Parole: Journal of Linguistics and Education, 12 (2), 2022, 104-111

Available online at http://ejournal.undip.ac.id/index.php/parole

Frequency of Occurrence of Phonemes in Hindi

Sreedevi. N[a*], Irfana. M[b], Anu Rose Paulson[c]

[a,b]*India Institute of Speech and Hearing, Manasagangothri, Mysore, India*

[c]*Department of Speech Pathology and Audiology, Sri Devaraj Urs Academy of Higher Education and Research, Kolar, India*

A B S T R A C T

Hindi, an Indo-Aryan language is the national language of India and the state language of various North Indian states of India such as Madhya Pradesh, Delhi, Uttar Pradesh etc. Statistics on the phonemes of a language provides useful information in the field of speech language pathology, audiology, linguistics and communication engineering. The data can be effectively used for the assessment and selection of target phonemes for treatment of various communication disorders, develop phonetically balanced word lists for audiological testing, and teach foreign language. It also provides valuable information to device text to speech systems and automatic speech recognition systems. The earlier data on frequently occurring phonemes in numerous Indian languages were derived from written sources. However, information from spoken language may be of more significance compared to written language. The aim of the present study was to determine the frequently occurring phonemes in spoken Hindi. Participants were native speakers of Hindi in the age range of 20 to 70 years. Eighteen group conversation samples were recorded. The samples were transcribed using IPA transcription. Systematic Analysis of Language Transcripts (SALT) software was used to analyse the samples in order to obtain the frequently occurring phonemes. Descriptive statistics was applied for the same. Results revealed that phonemes /n, a, e, f, h, k/ were the most frequently occurring phonemesin Hindi. Aspirated phonemes (/gh/, /ʈh/, /ph/, /ɖh/) were the least present phonemes in the data. High and front vowels were more frequently present in spoken Hindi. Considering the manner of articulation, nasals and stops had higher occurrence. Alveolar dominated considering the place of articulation of phonemes. The applications of the study are extensive and can be utilized efficiently in a variety of disciplines.

A R T I C L E I N F O

Paper type
Research Article

Article History
Received : 16/02/2022
Accepted : 27/10/2022

Keywords
- Phonemes
- Hindi
- Manner
- Articulation
- Place Online learning

1. Introduction

Language helps us to communicate effectively through speech by delivering and receiving meaningful messages in a structures and conventional way. It includes both spoken and written and also several non-verbal cues. The world's languages consist of a set of spoken or written symbols with a definite number of phonemes. Each language has its own set of phoneme inventory or phonological system and also has variations with respect to culture, geography etc. These variants of a standard form of language are known as dialects. The standard form of a language, mainly used for official purposes may be different from its dialectal variations. The spoken form of the language may vary.

Hindi is an Indo-Aryan language spoken in various states of India, namely, Madhya Pradesh, Delhi, Uttar Pradesh, Uttarakhand, Bihar, Rajasthan, Chattisgarh, Haryana, Himachal Pradesh and Jharkhand. A variety of

* Email addresses: sreedeviaiish@gmail.com (Sridevi), *fanairfana@gmail.com* (Irfana), *anurosepaulson@gmail.com* (Paulson)

dialects of Hindi are spoken widely across India. It is also spoken by individuals who do not have Hindi as their state language such as Maharashtra, North Eastern states of India etc. It is also the lingua franca of countries such as Fiji (known as Fiji Hindi), Nepal, Bangladesh and Pakistan and a minority language in Mauritius, Surinam, Guyana, South Africa, and Trinidad and Tobago (Meena, 2015). Modern Standard Hindi is the standardized form of Hindi. Khari boli, Haryanvi, Bagheli are few of the dialects of the language. Hindi phonology includes 12 vowels and 38 consonants. Among the vowels, [æ] and [ɒ], are borrowed from English. Consonants [f, z, ʃ] despite being loan phonemes, are well established in Modern Standard Hindi (Ohala, 2004). Hindi has an Akshara system, which uses a combination of alphabetic and syllabic systems (Pandey, 2014).

