɟ^{h}$/	0.08	/l/	5.25
/ʈ/	2.17	/v/	0.91
/$ʈ^{h}$/	0.13	/ʃ/	0.51
/ɖ/	1.97	/ʂ/	0.09
/$ɖ^{h}$/	0.04	/s/	5.79
/ɳ/	0.28	/h/	6.87
/p/	6.55	/t/	0.21

/p^{h}/	0.06	/z/	0.44
/b/	1.57	/w/	0.07
/b^{h}/	0.10	/f/	12.29
/m/	8.83	/ŋ/	0.10

Table 4.12

Mean percentage of occurrence of consonants in final word position in Hindi

Consonants	**Mean%**	**Consonants**	**Mean%**
/k/	4.49	/ʈ/	1.38
/k$^{h/}$	0.32	/ʈ^{h}/	0.00
/g/	2.78	/d/	5.67
/g^{h}/	0.02	/d^{h}/	0.00
/c/	3.07	/n/	28.57
/c^{h}/	0.01	/j/	0.70
/ɟ/	1.82	/r/	17.50
/ɟ^{h}/	0.03	/l/	4.81
/ʈ/	2.07	/v/	0.22
/ʈ^{h}/	0.19	/ʃ/	0.27
/ɖ/	1.53	/ʂ/	0.08
/ɖ^{h}/	0.00	/s/	4.88
/ɳ/	0.20	/h/	0.18
/p/	4.49	/t/	0.88
/p^{h}/	0.04	/z/	0.38
/b/	1.71	/w/	0.00
/b^{h}/	0.03	/f/	3.90
/m/	7.35	/ŋ/	0.40

According to Friedman test, a significant difference was present across various word positions, χ^2 (4) = 36, p< 0.05. Application of Wilcoxon signed rank test revealed a significant difference across initial, medial and final word positions (table 4.13). Word initial position had a higher occurrence followed by word medial position. Occurrence of consonants in the final position was less.

Table 4.13

Pair wise comparison of mean percentage occurrence across initial, medial and final word positions using Wilcoxon signed rank test

Manner of articulation	|Z|	p value
Initial- medial	3.724	0.000*
Medial- final	3.724	0.000*
Initial- final	3.724	0.000*

Note. *p<0.05

4. Frequently occurring consonant clusters in Hindi

On application of descriptive statistics, it was observed that consonant cluster /ky/ (10.56%) had a higher occurrence followed by cluster groups /sk/ (6.64%), /t̪n/ (4.98%), geminate /cc/ (4.98%), /rn/ (3.37%), /rʈ/ (3.17%) and /t̪l/ (3.17%). Two- consonant clusters were more frequent than three consonant clusters. Table 4.14 shows the first 56 consonant clusters from the data.

Table 4.14

Mean percentage of occurrence of consonants clusters in Hindi

Consonant clusters	Mean%	Consonant clusters	Mean%
/ky/	10.65	/rf/	0.82
/sk/	6.64	/ks/	0.79
/t̪n/	4.98	/ll/	0.72
/cc/	4.98	/pl/	0.72
/rn/	3.37	/rm/	0.72
/rʈ/	3.17	/ɳʈ/	0.72
/t̪l/	3.17	/sp/	0.67
/pk/	3.01	/hl/	0.64
/kʈ/	2.75	/fr/	0.59
/nk/	2.72	/hn/	0.58

/lʈ/	2.65	/gz/	0.57
/sm/	1.81	/sn/	0.54
/ʈʈ/	1.71	/sʈ/	0.54
/st/	1.56	/nʈ/	0.52
/pn/	1.51	/bj/	0.52
/hʈ/	1.48	/mm/	0.47
/gʈ/	1.44	/rɖ/	0.47
/ns/	1.39	/rv/	0.47
/pr/	1.31	/kr/	0.47
/rk/	1.28	/sl/	0.44
/ln/	1.23	/khn/	0.44
/mn/	1.19	/ɳɖ/	0.42
/rs/	1.16	/sʂ/	0.42
/nɖ/	1.48	/dr/	0.42
/kk/	0.99	/mk/	0.42
/lk/	0.94	/ss/	0.39
/rʈ/	0.94	/gr/	0.39
/kl/	0.91	/kw/	0.37

5. Frequently occurring syllable types in Hindi

Syllable types/ shapes V, CV, VC, CVC, CCV, CVCC, CCVC and VCC were analysed from the data. The most frequent syllable type observed was CV (58.66%) followed by CVC with a mean occurrence of 23.47%. The least occurrence was noted for syllable type VCC (0.03%). The mean percentage of occurrence of syllable types in Hindi is provided in figure 4.9. Similar results have been reported by Khan (1990) and Sinha (2015) for Hindi language. Conversely, De (1973) reported that CVC syllable structure had the highest occurrence (45%) while CV structure occurred for only 30% in frequency in Hindi.

It has been observed that these syllable types were consistently present in typically developing children by the age of 3 years. Shailaja, Manjula and Praveen (2011) studied phonotactics in Hindi speaking typically developing children and children with phonological impairment. Results revealed that syllable shapes CV, CVC, VC, V, CCV, CVCC and VCC were

evident in their speech. Syllable shape CV was predominant in both the groups followed by CVC type. The results of the present study are also in consonance with Rupela and Manjula (2006) and Priya and Manjula (2016) in Kannada and Neethipriya (2007) in Telugu. English language also favoured CV and CVC syllable shapes in almost equal proportions while French and Spanish favoured CV type in higher percentage (Delattre & Olsen, 1969). CV syllable structure also had highest occurrence in Kannada (Jayaram, 1985) Gujarati (Patel, 2004), Punjabi (Singh & Lehal, 2010), Bengali, Marathi and Tamil (Prakash, Prakash and Murthy, 2016). Therefore, it can be stated that CV is the most frequent syllable type in Hindi as reported in Kannada, Telugu, French and Spanish.

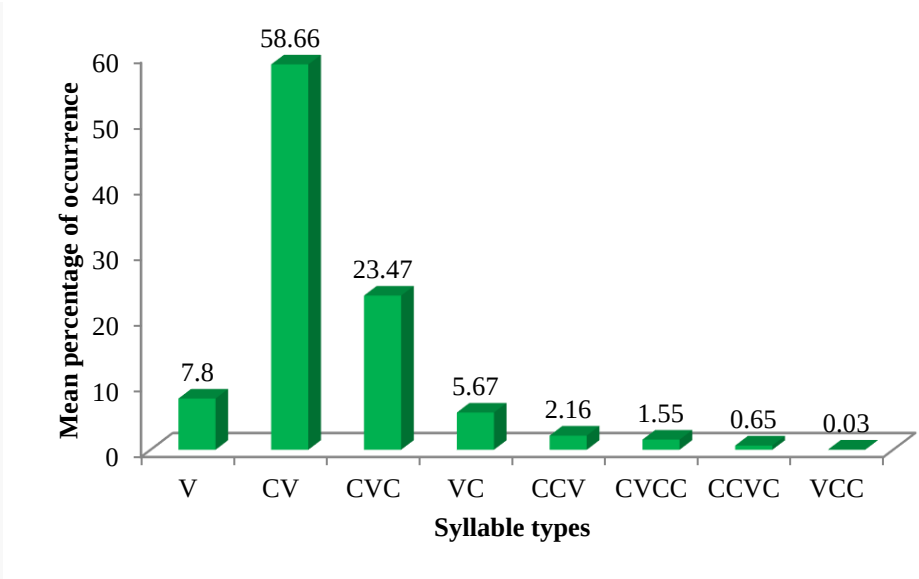

Figure 4.9. Mean percentage of occurrence of syllable types in Hindi

6. Frequently occurring word shapes in Hindi

The mean percentage of occurrence of word shapes in Hindi was computed and is provided in table 4.15. Monosyllables, disyllables, trisyllables and multisyllables were recorded

