Original Article

Spoken Word Frequency in the Hindi Language: A Preliminary Database for Psycholinguistic Studies

Himanshu Verma, Gouri Shanker Patil[1]

Department of Otolaryngology, Post Graduate Institute of Medical Education and Research, Chandigarh, [1]Department of Speech-Language Pathology, Ali Yavar Jung National Institute of Speech and Hearing Disability, Secunderabad, Telangana, India

Abstract

Objective: Limited studies related to spoken word corpus in the Indian context are available in the literature. To fulfill the demands of the spoken word frequency database in Hindi for advance psycholinguistic and cognitive studies, we tried to establish the preliminary spoken word database of Hindi language for children studying in Grade VI to Grade IX. **Methods:** To create the spoken word corpus a recorder was given to subjects to record their conversation. The recorded sample was transcribed into Hindi text using voice note II software. The transcribed sample was uploaded into Text Analyzer software, and word frequency, the number of syllables, and lexical density were computed. **Results:** Spoken word corpus consists of a total of 49,476 words. Lexical density was higher for females than males because the female database contains more unique words. The study also revealed that subjects used functional words and verbs more frequently, followed by nouns. **Conclusion:** We can conclude that the current database provides information about the high-frequency and low-frequency words used by children studying in Grade VI to Grade IX. This database will be helpful in psycholinguistic and cognitive experiments; however, the present corpus included data from the middle socioeconomic group and contained fewer words. The present study is the preliminary study future study demands and requires an extensive word database.

Keywords: Corpus linguistic, mental lexicon, spoken word database

Date of Submission: 23-08-2020 **Date of Revision:** 09-03-2021 **Date of Acceptance:** 10-05-2021 **Date of Web Publication:** 24-12-2021

Introduction

A language is a code whereby ideas about the world are expressed through a conventional system of arbitrary signals for communication. Languages are described qualitatively in terms of grammatical units such as nouns, verbs, noun phrases, verb phrases, subject, object, agent, and goal to explain the structure of the language. With new advancements and ideas in the field of linguistics, many researchers study language quantitatively. The quantitative study is conducted using a new method of language study called corpus linguistics that has emerged in recent years. Corpus is a large collection of written or spoken texts available in machine-readable form accumulated in a scientific way to represent a particular variety or use of a language.[1] The size, text type, organization, accessing method, etc., are some of the basic features of a corpus that have to be carefully decided while generating a corpus. Word frequency is one of the most studied domains in the field of corpus linguistics.

The study of word frequencies (i.e. the question of how often particular words or word forms occur in a given text or corpus of texts) is one of the favorite and most traditional issues in the history of quantifying approaches to language. It is no intrinsic property because it cannot be measured directly on the word using some operational definitions. It can be determined by counting occurrences of the word in a finite specimen of text. The counting can be performed mechanically, and the results of counting can be used in typography, stenography, psychology, psychiatry, language teaching, cryptography, software production, etc. Word counts, which are a necessary precondition for any theoretical study of lexical frequency behavior, became popular quite early, and in any case, long before quantitative linguistics became established as a discipline in its own right.

Address for correspondence: Mr. Himanshu Verma, Department of Otolaryngology, Post Graduate Institute of Medical Education and Research, Chandigarh, India.
E-mail: himanshu.v. 91@gmail.com

Access this article online

Quick Response Code:

Website: www.jisha.org

DOI: 10.4103/jisha.JISHA_24_20

How to cite this article: Verma H, Patil GS. Spoken word frequency in the Hindi language: A preliminary database for psycholinguistic studies. J Indian Speech, Language Hearing Assoc 2021;35:27-32.

Word frequency and the mental lexicon

Word frequency plays an essential role in our mental access to lexical information. Word frequency refers to "how often the word occurs in normal use of the language."[2] Naturally, some words occur more often than other words in our daily conversation or certain situations. Carroll (1938)[3] regards word frequency as one of the significant factors which influence the process of accessing or retrieving lexical information from memory. Studies have demonstrated that phoneme recognition is accelerated using high-frequency words compared to low-frequency words,[4] as are visual word recognition tasks.[5,6] High-frequency words tend to be recognized more accurately and accessed faster in our mental lexicon.