There are several literature reports on frequency of occurrence of phonemes in various languages. The studies are found as early as 1930s. The earliest study by Whitney (1874) and Dewey (1923) were on English. Dewey (1923) reported 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%) as the frequently present phonemes in English. Delattre (1965) determined frequency of phonemes from connected speech in English. Among the 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%). There are studies in several other languages such as French (Malecot, 1974), Spanish (Sandoval, Toledano, Torre, Garrote & Guirao, 2008), Cantonese, Mandarin, Italian, German and English (Thomas, 2005), Thai (Munthuli, Tantibundhit, Onsuwan, Kosawat & Wutiwiwatchai, 2015) etc. Research in Indian languages was initially by Bhagwat (1961) in Marathi, Ghatage and Madhav (1964) and Khan (1990) in Hindi and Jayaram (1985) in Kannada. The highest occurring phoneme was /k/ in Hindi. Phonemes /k, h, s, m, p, n, ɟ, b, d, w/ and /r, n, k, t̪, s, j, h, l, m/ had the highest occurrence in initial and final positions. Dentals (16.32%) had a higher occurrence followed by velars (12.59%) and labials (9.65%). Chourasia, Samudravijaya, Ingle and Chandwani (2007) reported vowel /a/ to be the most frequent in Hindi. According to De (1973) long vowel /a:/ had highest percentage of occurrence among vowels while /b, s, p/ had frequent occurrence among consonants. Malviya, Mishra and Tiwary (2016) reported phonemes /a: k, r, e, i, n, i:, t̪, s/ to be most commonly present in Hindi corpus. All these studies used written materials as source; newspapers, scripts of dramas, books and dictionaries (Whitney, 1874; Ramakrishna et al., 1957; Ghatage, 1994; Tamaoka & Makioka, 2004). Also spoken materials such as interviews, lectures, radio announcements, telephone conversations etc. were utilized to determine the frequency count of phonemes (French, Carter & 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. The data obtained from these sources have been used to determine frequently occurring phonemes, consonant cluster groups, syllables, syllable type frequency, word frequency (in different contexts, grammatical categories) and morphophonemic categories.

The information is used extensively in areas such of speech language pathology, audiology, linguistics, and speech engineering. In speech language pathology and audiology, the data on frequently occurring phonemes are used to develop various assessment tools (e.g., PB word list) and speech therapy materials (e.g., articulation drill materials). The information can be used by speech engineers in devising speech recognition and text-to-speech systems which are used in Augmentative and Alternative Communication for the rehabilitation of individuals with communication disorders (Cerebral palsy, aphasia etc.). It can also be used effectively to teach a foreign language.

Hindi is a widely used language in India, spoken by over three million people (Kachru, 2006). As discussed earlier, it has several variations as well. Spoken form of a language is different from written form. Studies such as those by Ghatage and Madhav (1964) were from written materials. Moreover, recently with the wide use of English, there are many new modified and borrowed words in the spoken form of any language. Also, there may be differences in the frequency count of phonemes in written and spoken languages. There is limited research on spoken Hindi. Hence, arises the need to create a database of spoken Hindi and gather information on frequently occurring phonemes in conversational Hindi.

2. Methods

Participants: A total of 91 native speakers of Hindi in the age range of 20-to-70 years participated in the study. The participants were exposed to Hindi and use the language in daily conversation. The data was collected from individuals of major Hindi speaking belts - Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. A minimum of 4-5 participants were considered in a group recording and the conversations were recorded for 20 minutes each. From a total of 91 participants, 33 were males and 58 were females.

Instrumentation: Olympus (LS 100) digital recorder was used to record the group conversations. Transcription of the recordings was performed using Toshiba (Satellite C665) laptop and Philips (Shl3095) headphones and Systematic Analysis of Language Transcripts (SALT- Clinical demo version 2012.4.5) was used to carry out the analysis.

Procedure: The selected participants, in groups of 4-5, were asked to sit in a circle and the digital recorder was placed at the centre, equidistant from each of the participant. As there was no specific topic provided for conversation, the participants were encouraged to speak freely on any topic of their interest. Also, conversations had to be carried out as naturally as possible in Hindi only, sometimes using loan words from English when necessary. Totally 18 spoken Hindi recordings were carried out.

Data analysis: Conversation samples were transcribed with the help of International Phonetic Alphabet (IPA) by Ohala (1994) for Hindi. 10% of each recording sample was selected, transcribed and analysed for testing both inter and intra judge reliability measures. Cronbach alpha index of 0.87 and 0.90 were obtained for inter-judge and intra- judge reliability respectively.

Statistical analysis: Mean percentage of occurrence of various phonemes was determined by employing descriptive statistics. Subsequently, Wilcoxon's sign language test was performed to establish pair-wise significance.