from the data. Monosyllables accounted for more than 50% of the total data. Disyllables were next in line followed by trisyllables. Overall, the most frequent word shape observed was CV (32.22%) followed by CVCV and CVC with 18.70% and 12.89% of mean occurrence respectively. In monosyllable structure, CV and CVC word shapes occurred more frequently, whereas in disyllables, CV, CV chain was the most frequent word shape. In trisyllables, CVCVCV chain (3.66%) had the highest occurrence. The least occurring word shapes had trisyllabic or multisyllabic structure such as CVCVCVCV (0.13%), CVCVCVCVC (0.01%). The results from the study are in consonance with Khan (1990) in Hindi. Study by Shailaja, Manjula and Praveen (2011) in Hindi speaking typically developing children and children with phonological impairment aged 3-to-5 years yielded similar results. Word shapes CV and CVC had higher occurrence among the monosyllables and CVCV had a higher percentage of occurrence among the disyllables. CVCVCV was highly present among the trisyllables. The VC structure was seen much less in Hindi speaking children. The present study in spoken Hindi reported highest occurrence of monosyllabic word shapes whereas in Kannada disyllabic word shapes predominate as reported by Hiremath (1961).

Table 4.15

Mean percentage of occurrence of word shapes in conversational Hindi

Word length	**Word shapes**	**Mean%**
Monosyllables	V	4.29
	CV	32.22
	VC	3.12
	CVC	12.89
	CCV	1.19
	CVCC	0.85
	CCVC	0.36
	CCV, CVC	0.13
	CCV, CC	0.10
	CV,CV	18.70
	CVC, CV	5.86

Disyllables	CVC, VC	3.58
	CVC, CVC	1.58
	CCV, CV	0.38
	C, CV	1.19
	VC, CV	3.13
	VC, VC	0.76
	VC, VCC	0.01
	CV, V	2.29
Trisyllables	CV,CV,CV	3.66
	CVV, CV	0.68
	CV, CV, CVC	0.33
	CV, CCV, CV	0.36
	CV, CV, CCV	0.21
Polysyllables	CV,CV,CV,CV	0.13
	CV, CV, CV, CVC	0.01

Therefore the present study on frequency of occurrence of phonemes in conversational Hindi revealed higher occurrence of consonants (54.42%) than vowels and diphthongs. The first ten frequently occurring phonemes in descending order are as follows: /i/ > /n/> /a/ > /e/ > /f/> /h/ > /k/> /m/ > /r/ > /o/. Phoneme /i/ had higher occurrence among the vowels while phoneme /n/ was predominant among the consonants. Diphthongs had least occurrence of 1.08%. Non- nasal vowels were more frequent than nasal vowels in Hindi. Considering manner and place of articulation, nasals (34.67%) and alveolars (42.79%) had higher mean percentage in spoken Hindi. Word initial position had a higher occurrence of consonants /n/> /h/>/f/> /k/> /d/ whereas /f/> /n/> /m/> /r/> /h/ were commonly present in word medial position. Phonemes /r/> /m/> /d/> /s/> /l/ were prominent in word final position. Two-consonant clusters were more frequent than three consonant clusters. Consonant cluster /kj/ had a higher occurrence followed by cluster groups /sk/, /ʈn/, geminate /cc/, /rn/, /rʈ/ and /ʈl/. CV (58.66%) was the most frequently observed syllable type in Hindi. Monosyllables accounted for more than 50% of the total data. Word shape CV was highly present while VC was least observed in spoken Hindi.

CHAPTER V

SUMMARY AND CONCLUSIONS

The present study aimed to determine the frequency of occurrence of phonemes, consonant clusters, syllable types and word shapes in conversational Hindi from major Hindi speaking regions, such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. The data on frequently occurring phonemes have widespread applications in various fields such as audiology, speech language pathology, speech engineering and linguistics.

The study included a total of 91 native speakers of Hindi, both males (n = 33) and females (n = 58) in the age range of 20-to-70 years. Data consisted of 18 recordings from individuals native to major Hindi speaking regions of India such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. Each recording has 4-5 participants and was recorded for about 20 minutes. The participants were assessed for any speech, language and hearing difficulties.

The speech samples were collected in controlled natural environments. The participants were made to sit in a circle with the digital audio recorder (Olympus LS 100) placed at the center, equidistant from all the participants. The individuals were encouraged to initiate and speak on any topic of common interest to the group. The conversation samples were transcribed using International Phonetic Alphabet transcription by Ohala (1994) for Hindi language. Commonly used loan English words were included in the data.

The raw data were then coded and analyzed in Systematic Analysis of Language Transcripts (SALT) software (Clinical Demo Version 2012.4.5) to determine the frequency count of phonemes, consonant clusters, syllable types and word shapes. The output obtained

from SALT analysis was further statistically anayzed. Descriptive statistics was employed. The coprus consisted of a total of 1, 48,862 phonemes from the 18 recordings. Mean occurrence of consonants and vowels were 54.42% and 44.50% respectively. Phonemes /i, a, e, o, a:/ and /n, f, h, k, m/ had higher occurrences among vowels and consonants respectively. Considering the total phonemes in spoken Hindi, vowel /i/ had the highest mean percentage of 19.43%. The mean percentage of occurrence of the phonemes in descending order is as follows: /n/ (15.99%)> /a/ (9.17%)> /e/ (6.12%)> /f/ (5.22%)> /h/ (3.99%)> /k/ (3.80%). Diphthongs and aspirated consonants had the least occurrence in the data.

Front vowels (51.8%) and high vowels (63.16%) were predominant in the data. Occurrence of oral vowels was higher compared to nasal vowels. Considering the place of articulation, nasals (34.67%) had higher percentage of occurrence. Stops (25.3%) and fricatives (24.25%) had almost similar mean percentage scores. Affricates had least occurrence of 3.93%. Alveolars (42.79%) occurred in higher frequency while considering place of articulation. This was followed by bilabials and labiodentals. Mean occurrence for retroflexes were the least, only about 2.5%. Hindi had relatively higher occurrence of glottal /h/ when compared to languages like Marathi, Telugu and Malayalam. Consonants /n/ > /h/ > /f/ > /k/ > /d/ > /s/ and /f/ > /n/ > /m/ > /r/ > /h/ > /k/ were commonly present in initial and medial word positions respectively. In the final word position, consonants /n/> /r/> /m/> /d/> /s/> /l/> /k/ were prominent. Consonants were predominant more in the initial position and least in the final position.

Two consonant clusters occurred more in Hindi than three or four consonant clusters. The commonly occurring consonant clusters were /ky/ (eg, /kja/), /sk/ (eg, /kiska/), /ʈn/ (eg, /kiʈna/), /cc/ (eg, /acca/), /rn/ (eg, /karna/), /rʈ/ (eg, /karʈa/) and /ʈl/ (eg, /maʈlab/). The most frequent syllable types observed were CV (58.66%) and CVC (23.47%) while VCC occurred least in the data. Monosyllables accounted for more than 50% of the total data. On the whole, word shape

CV (32.22%) had highest mean percentage followed by CVCV (18.70%) and CVC (12.89%). Word shape VC occurred least in spoken Hindi. Among monosyllables and disyllables, CV and CVCV were predominant respectively. The least occurring word shapes had trisyllabic or multisyllabic structure.

The results of the current study will enable audiologists and speech language pathologists in developing assessment (eg, PB word lists) and intervention (eg, speech sound targets for articulation therapy) tools for the rehabilitation of individuals with communication disorders. It will help explain the speech sound developmental sequence in typically developing children and speech error patterns in individuals with communication disorders. The information can be utilized by speech engineers to design text-to-speech and speech to text systems, translation systems and speech synthesis systems which have become a part of the present day electronic devices and can go a long way in rehabilitation of communication disabled. Learning foreign language may be a simpler task with the utilization of information regarding frequency of phonemes in a language. Hindi being a language with large number of native and non- native speakers, it is necessary to create a database of the phonemes of the language for clinical, linguistic and rehabilitation purposes.

References

Berkson, K. H., & Nelson, M. (2017). Phonotactic frequencies in Marathi. *IULC Working Papers, 17*(1).

Bhagwat, S. V. (1961). *Phonemic frequencies in Marathi and their relation to devising a speed script*. Pune: Deccan College.

Bharadwaja Kumar, G., Murthy, K. N., & Chaudhuri, B. (2007). Statistical analyses of Telugu text corpora. *IJDL. International journal of Dravidian linguistics, 36*(2), 71-99.

Bharati, A., Rao, P., Sangal, R., & Bendre, S. M. (2002). Basic statistical analysis of corpus and cross comparison. *Proc. ICON*.

Black, J. W., & Singh, Sadanand. (1968). The psychological basis of phonetics. *Malmberg. Manual of phonetics. Amsterdam*, 105.

Bridgeman, L. I. (1961). Kaiwa (Guarani) phonology. *International Journal of American Linguistics, 27*(4), 329-334.

Broeder, P., & Murre, J. (Eds.). (1999). *Language and thought in development: cross-linguistic studies* (Vol. 26). Gunter Narr Verlag.

Carroll, J. B. (1952) *Transitional probabilities of English phonemes*. Unpublished Progress Report on Project, 52.

Chourasia, V., Samudravijaya, K., Ingle, M., & Chandwani, M. (2007). 'Statistical analysis of phonetic richness of Hindi text corpora'. *Proc. of Frontiers of Research on Speech and Music Signal Processing AIISH*. 58-63.

Crystal, D. (1981). *Clinical linguistics*. New York: Springer-Verlag Wien.

Dash, N. S. (2009). *Language corpora: past, present and future*. New Delhi: Mittal Publications.

Dauer R.M., 1983, "Stress-timing and syllable-timing reanalyzed", *Journal of Phonetics 11*, 51-62.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of phonetics*.

De, N. S. (1973). Hindi PB list for speech audiometry and discrimination test. *Indian Journal of Otolaryngology, 25*(2), 64-75.

Delattre, P. (1965). *Comparing the phonetic features of English, French, German and Spanish: An interim report*. Groos.

Delattre, P. and Olsen, C, (1969), Syllabic features and phonic impression in English, German, French and Spanish. *Lingua, 22*: 160-175.

Denes, P. B. (1963). On the statistics of spoken English. *The Journal of the Acoustical Society of America, 35*(6), 892-904.

Dewey, G. (1923). *The Relative Frequency of English Speech Sounds* (Cambridge: Harvard University Press).

Edwards, H. T. (2003). *Applied phonetics: the sounds of American English*. United Nations Publications.

Fant, G., & Richter, M. (1958). Some notes on the relative occurrence of letters, phonemes, and words in Swedish. *In Proc. of the VIIIth Intl. Congress of Linguists, Oslo* (pp. 815-816).

Ferguson, C. A., & Chowdhury, M. (1960). The phonemes of Bengali. *Language*, *36*(1), 22-59.

Fowler, M. (1957). Herdan's statistical parameter and the frequency of English phonemes. *Studies presented to Joshua Whatmough (The Hague)*, 45.

French, N. R., Carter, C. W., & Koenig, W. (1930). The words and sounds of telephone conversations. *Bell Labs Technical Journal, 9(2)*, 290-324.

Fry, D. B. (1947). The frequency of occurrence of speech sounds in Southern English. *Archives néerlandaises de phonétique expérimentale, 20*, 103-106.

Gerber, S. E., & Vertin, S. (1969). Comparative frequency counts of English phonemes. *Phonetica, 19*(3), 133-141.

Ghatage, A. M. (1964). *Phonemic and Morphemic frequencies in Hindi*. Poona: Deccan College Postgraduate and Research Institute.

Ghatage, A. M. (1994). *Phonemic and morphemic frequencies in Malayalam*. Mysore: Central Institute of Indian Languages.

Ghazali, M. (2002). Urdu Syllable Templates. *Annual Report of Center for Research in Urdu Language Processing (CRULP).*

Giegerich, H. J. (1992). *English phonology: An introduction*. England: Cambridge University Press.

Guirao, M. and Borzone de Manrique, A. M., (1972). Fonemas, silabas y palabras del espanol de Buenos Aires. *Filologia*, XVI: 135-165.

Guirao, M., & García Jurado, M. (1990). Frequency of Occurrence of Phonemes in American Spanish. *Revue quebecoise de linguistique, 19*(2), 135-149.

Hanna, P. R., Hanna, J. S., Hodges, R. E., & Rudorf, E. H. (1966). *Phoneme-grapheme correspondences as cues to spelling improvement*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Hayden, R. E. (1950). The relative frequency of phonemes in general-American English. *Word, 6*(3), 217-223.

Hiremath, R. C. (1961). *The structure of Kannada*. Karnatak University.

Jayaram, M. (1985). Sound and Syllable distribution in written Kannada and their application to Speech and Hearing. *Journal of All India Institute of Speech and Hearing, 16*, 19-30.

Kachru, Y. (2006). *Hindi* (Vol. 12).Philadelphia: John Benjamins Publishing.

Kalyani, N., & Sunitha, D. K. (2009). Syllable analysis to build a dictation system in Telugu language. *arXiv preprint arXiv:1001.2263*.

Kelkar, A. R. (1994). *Phonemic and Morphophonemic frequency in Oriya*. Mysore: Central Institute of Indian Languages.

Kelkar, A. R. (1994). *Phonemic and Morphophonemic frequency in Oriya*. Mysore: Central Institute of Indian Languages.

Kenyon, J. S., & Knott, T. A. (1951). *A Pronunciation Dictionary of American English*. G & C. Merriam.

Khan, I. (1990). *Statistical study of Hindi speech sounds* (Doctoral dissertation, Aligarh Muslim University).

Kumar, S. R., & Mohanty, P. (2012). Speech recognition performance of adults: A proposal for a battery for Telugu. *Theory and Practice in Language Studies, 2*(2), 193.

Kumar, S. R., & Mohanty, P. (2012). Speech recognition performance of adults: A proposal for a battery for Telugu. *Theory and Practice in Language Studies, 2*(2), 193.

Leung, M. T., Law, S. P., & Fung, S. Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, 36*(3), 500-505.

Lotz, J. (1952). Vowel frequency in Hungarian. *Word, 8*(3), 227-235.

Mader, J. B. (1954). The relative frequency of occurrence of English consonant sounds in words in the speech of children in grades one, two, and three. *Communications Monographs, 21*(4), 294-300.

Malécot, A. (1974). Frequency of occurrence of French phonemes and consonant clusters. *Phonetica, 29*(3), 158-170.

Malviya, S., Mishra, R., & Tiwary, U. S. (2016, October). Structural analysis of Hindi phonetics and a method for extraction of phonetically rich sentences from a very large Hindi text corpus. In *Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA), 2016 Conference of The Oriental Chapter of International Committee for* (pp. 188-193). IEEE.

Manjula, P., Sharathkumar K. S., Antony, J., & Geetha, C. (2015). Development of Phonemically Balanced Word List in Kannada Language for Adults. *Journal of Hearing Science, 5*(1), 22-30.

McMahon, A. (2002). *An introduction to English phonology*. Edinburgh: Edinburgh University Press.