3. Results and Discussion

The study aimed at identifying the frequency of occurrence of phonemes in conversational Hindi from major Hindi speaking states such as Madhya Pradesh, Delhi, Chhattisgarh, Jharkhand, Uttar Pradesh, and Uttarakhand. There was a total of 1,48,862 phonemes in the corpus from 18 recordings and the number of total phonemes recorded across ach recording varied from 6000 to 12000 phonemes. Figure 1 provides information about the total phonemes recorded in each recording session. The total corpus included consonants, vowels and diphthong. The mean percentage of vowels (54.42%) was higher than consonants (44.50%) which are depicted in figure 2. 1.08% of the total phonemes accounted for the diphthongs.

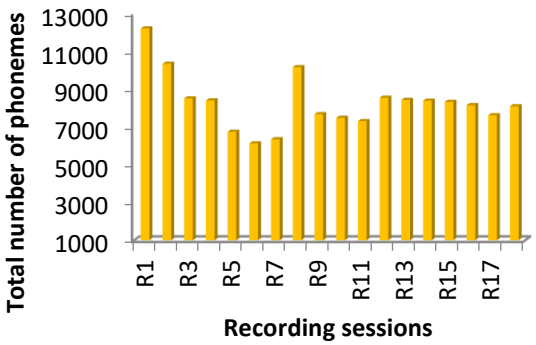

Figure 1. Total number of phonemes in each recording session

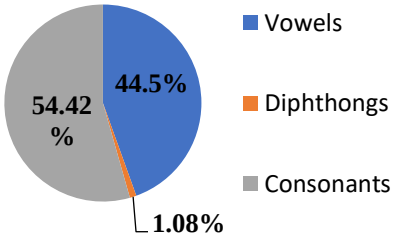

Figure 2. Total number of phonemes in each recording session

Similar results were obtained in spoken English (Denes, 1959 &Delattre, 1965), American English, spoken Cantonese, Mandarin and Italian (Thomas, 2005) and written English (2011). The study is also in consonance with several Indian languages- Ghatage & Madhav (1964) in Hindi, Pandit (1965) in written and spoken Gujarati Ranganatha (1982), Jayaram (1985) and Sreedevi et al, (2012) in Kannada, Vasanthakumari (1989) in Tamil, Kumar & Mohanty (2012) in Telugu and Sreedevi & Irfana (2013) in Malayalam.

In the present study, in conversational Hindi, vowel /i/ had the highest occurrence of 19.43% followed by consonant /n/ (15.99%), vowels /a/ (9.17%) and /e/ (6.12%) and consonants /f/ (5.22%), /h/ (3.99%) and /k/ (3.80%). Table 1 depicts the mean percentage of consonants obtained in conversational Hindi. The results of the study however are contradictory to earlier studies in written Hindi. Chourasia, Samudravijaya, Ingle and Chandwani (2007) reported vowel /a/ and consonant /r/ as frequently occurring phonemes in Hindi. Phonemes /a:/, /a/, /e/, /i:/, /i/, /b/, /s/, /p/, /m/ and /k/ were the highly occurring phonemes as determined by De (1973). The least frequently occurring phonemes in Hindi were the aspirated consonants /ʈʰ/, /t̪ʰ/, /gʰ/, /bʰ/, /ɾʰ/ and /ɖʰ/. This is in consonance with previous studies in Hindi and also with various other Indian languages.

Table 3: *Mean percentage of occurrence of phonemes in spoken Hindi in descending order obtained in the present study*

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	/ɾ/	0.87	/t/	0.13
/f/	5.22	/v/	0.87	/ã:/	0.13
/h/	3.99	/ʈ/	0.64	/cʰ/	0.12
/k/	3.80	/ã/	0.64	/ʈʰ/	0.08
/m/	3.11	/ɖ/	0.56	/ɳ/	0.06
/r/	3.02	/pʰ/	0.44	/ŋ/	0.04
/o/	2.94	/kʰ/	0.36	/t̪ʰ/	0.04
/s/	2.33	/i:/	0.31	/gʰ/	0.04
/p/	2.29	/ə/	0.30	/bʰ/	0.03
/a:/	2.11	/ũ:/	0.30	/õ/	0.03
/l/	2.01	/dʰ/	0.27	/ĩ:/	0.03
/d/	1.71	/u:/	0.22	/ɾʰ/	0.02
/j/	1.52	/i:/	0.22	/ʂ/	0.02
/t̪/	1.37	/au/	0.18	/w/	0.01
/g/	1.23	/ẽ/	0.18	/ɖʰ/	0.01
/c/	1.16	/ʃ/	0.18		

Short vowels occurred more frequently than long vowels (figure 3) as observed in other Indian languages as well Sreedevi, Smitha & Vikas (2012) in Kannada; Sreedevi & Irfana (2013) in Malayalam.