Meena, R. L . (2015, September). Re: Learning of Hindi Phonology as a Foreigner. Retrieved from https://bhashhiki.blogspot.com/2017/01/learning-of-hindi-phonology-as.html

Miller, J. & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Clinical Demo Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Miller, J. F., & Iglesias, A. (2008). Systematic Analysis of Language Transcripts (SALT), English & Spanish (Version 9) [Computer software]. Madison: University of Wisconsin—Madison, Waisman Center. *Language Analysis Laboratory*.

Mines, M. A., Hanson, B. F., & Shoup, J. E. (1978). Frequency of occurrence of phonemes in conversational English. *Language and Speech*, *21*(3), 221-241.

Mohanan, T. (1989). Syllable structure in Malayalam. *Linguistic Inquiry*, 589-625.

Munthuli, A., Tantibundhit, C., Onsuwan, C., Kosawat, K., & Wutiwiwatchai, C. (2015). Frequency of occurrence of phonemes and syllables in Thai: Analysis of spoken and written corpora. *In Proceedings of 18th International Congress of Phonetic Sciences*.

Nair, K. K., & Ramachandran, V. (1958). A note on the statistical studies in Kannada speech sounds. *IETE Journal of Research*, *4*(4), 197-198.

Navarro Tomás, T. (1946). Escala de frecuencia de los fonemas españoles. *Estudios de fonología española*, 15-30.

Nāyaka, H. M. (1967). *Kannada, Literary and Colloquial: A study of two styles*. Mysore: Rao and Raghavan.

Neethipriya, N. (2007). *Aspects of phonotactics in Typically Developing Telugu Speaking Children*. Unpublished Masters Dissertation submitted to University of Mysore.

Ohala, M. (1983). Aspects of Hindi phonology (Vol. 2). Motilal Banarsidass Publishe.

Ohala, M. (1994). Hindi. *Journal of the International Phonetic Association*, *24*(1), 35-38.

Palai, E. B., & O'Hanlon, L. (2004). Word and phoneme frequency of occurrence in conversational Setswana: a clinical linguistic application. *Southern African Linguistics and Applied Language Studies, 22*(3-4), 125-142.

Pandey, P. (2007). Phonology–orthography interface in Devanāgarī for Hindi. *Written Language & Literacy*, *10*(2), 139-156.

Pandey, P. (2014). Akshara-to-sound rules for Hindi. *Writing Systems Research*, *6*(1), 54-72.

Pandit, P. B. (1965). *Phonemic and morphemic frequencies of the Gujarati language*. Deccan College Postgraduate and Research Institute.

Patel, P. G. (2004). *Reading acquisition in India: Models of learning and dyslexia* (Vol. 6). Sage.

Prakash, A., Prakash, J. J., & Murthy, H. A. (2016). Acoustic analysis of syllables across Indian languages. In *INTERSPEECH* (pp. 327-331).

Priya, M. B., & Manjula, R. (2016a). Articulation judgment abilities and phonological representations in typically developing preschoolers. *Journal of Advanced Linguistic Studies, 5*(1-2), 175-197.

R. Sagon, (2006). "The development of a phonetically balanced word recognition test in the Ilocano language," Ph.D. dissertation, Washington University School of Medicine, Washington.

Ramakrishna, B. S., Nair, K. K., Chiplunkar, V. N., Atal, B. S., Ramachandran, V., & Subramanan, R. (1962). *Some aspects of the relative efficiencies of Indian languages: A study from information theory point of view*. Bangalore: Indian Institute of Science.

Ramakrishna, B. S., Nair, K. K., Chiplunkar, V. N., Atal, B. S., & Rajaraman, V. (1957). Statistical studies in some Indian languages with applications to communication engineering. *IETE Journal of Research, 4*(1), 25-35.

Ramaswami, N. (1999). *Common linguistic features in Indian languages: Phonetics* (No. 447). Central Institute of Indian Languages.

Ranganatha, M. R. (1982). *Morphophonemic analysis of the Kannada language: Relative frequency of phonemes and morphemes in Kannada*. Central Institute of Indian Languages.

Ranganatha, M. R. (1982). *Morphophonemic analysis of the Kannada language: Relative frequency of phonemes and morphemes in Kannada*. Mysore: Central Institute of Indian Languages.

Rao, G. U. & Thenarasu, S. (2007). PGDCAIL: CAIL-421. *Corpus Linguistics*, University of Hyderabad.

Renwick, M. E. (2011). Phoneme Type Frequency in Romanian. University of Pennsylvania *Working Papers in Linguistics, 17*(1), 22.

Roberts, A. H. (1965). *A statistical linguistic analysis of American English*. Mouton.

Robson, J., Pring, T., Marshall, J., & Chiat, S. (2003). Phoneme frequency effects in jargon aphasia: A phonological investigation of nonword errors. *Brain and language*, *85*(1), 109-124.

Rovenchak, A. (2011). Phoneme Distribution, Syllabic Structure, and Tonal Patterns in Nko Texts. *Mandenkon*, *47*, 77-96.

Rupela, V., & Manjula, R. (2006). Phonotactic development in Kannada: some aspects and future directions. In *Language Forum: A Journal of Language and Literature* (Vol. 32, No. 1-2, pp. 83-93).

Sandoval, A. M., Toledano, D. T., de la Torre, R., Garrote, M., & Guirao, J. M. (2008). Developing a phonemic and syllabic frequency inventory for spontaneous spoken Castilian Spanish and their comparison to text-based inventories. *In Language Resource and Evaluation Conference* (pp. 1097-1100).

Shin, J. (2008). Phoneme and syllable frequencies of Korean based on the analysis of spontaneous speech data. *Communication Sciences & Disorders*, *13*(2), 193-215.

Shin, J. (2010). Phoneme and syllable frequencies based on the analysis of entries in the Korean dictionary. *Communication Sciences & Disorders*, *15*(1), 94-106.

Shin, J., Kiaer, J., & Cha, J. (2012). *The sounds of Korean*. New York: Cambridge University Press.

Shukla, S., Manjula, R., & Praveen, H. R. (2011). Phonotactic patterns in conversational speech of typically developing children and children with phonological impairment: a comparison. *Journal of the All India Institute of Speech & Hearing*, *30.*

Sigurd, B. (1968). Rank-frequency distributions for phonemes. *Phonetica, 18*(1), 1-15.

Singh, P., & Lehal, G. S. (2010). Corpus based statistical analysis of Punjabi syllables for preparation of Punjabi speech database. *International Journal of Intelligent Computing Research (IJICR), 1*(3).

Singh, P., & Lehal, G. S. (2010). Syllables Selection for the Development of Speech Database for Punjabi TTS System. *IJCSI International Journal of Computer Science Issues, 7*(6), 164-168.

Sinha, S. (2015). Analysis and Recognition of Dialects of Hindi Speech. *International Journal of Scientific Research in Computer Science and Engineering, 3*(5), 1-5.

Sinha, S., Jain, A., & Agrawal, S. S. (2014). Speech processing for Hindi dialect recognition. In *Advances in Signal Processing and Intelligent Recognition Systems* (pp. 161-169). Springer, Cham.

Smirnova, N., & Chistikov, P. (2011). Statistics of Russian monophones and diphones. *Proc. of Specom-2011. Kazan, Russia*, 218-223.

Sreedevi, N., & Irfana, M. (2013). *Frequency of occurrence of phonemes in Malayalam.* ARF Project. AIISH, Mysore.

Sreedevi, N., Smitha, N., & Vikas, M.D. (2012). *Frequency of phonemes in Kannada.* ARF Project. AIISH, Mysore.

Tamaoka, K., & Makioka, S. (2004). Frequency of occurrence for units of phonemes, morae, and syllables appearing in a lexical corpus of a Japanese newspaper. *Behavior Research Methods, 36*(3), 531-547.

Tambovtsev, Y. (2007). How Can Typological Distances between Latin and Some Indo-European Language Taxa Improve Its Classification?. *Prague Bull. Math. Linguistics*, *88*, 73-90.

Tarnóczy, T. (1961). Phonetische Gesichtspunkte bei der Zusammenstellung von Texten für Verständlidikeitsmessungen. *STUF-Language Typology and Universals, 14(1-4)*, 74-87.

Thomas, T. W. C. (2005). The effects of occurrence frequency of phonemes on second language acquisition: A quantitative comparison of Cantonese, Mandarin, Italian, German and American English. *Chinese University of Hong Kong. Available at http://www.thomastsoi.com/wpcontent/downloads/The%20Effects%20of%20Occurrence %20Frequency%20of%20Phonemes% 20on% 20SLA.pdf (Last viewed 30 September 2015).*

Tobias, J. V. (1959). Relative occurrence of phonemes in American English. *The Journal of the Acoustical Society of America*, *31*(5), 631-631.

Travis, L. E. (1931). *Speech pathology; a dynamic neurological treatment of normal speech and speech deviations*.

Trnka. (1935). A Phonological Analysis Of Present-Day Standard English, By B. TRANKA, Studies in English by Members of the English Seminar of the Charles University, Prague, Fifth Volume, 1935 (批評紹介). 英文学研究, 16(2), 288-291.

Vasanthakumari, T. (1989). *Generative phonology of Tamil*. New Delhi: Mittal Publications.

Voelker, C. H. (1934). Phonetic distribution in formal American pronunciation. *The Journal of the Acoustical Society of America, 5*(4), 242-246.

Voelker, C. H. (1937). A comparative study of investigations of phonetic dispersion in connected American English. *Arch. Nderlandaises de Phondtique Expdri-~ tentale,* 13, 138-152.

Wang, W. S., & Crawford, J. (1960). Frequency studies of English consonants. *Language and Speech, 3*(3), 131-139.

Whitney, W. D. (1874). *The proportional elements of English utterance. Proceedings of the American Philological Association,* 6, 14-7.

Yadav, R. (1976). Generative Phonology and the Aspirated Consonants of Colloquial Maithili'. *Contributions to Nepalese Studies, 4*(1), 77-91.

Yegerlehner, J., & Voegelin, F. M. (1957). Frequencies and inventories of phonemes from nine languages. *International Journal of American Linguistics, 23*(2), 85-93.

Zangwill, O. L. (1975). A phonological investigation of aphasic speech. *Linguistics,* 154-155, 163-164.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Oxford, England: Addison-Wesley Press.

Zipf, G. K., & Rogers, F. M. (1939). Phonemes and Variphones in four present-day romance Languages and Classical Latin from the viewpoint of dynamic Philology. *Archives néerlandaises de phonétique expérimentale, 15*, 111-147.

Zuidema, W (2009). *A syllable frequency list for Dutch*. *Geo Journal,* 1-9.

Appendix A

Phonemes of Hindi

Table A1

Depicts vowels in Hindi (Ohala, 2004)

Non- nasal vowels	Nasal vowels
/a/, /a:/	/ã/, /ã:/
/i/, /i:/	/ĩ/, /ĩ:/
/u/, /u:/	/ẽ/
/e/	/ũ/, /ũ:/
/o/	/õ/
/æ/	
/ə/	

Dipthongs: /ai/, /au/

Table A2

Depicts consonants in Hindi (Ohala, 2004)

	Bilabial	**Labio-dental**	**Dent-al**	**Alveolar**	**Post-alveolar**	**Retro-flex**	**Palatal**	**Velar**	**Glottal**
Stop	p, b		t̪, t̪^{h} d̪, d̪^{h}			ʈ, ʈ^{h} ɖ, ɖ^{h}		k, k^{h} g, g^{h}	
Affricate					c, c^{h} ɟ, ɟ^{h}				
Nasal	m			n			ŋ		
Fricative		f		s, z	ʃ				h
Trill						r			
Approxi-mant		v		l			j		

Appendix B

Loan words included in the study

Real	Standard	Recent	Maths
Nurse	Non-stop	Change	Board
Stage	Continue	Tax	Toper
Public	Cartoon	Choice	Top
Nursing	Love	Late	Number
Experience	Like	Group	Result
Important	Ward	Hospital	State
College	Family	Percent	Seat
Class	Afternoon	North India	Exams
Topic	Duty	Golden	First
Start	Last	Idea	Point
Actually	Second	Car	Normal
Show	Year	Police	Page
Next	Doctors	Smile	Date
Power	Colour	Operation	Diploma
Portion	Painting	Sir	Kilo
First	Type	Best	Company
Time	Hostel	Garden	Water
Mam	Senior	Dam	Lunch
Madam	Signal	Newspaper	Doctor
Competition	Ball	Paper	Teacher
Dress	Cream	Bridge	Guest
Quiz	Rules	News	First
Level	Clinics	Plan	Pressure
Rank	Teacher	Special	Reservation
Party	White	Enjoy	Caste
Participate	Medical	Same	Last
Person	Science	Course	Tension
School	South	South	Extra
Minute	Train	Good	Group
Dance	Holiday	Market	Hospital
Job	Day	Card	Friends
Students	Viva	System	Smoke
Explain	Photo	Salary	Passport
Day	Ticket	Tough	Different colours,
Auto	Room	Engineering	numbers, days of
Road	Vacation	Problem	week, months
Steps	Posting	Marks	
Morning	Facebook	Cancel	
Patients	Side	Coaching	

Appendix C

Table C1

Depicts codes used for vowels, diphthongs and consonants in Hindi

Phonemes	Codes	Phonemes	Codes
/a/	a	/p^h/	ph
/aː/	aa	/b/	b
/i/	i	/b^h/	bh
/iː/	ii	/m/	m
/u/	u	/t̪/	th
/uː/	uu	/$t̪^h$/	thh
/e/	e	/d/	d
/o/	o	/d^h/	dh
/æ/	ae	/n/	n
/ə/	aee	/j/	y
/ã/	am	/r/	r
/ãː/	aam	/l/	l
/ĩ/	im	/v/	v
/ĩː/	iim	/ʃ/	sh
/ẽ/	em	/ʂ/	shh
/ũ/	um	/s/	s
/ũː/	uum	/h/	h
/õ/	om	/t/	tt
/ai/	ai	/z/	z
/au/	au	/w/	w
/k/	k	/f/	f
/$k^{h/}$	kh	/ŋ/	ng
/g/	g		
/g^h/	gh		
/c/	ch		
/c^h/	chh		
/ɟ/	j		
/$ɟ^h$/	jh		
/ʈ/	T		
/$ʈ^h$/	TTh		
/ɖ/	DD		
/$ɖ^h$/	DDh		
/ɳ/	nn		
/p/	p		