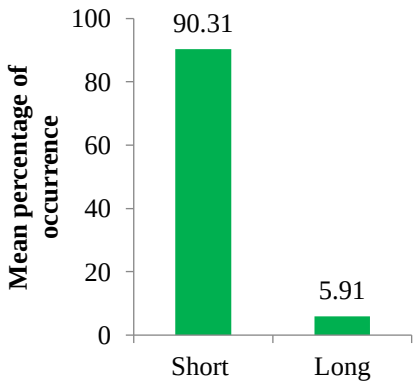

Figure 3: Types of vowels

Among the vowels, high vowel /i/ had most occurrence followed by vowels /a/, /e/ and /o/ (Figure 4). Vowel /i/ showed highest occurrence in English as well (Dewey, 1923; Voelker, 1935; Tobias, 1959). On the other hand, reports by Ghatage and Madhav (1964), Chourasia, Samudravijaya, Ingle and Chandwani (2007), De (1973), Khan (1990) and Malviya, Mishra and Tiwary (2016) reported vowel /a/ to be the most predominant vowel in Hindi. Other Dravidian and Indo-Aryan languages such as Marathi (Berkson & Nelson, 2015), 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) also had vowel /a/ as the highly occurring vowel.

Front vowels had higher occurrencethan central and back vowels (Figure 5). Literature reports English (Denes, 1959; Thomas, 2005), Cantonese, Mandarin, German, Italian (Thomas, 2005) and Telugu (Kalyani&Sunitha, 2009) to have high occurrence of front vowels.

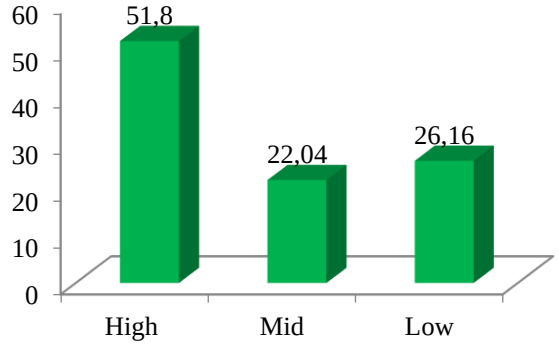

Figure 4. Mean percentage of occurrence of high, mid and low vowels

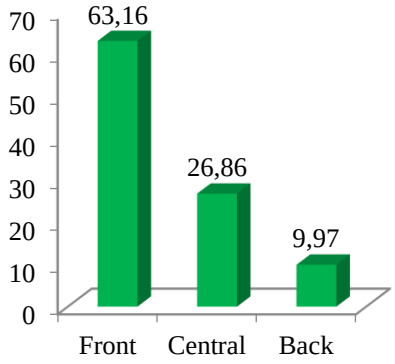

Figure 5. Mean percentage of occurrence front, central and back vowels

Nasal vowel /ã/ (1.42%) had relatively higher occurrence in the present study. De (1973) also reported vowels /õ/ and /ã/ as the most frequently occurring nasal vowels in Hindi. Among the diphthongs, /ai/ showed more occurrences.

The frequently occurring consonants in Hindi were /n/, /f/ , /h/ , /k/ , /m/ and /r/. Similar to Hindi, Kannada (Nayaka, 1967; Sreedevi, Smitha and Vikas, 2012), English (Voelker, 1935; Mader, 1954; Crystal, 1981; Thomas, 2005), Swedish (Sigurd, 1968), German (Thomas, 2005) and Italian (Thomas, 2005) had nasal phoneme /n/ as the highly occurring consonant. Earlier studies in Hindi are conflicting to the results obtained in the present study. Khan (1990) and Malviya, Mishra and Tiwary (2016) reported phoneme /k/ to be the highly occurring consonant in Hindi. The aspirated consonants had least occurrence which was a similar result obtained in many other Indian languages such as Kannada (Nayaka, 1967; Sreedevi, Smitha and Vikas, 2012), Malayalam (Sreedevi & Irfana, 2013) etc.

Considering manner of articulation, nasals were predominant followed by stops and fricatives in the present study. Phoneme /n/ occurred highest among nasals, phonemes /f/ and /k/ among fricatives and stops. The study has results similar to other languages such as Malayalam, Kannada, Tamil, Telugu (Ramakrshna, Nair, Chiplunkar, Atal & Rajaraman, 1957) and Cantonese and Mandarin (Thomas, 2005) except for the occurrence of stops. Occurrence of stops was higher in many Indian languages (Jayaram, 1985; Kalyani & Sunitha, 2009) and non- Indian languages (Denes, 1957; Guirao & Jurado, 1990; Thomas, 2005). Figure 6 depicts the percentage of occurrence of consonants based on manner of articulation. Application of Friedman test revealed significant difference across the categories. A pair-wise test of the same revealed all the pairs to have significant difference except fricatives and stops ($|Z|= 1.786$; $p= 0.074$).