Table C2

Depicts consonant clusters and their codes in the present study

Clusters	Codes	Clusters	Codes	Clusters	Codes	Clusters	Codes	Clusters	Codes
/kj/	ky	/dk/	dk	/k^{h}b/	khb	/pc/	pch	/rɟj/	rjy
/kv/	kv	/dl/	dl	/k^{h}m/	khm	/bd/	bd	/rʈj/	rthy
/k^{h}j/	khy	/d^{h}k/	dhk	/gz/	gz	/ɟm/	jm	/rʈm/	rthm
/gj/	gy	/d^{h}m/	dhm	/gm/	gm	/ɟn/	jn	/rʈ^{h}m/	rthhm
/gv/	gv	/pʈ/	pT	/ʈb/	thb	/fl/	fl	/rʈ^{h}j/	rthhy
/ɟj/	jy	/pk/	pk	/ʈf/	thf	/fs/	fs	/rdr/	rdr
/ɟv/	jv	/pɟ/	pj	/db/	db	/fr/	fr	/rd^{h}v/	rdhv
/ɖj/	DDy	/pʈ/	pth	/ds/	ds	/nɖ/	nDD	/rhj/	rhy
/ʈj/	thy	/ps/	ps	/bʈ/	bth	/nʈ/	nT	/rmj/	rmy
/pj/	py	/pn/	pn	/bg/	bg	/ʃk/	shk	/lkj/	lky
/p^{h}j/	phy	/p^{h}r/	phr	/bz/	bz	/ʂk/	shhk	/ŋɖj/	nnDDy
/bj/	by	/bʈ/	bT	/bʃ/	bsh	/ʃw/	shw	/ŋɖr/	nnDDr
/nj/	ny	/b^{h}k/	bhk	/bʂ/	bshh	/ʃn/	shn	/nʈj/	nthy
/mj/	my	/cʈ/	chT	/ɟb/	jb	/hw/	hw	/ndv/	ndv
/sj/	sy	/ck/	chk	/ɟd/	jd	/vʈ/	vth	/ndj/	ndy
/sv/	sv	/cr/	chr	/zb/	zb	/wʈ/	wth	/rʈsj/	rthsy
/sw/	sw	/cl/	chl	/zd/	zd	/hn/	hn		
/ʈr/	Tr	/cn/	chn	/zk/	zk	/hl/	hl		
/ɖr/	DDr	/c^{h}ʈ/	chhth	/zl/	zl	/hs/	hs		
/bl/	bl	/c^{h}l/	chhl	/zr/	zr	/rʈ/	rT		
/sʈ/	sT	/c^{h}m/	chhm	/zh/	zh	/rd/	rd		
/spl/	spl	/c^{h}n/	chhn	/zn/	zn	/rɟ/	rj		
/skr/	skr	/ɟr/	jr	/zm/	zm	/rʃ/	rsh		
/kʂ/	kshh	/ɟl/	jl	/fʈ/	fth	/rʂ/	rshh		
/kʃ/	ksh	/ɟw/	jw	/fg/	fg	/rz/	rz		
/g^{h}r/	ghr	/ɟ^{h}k/	jhk	/fv/	fv	/lg/	lg		
/cj/	chy	/ɟ^{h}l/	jhl	/fw/	fw	/lʃ/	lsh		
/ʈr/	thr	/rk/	rk	/fn/	fn	/lɟ/	lj		
/ʈv/	thv	/rk^{h}/	rkh	/sb/	sb	/jr/	yr		
/dr/	dr	/rg/	rg	/sd/	sd	/mʈ/	mth		
/dj/	dy	/rg^{h}/	rgh	/sɟ/	sj	/nʃ/	nsh		
/dv/	dv	/rʈ^{h}/	rTTh	/sf/	sf	/rʈg/	rttg		
/d^{h}r/	dhr	/rʈ/	rth	/ʃʈ/	shth	/rʈg/	rthg		
/d^{h}j/	dhy	/rp/	rp	/ʂʈ/	shhth	/mpr/	mpr		
/b^{h}r/	bhr	/rb/	rb	/ʃg/	shg	/nfl/	nfl		
/nr/	nr	/rc/	rch	/jd/	yd	/sʈm/	sthm		
/mr/	mr	/rc^{h}/	rchh	/jb/	yb	/mbr/	mbr		

/ml/	ml	/rɟʰ/	rjh	/hf/	hf	/mbʰl/	mbhl
/vr/	vr	/rl/	rl	/hʃ/	hsh	/nggr/	nggr
/vj/	vy	/rs/	rs	/hd/	hd	/nggl/	nggl
/hr/	hr	/rv/	rv	/hb/	hb	/ngʰr/	nghr
/ʂm/	shhm	/rw/	rw	/hʈ/	hth	/ŋcm/	nnchm
/ʃm/	shm	/rh/	rh	/hk/	hk	/ŋɟʰr/	nnjhr
/ʃv/	shv	/rm/	rm	/rf/	rf	/ŋɟʰl/	nnjhl
/ʃl/	shl	/rn/	rn	/lz/	lz	/ŋɟr/	nnjr
/ʂl/	shhl	/lk/	lk	/lf/	lf	/ŋɖl/	nnDDl
/ʃr/	shr	/lʈ/	lT	/md/	md	/ndʰl/	ndhl
/ʂr/	shhr	/lɖ/	lDD	/mg/	mg	/ntʰr/	nthr
/sr/	sr	/lʈ/	lth	/mz/	mz	/ndr/	ndr
/skʰ/	skh	/lʈʰ/	lthh	/nz/	nz	/ndʰr/	ndhr
/sʈʰ/	sthh	/ld/	ld	/gd/	gd	/kʃm/	kshm
/sʈ/	sth	/lp/	lp	/gdʰ/	gdh	/ʈkr/	thkr
/spʰ/	sph	/lb/	lb	/ʈp/	thp	/ʈsn/	thsn
/sk/	sk	/lc/	lch	/ʈs/	ths	/ʈsj/	thsy
/sʈr/	sthr	/lcʰ/	lchh	/ʈʰj/	thhy	/ʈpr/	thpr
/smr/	smr	/lɟʰ/	ljh	/ʈʰw/	thhw	/rʈr/	rthr
/ʃj/	shy	/lr/	lr	/ʈʰv/	thhv	/rʈʰkn/	rthhn
/kr/	kr	/ls/	ls	/ʈʰm/	thhm	/ngkʰj/	ngkhy
/kl/	kl	/lv/	lv	/dg/	dg	/nggj/	nggy
/gr/	gr	/lw/	lw	/dgʰ/	dgh	/ngkr/	ngkr
/gl/	gl	/lh/	lh	/dbʰ/	dbh	/ndʰj/	ndhy
/pr/	pr	/lm/	lm	/dw/	dw	/kk/	kk
/pl/	pl	/ln/	ln	/dʰw/	dhw	/kkh/	kkh
/br/	br	/sl/	sl	/dʰv/	dhv	/gg/	gg
/sp/	sp	/jk/	yk	/dʰn/	dhn	/cc/	chch
/sm/	sm	/jʈ/	yth	/pd/	pd	/ɟɟ/	jj
/sn/	sn	/jc/	ych	/pw/	pw	/ʈʈ/	TT
/spr/	spr	/jl/	yl	/pv/	pv	/ɖɖ/	DDDD
/st/	stt	/js/	ys	/pm/	pm	/ŋŋ/	nn
/kʈ/	kT	/vk/	vk	/bj/	bj	/mm/	mm
/kʈ/	kth	/wk/	wk	/bdʰ/	bdh	/pp/	pp
/kc/	kch	/wc/	wT	/cm/	chm	/bb/	bb
/kw/	kw	/vʈ/	vT	/jg/	jg	/ll/	ll
/km/	km	/wɖ/	wDD	/shʈʰ/	shTTh	/ss/	ss
/kn/	kn	/vɖ/	vDD	/shp/	shp	/ʈʈ/	thth
/kʰʈ/	khT	/wr/	wr	/sʈʰ/	sTTh	/kʰs/	khs
/kʰr/	khr	/wl/	wl	/rdʰ/	rdh	/kʰʃ/	khsh
/kʰl/	khl	/vl/	vl	/rbʰ/	rbh	/lkʰ/	lkh

/k^{h}n/	khn	/mk/	mk	/ɾʈ^{h}/	rthh	/nf/	nf
/gʈ/	gT	/mʈ/	mT	/rj/	ry	/g^{h}n/	ghn
/gʈ/	gth	/md^{h}/	mdh	/ld^{h}/	ldh	/ʧj/	Ty
/gn/	gn	/mc/	mch	/lj/	ly	/lp^{h}/	lph
/g^{h}ʈ/	ghT	/mJ^{h}/	mjh	/vd/	vd	/lb^{h}/	lbh
/g^{h}ʈ/	ghth	/mʃ/	msh	/wd/	wd	/ʃʈ/	shT
/g^{h}l/	ghl	/mʂ/	mshh	/vd^{h}/	vdh	/nʂ/	nshh
/ʈk/	Tk	/mh/	mh	/vn/	vn	/ŋʈ/	nnT
/ʈl/	Tl	/mn/	mn	/wn/	wn	/ŋɖ/	nnDD
/ʈv/	Tv	/nk/	nk	/ws/	ws	/ks/	ks
/ʈw/	Tw	/nk^{h}/	nkh	/vs/	vs	/kʃr/	kshr
/ʈn/	Tn	/nb/	nb	/jn/	yn	/kʃj/	kshy
/ʈ^{h}k/	TThk	/nw/	nw	/hj/	hy	/gd^{h}j/	gdhy
/ʈ^{h}r/	TThr	/nh/	nh	/nm/	nm	/ʈrj/	thry
/ʈ^{h}l/	TThl	/ns/	ns	/nl/	nl	/ʃʈr/	shTr
/ʈl/	thl	/ʈf/	Tf	/ms/	ms	/ʃʈ^{h}j/	shTThy
/ʈw/	thw	/bm/	bm	/ʈm/	thm	/sʈj/	sthhy
/ʈn/	thn	/rɖ/	rDD	/dn/	dn	/sʈj/	sthy
/ʈ^{h}r/	thhr	/kb/	kb	/dm/	dm	/rkj/	rky
/ʈ^{h}l/	thhl	/kd/	kd	/dʃ/	dsh	/rk^{h}j/	rkhy
/ ʈ^{h}n/	thhn	/kf/	kf	/dʂ/	dshh	/rg^{h}j/	rghy

Table C3

Depicts syllable types in conversational Hindi

Syllable types	
V	CCV
CV	CVCC
CVC	CCVC
VC	VCC

Table C4

Depicts word length and word shapes in conversational Hindi

Word length	**Word shapes**
Monosyllables	V CV VC CVC CCV CVCC CCVC
Disyllables	CCV, CVC CCV, CC CV,CV CVC, CV CVC, VC CVC, CVC CCV, CV C, CV VC, CV VC, VC VC, VCC CV, V
Trisyllables	CV,CV,CV CVV, CV CV, CV, CVC CV, CCV, CV CV, CV, CCV
Polysyllables	CV,CV,CV,CV CV, CV, CV, CVC

Appendix D

Table D.1

Median and standard deviation of occurrence of phonemes in spoken Hindi in descending order

Phonemes	Median	SD	Phonemes	Median	SD	Phonemes	Median	SD
/i/	1613.0	317.7	/u/	88.5	23.5	/ũ/	10.0	9.9
/n/	1330.0	268.5	/b/	96.0	21.3	/z/	12.0	6.4
/a/	724.5	160.8	/ai/	70.5	22.9	/æ/	9.5	7.7
/e/	507.5	77.8	/ɟ/	9.0	4.8	/t/	8.00	9.0
/f/	420.0	97.5	/v/	71.0	13.9	/ã:/	9.5	8.6
/h/	314.0	75.5	/ʈ/	46.5	22.9	/c^h/	9.0	4.8
/k/	320.0	48.7	/ã/	50.5	20.6	/$ʈ^h$/	6.0	4.8
/m/	260.0	55.2	/ɖ/	46.5	12.9	/ɳ/	3.5	6.5
/r/	240.0	52.4	/p^h/	24.0	31.3	/ŋ/	2.0	4.7
/o/	243.5	52.3	/k^h/	25.0	15.5	/$ʈ̤^h$/	2.0	2.9
/s/	183.0	50.4	/i:/	25.0	7.8	/g^h/	2.0	2.2
/p/	179.5	38.9	/ə/	25.0	6.9	/b^h/	3.0	2.3
/a:/	175.0	40.5	/ũ:/	22.5	22.5	/õ/	0.00	7.3
/l/	154.5	41.6	/d^h/	19.5	19.4	/ĩ:/	0.5	4.2
/d/	131.5	34.6	/u:/	19.0	6.1	/$ɟ^h$/	2.0	2.2
/j/	129.0	25.5	/i/	15.5	10.6	/ʂ/	0.00	3.7
/ʈ̤/	119.0	27.9	/au/	14.5	8.1	/w/	0.50	1.9
/g/	96.0	26.21	/ẽ/	13.5	6.65	/$ɖ^h$/	0.50	1.2
/c/	97.5	32.51	/ʃ/	14.0	5.9			

Table D.2

Median and standard deviation of occurrence of vowels and diphthongs in Hindi

Vowels	Median	SD	Vowels	Median	SD	Diphthongs	Median	SD
/a/	724.50	160.89	/ã/	50.50	20.65	/ai/	70.50	22.95
/a:/	175.00	40.55	/ã:/	9.50	8.64	/au/	14.50	8.10
/i/	1613.00	317.76	/ĩ/	15.50	10.65			
/i:/	25.00	7.815	/ĩ:/	0.50	4.27			
/u/	88.50	23.58	/ẽ/	13.50	6.65			
/u:/	19.00	6.18	/ũ/	10.00	9.98			
/e/	507.50	77.85	/ũ:/	22.50	22.51			
/o/	243.50	52.38	/õ/	0.00	7.38			
/æ/	9.50	7.73						
/ə/	25.00	6.94						

Table D.3

Median and standard deviation of occurrence of consonants in Hindi

Consonants	Median	SD	Consonants	Median	SD
/k/	320.00	48.71	/t̪/	119.00	27.95
/kʰ/	25.00	15.53	/t̪ʰ/	2.00	2.92
/g/	96.00	26.21	/d/	131.50	34.69
/gʰ/	2.00	2.21	/dʰ/	19.50	19.48
/c/	97.50	32.51	/n/	1330.00	268.54
/cʰ/	9.00	4.80	/j/	129.00	25.56
/ɟ/	67.50	14.34	/r/	240.00	52.43
/ɟʰ/	2.00	2.22	/l/	154.50	41.68
/ʈ/	46.50	22.97	/v/	71.00	13.99
/ʈʰ/	6.00	4.84	/ʃ/	14.00	5.94
/ɖ/	46.50	12.97	/ʂ/	0.00	3.78
/ɖʰ/	0.50	1.23	/s/	183.00	50.41
/ɳ/	3.50	6.56	/h/	314.00	75.58
/p/	179.50	38.98	/t/	8.00	9.01
/pʰ/	24.00	31.35	/z/	12.00	6.44
/b/	96.00	21.34	/w/	0.50	1.90
/bʰ/	3.00	2.34	/f/	420.00	97.50
/m/	260.00	55.22	/ŋ/	2.00	4.70

Table D.4

Median and standard deviation of occurrence of consonants in initial word position in Hindi