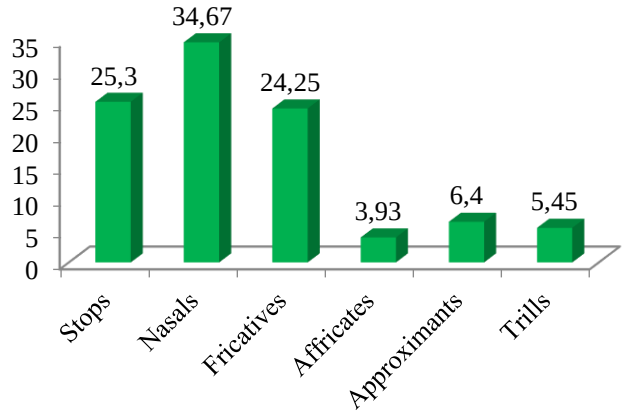

Figure 6. Mean percentage of occurrence of consonants based on manner of articulation

Considering place of articulation, alveolars occurred maximally, while retroflex had least occurrence. 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 velars, phoneme /k/ had higher occurrence. Unlike Malayalam (Sreedevi & Irfana, 2013), Telugu (Kalyani & Sunitha, 2009) and Marathi (Berkson & Nelson, 2015), Hindi had relatively higher occurrence of glottal sound /h/. As in Hindi, Telugu (Kalyani & Sunitha, 2009; Kumar & Mahanty, 2012), Cantonese, Mandarin, Italian, German and American English (Thomas, 2005) had higher occurrence of alveolars. Dentals were more frequent in Malayalam and Kannada. However, Khan (1990) reported dentals to be frequent in Hindi than alveolar. Figure 7 illustrates the mean percentage of occurrence of consonants based on place of articulation. Friedman test revealed statistical significance among varies types of consonants. Pair wise comparison of the same revealed palatals- dentals (|Z|= 2.345; p= 0.019), bilabials- labiodentals (|Z|= 1.024; p= 0.306) and glottal-palatals (|Z|= 1.847; p= 0.065) did not have a significant difference which indicates these categories had a similar percentage of occurrences in Hindi.

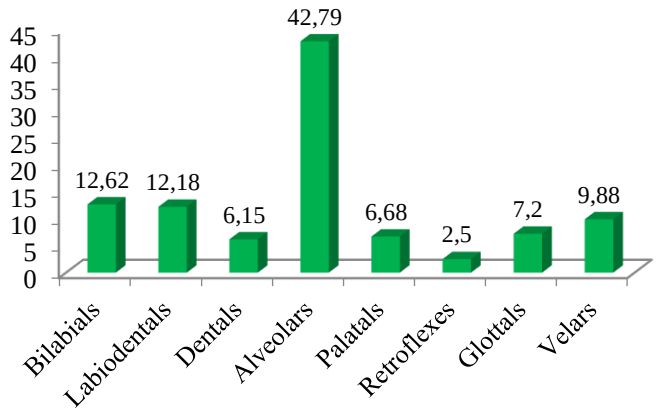

Figure 7. Mean percentage of occurrence of consonants based on place of articulation

4. Conclusions

To conclude, consonants had higher occurrence than vowels. Vowel /i/ and consonant /n/ had maximum occurrences among vowels and consonants respectively. The frequency count of diphthongs and aspirated consonants were the least in the data. Nasals occurred more frequently considering the manner of articulation while stops and fricatives were more frequent considering place of articulation. Unlike other Indian languages

like Malayalam, Kannada and Telugu, Hindi had a higher occurrence of glottal stop /h/. Consonants were predominant in the initial position and least in the final position. The results of the current study will enable audiologists and speech language pathologists in developing assessment (PB word lists) and assessment and intervention (speech sound targets for articulation therapy) tools for the rehabilitation of individuals with communication disorders. The information is paramount to speech engineers and linguists as well. 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.

References

Berkson, K. H., & Nelson, M. (2015). 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.

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

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

Ghatage, A. M. &Madhav. A. (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.

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

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*.

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

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.

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, WaismanCenter. *Language Analysis Laboratory*.

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*.

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

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

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.

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.

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).

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.

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)*.

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