Consonants	Median	SD	Consonants	Median	SD
/k/	200.00	36.69	/ʈ/	76.50	24.42
/kʰ/	15.50	13.73	/ʈʰ/	2.00	2.97
/g/	23.00	12.91	/d/	87.00	28.91
/gʰ/	2.00	2.28	/dʰ/	18.50	15.86
/c/	30.50	28.14	/n/	920.50	309.34
/cʰ/	5.50	4.17	/j/	26.00	12.80
/ɟ/	43.50	14.78	/r/	42.00	18.77
/ɟʰ/	0.00	1.62	/l/	63.00	26.35
/ʈ/	13.00	5.21	/v/	57.50	13.99
/ʈʰ/	2.00	3.36	/ʃ/	6.00	3.47
/ɖ/	10.00	7.39	/ʂ/	.00	.548
/ɖʰ/	0.00	0.46	/s/	80.00	22.90
/ɳ/	0.00	0.70	/h/	214.50	52.40
/p/	66.00	22.33	/t/	1.00	3.31
/pʰ/	22.50	30.52	/z/	4.00	4.28
/b/	60.50	15.18	/w/	.00	.69
/bʰ/	0.50	1.96	/f/	173.50	126.59
/m/	91.50	41.57	/ŋ/	.00	.00

Table D.5

Median and standard deviation of occurrence of consonants in medial word position in Hindi

Consonants	Median	SD	Consonants	Median	SD
/k/	90.00	15.99	/t̪/	25.50	12.04
/kʰ/	7.00	6.57	/t̪ʰ/	0.00	0.23
/g/	57.00	20.97	/d/	19.50	8.83
/gʰ/	.00	.00	/dʰ/	0.00	4.69
/c/	33.00	20.67	/n/	138.50	107.61
/cʰ/	2.50	2.64	/j/	88.00	21.08
/ɟ/	16.50	5.12	/r/	95.00	23.64
/ɟʰ/	.50	1.90	/l/	65.00	21.43
/ʈ/	28.50	14.36	/v/	12.50	4.35
/ʈʰ/	1.00	1.95	/ʃ/	6.00	5.09
/ɖ/	28.00	7.04	/ʂ/	.00	2.98
/ɖʰ/	.00	1.09	/s/	78.00	23.93
/ɳ/	1.00	5.87	/h/	82.50	45.08
/p/	81.50	25.05	/t/	2.00	1.65
/pʰ/	.00	1.16	/z/	4.50	4.79
/b/	18.00	11.96	/w/	.00	1.62
/bʰ/	1.00	1.45	/f/	177.50	126.02
/m/	113.00	37.77	/ŋ/	.00	3.31

Table D.6

Median and standard deviation of occurrence of consonants in final word position in Hindi

Consonants	Median	SD	Consonants	Median	SD
/k/	25.00	5.49	/ʈ/	7.50	3.97
/kʰ/	1.00	2.48	/ʈʰ/	0.00	0.00
/g/	13.00	7.95	/d/	27.50	20.40
/gʰ/	.00	.32	/dʰ/	0.00	0.00
/c/	16.00	8.92	/n/	146.00	116.77
/cʰ/	.00	.23	/j/	2.50	4.22
/ɟ/	9.50	2.97	/r/	95.00	24.09
/ɟʰ/	.00	.38	/l/	26.00	8.18
/ʈ/	10.50	6.98	/v/	0.50	1.66
/ʈʰ/	.50	1.92	/ʃ/	0.50	2.00
/ɖ/	5.50	6.24	/ʂ/	0.00	1.24
/ɖʰ/	.00	0.00	/s/	25.00	16.63
/ɳ/	1.00	1.23	/h/	0.00	1.78
/p/	23.00	5.20	/t/	2.00	6.70
/pʰ/	.00	0.42	/z/	1.00	2.42
/b/	9.00	3.11	/w/	0.00	0.00
/bʰ/	.00	0.51	/f/	11.00	26.96
/m/	40.00	10.50	/ŋ/	1.50	3.43

Table D.7

Median and standard deviation of occurrence of consonants clusters in Hindi

Consonant clusters	Median	SD	Consonant clusters	Median	SD
/ky/	17.50	14.09	/rf/	2.00	1.68
/sk/	13.00	6.17	/ks/	1.50	1.66
/ʂn/	11.50	4.96	/ll/	.50	2.87
/cc/	9.50	9.02	/pl/	1.00	1.88
/rn/	7.50	3.66	/rm/	1.00	1.91
/rʂ/	7.00	2.42	/ɳʈ/	1.00	1.61
/ʂl/	7.00	3.41	/sp/	1.00	1.50
/pk/	5.50	5.01	/hl/	1.00	1.78
/kʂ/	7.00	3.43	/fr/	1.00	1.37
/nk/	5.00	4.68	/hn/	.00	1.29
/lʂ/	6.00	2.73	/gz/	1.00	1.52
/sm/	3.50	3.68	/sn/	.00	2.07
/ʈʈ/	3.50	3.80	/sʈ/	.00	3.78
/st/	2.50	3.51	/nʈ/	1.00	1.61
/pn/	3.00	2.00	/bj/	.00	3.07
/hʂ/	3.00	2.42	/mm/	.00	1.86
/gʂ/	3.00	2.07	/rɖ/	1.00	1.39
/ns/	1.00	5.06	/rv/	.00	.32
/pr/	3.50	2.26	/kr/	1.00	1.16
/rk/	2.50	1.64	/sl/	1.00	1.64
/ln/	2.00	3.05	/khn/	.00	2.02
/mn/	3.00	1.39	/ɳɖ/	1.00	2.80
/rs/	1.50	3.12	/sʂ/	.50	1.10
/nɖ/	1.00	2.80	/dr/	.50	1.10
/kk/	2.00	1.76	/mk/	.50	1.34
/lk/	2.0000	1.52	/ss/	.50	1.07
/rʈ/	1.00	3.69	/gr/	.50	1.13
/kl/	1.00	2.57	/kw/	.00	1.20

Table D.8

Median and standard deviation of occurrence of word shapes in conversational Hindi

Word length	Word shapes	Median	SD
Monosyllables	V	65.00	16.81
	CV	491.00	96.08
	VC	46.50	13.43
	CVC	183.50	59.33
	CCV	17.00	8.07
	CVCC	10.50	10.23
	CCVC	4.50	5.43
	CCV, CVC	1.00	1.92
	CCVCC	1.00	2.03
Disyllables	CV,CV	279.00	81.52
	CVC, CV	96.50	17.79
	CVC, VC	53.00	16.04
	CVC, CVC	25.50	9.03
	CCV, CV	6.00	4.29
	C, CV	17.00	8.07
	VC, CV	45.00	14.91
	VC, VC	9.50	6.43
	VC, VCC	.00	.47
	CV, V	38.00	20.39
Trisyllables	CV,CV,CV	59.00	15.00
	CVV, CV	12.50	6.14
	CV, CV, CVC	5.50	2.79
	CV, CCV, CV	3.50	5.14
	CV, CV, CCV	2.50	2.74
Polysyllables	CV,CV,CV,CV	2.00	1.53
	CV, CV, CV, CVC	.00	.42

Appendix E

Frequency of occurrence of phonemes in Hindi as given by Ghatage (1964)

List of Phonemes Arranged in the Descending Order of Frequency

Phoneme	Frequency	Phoneme	Frequency	Phoneme	Frequency
a	61,333	v	9,589	ṇ	2,537
ā	47,142	d	9,534	ch	1,486
k	36,131	j	7,777	ṅ	1,217
r	34,730	ś	7,719	ph	1,168
e	31,837	g	5,962	ṭh	1,051
n	24,849	ai	5,195	ã	743
t	21,226	b	5,168	gh	722
i	21,099	õ	5,160	jh	634
ī	19,996	c	5,013	ḍh	487
s	17,708	th	4,654	ẽ	431
m	15,958	bh	4,581	ũ	206
h	15,519	au	3,822	ñ	180
p	14,167	dh	3,607	ĩ	137
y	12,209	ū	3,158	ai	130
l	11,816	kh	3,024	au	45
u	11,336	ḍ	2,973	ā	3
o	9,859	ṭ	2,876		

****