Word frequency plays an active role in lexicon acquisition. Even children with language impairment learn verbs more efficiently if they are presented frequently and in an appropriate time spacing.[7] Brown[8] pointed out that words frequently used in speech to children tend to match the children's cognitive predilections. Moreover, studies have shown that when people of different ages are asked to write definitions of words, word frequency shows a strong influence on their definitions of adjectives[9] and nouns and verbs.[10] Marinellie *et al*.)[10] revealed that subjects were more familiar with high-frequency nouns than low-frequency nouns. They further found that the gap between the high and low-frequency nouns familiarity becomes smaller with age. This result suggests that the mental lexicon progresses and organizes high-frequency and low-frequency words differently. Specifically, research has demonstrated that high-frequency words were responded to more quickly and more accurately than low-frequency words at lexical decision tasks using the dual-task paradigm.[11] High-frequency words are important and valuable because they "cover a large proportion of the running words in spoken and written texts and occur in all kinds of uses of the language."[12]

High-frequency words facilitate target recognition in lexical decision tasks, whereas the opposite is observed for low-frequency words.[13-15] This effect was observed for the reaction times (RTs) in both the English Lexicon Project[16] and the British Lexicon Project.[17] Furthermore, the same pattern of faster RTs for high-frequency words was observed by Duyck *et al*.[18] for Dutch–English bilinguals when recognizing words in their second language. Other differential effects include observations that low-frequency words produce more phonological errors in speech than do high-frequency words.[19] Low-frequency words are recognized better in recognition memory experiments than high-frequency words (known as the mirror effect),[20,21] and that pictures are named faster when they correspond to high rather than low-frequency words.[22]

Need of the study

India is one of the fastest-growing countries in the world. With recent advances, many psycholinguistic and cognitive experiments are taking place, which will help in exploring the many unrevealed facts about language processing. Many psycholinguistic experiments need the word database to prove the hypothesis of these experiments, but due to the nonavailability of Indian language databases, these experiments[23] had many limitations, such as the choice of words they used in the experiment is not standardized and hence may affect the results.

India being a multicultural and multilinguistic background, there is a dearth of studies about the frequency of words in Indian languages. In India majority of the population (i.e. 528,347,193) speaks the Hindi language.[24] Majorly Hindi is speaking by Northern, Central, West, and some regions of East India.

As Hindi is spoken in a different region of India and some other countries (e.g. Fiji), it has many dialectal and vocabulary variations. Hindi belongs to the Indo-Aryan family of languages and has a Devanāgari script. It contains 13 vowels, 33 consonants, and additional diacritical signs. The script has syllabic as well as alphabetic properties. However, unlike most alphabetic scripts in which consonants typically stand-alone as phonemes, consonants in Hindi have an inherently associated vowel, and Hindi resembles a syllabary.[25] As the Hindi language differs from other foreign languages in many aspects, we need to establish a database for Hindi and other Indian languages.

The center of Indian language technology,[26] IIT Bombay,[27] Annamalai University,[28] and Dr. Babasaheb Ambedkar Marathawada University[29] developed the written database for Kannada, Hindi, Tamil, and Marathi language, respectively. Indian Institute of Informational Technology, Hyderabad, and four other universities (i.e. University of Washington, Columbia University, University of Colorado Boulder, and the University of Massachusetts at Amherst) developed a Hindi-Urdu Treebank database.[30] However, there is still not an adequate database in Hindi spoken language. There are some preliminary[27] and advanced studies[30] in the Hindi written frequency database, but there is a lack of a spoken frequency database. As many studies are available on the written frequency database, so many evidence are available on the influence of word familiarity on word recognition. However, the use of spoken word frequency counts is conspicuously absent or limited in the literature due to a lack of appropriate frequency counts. Tillmann[31] beautifully described the differences between the written and spoken databases. He stated that a written database is different from a spoken database, and both have their significance. Studies[31,32] suggested that a spoken database is more natural and provides better insight into human language use and processing. Hence, there is a need to develop a spoken-word database.

In the present study, we targeted the specific age group (i.e. 11–14 years) to develop the Hindi database, as this is the transitional phase for higher language development. According to Piaget's theory,[33] during this age, children transit from the concrete operational stage to the formal operational stage. During this age, there will be a growth of the white matter in the language-related areas of the brain,[34,35] and the language development during this age range is connected with

the development of other nonlinguistic abilities (i.e. social skills, attention, memory, etc.).[36] During the selected age, children start to think beyond the literal interpretations and will comprehend figurative language, which is the essence of language.[34] Further, they were able to handle the voice recorder carefully with all instructions. Here, we tried to develop a database across age groups as many studies in past literature suggested that males and females process and use language differently.[37,38] We also formulated the database across grammatical categories as this will help to know what grammatical category is mostly used as described by Li and Fang.[39]

The present study would provide a database for developing speech materials for assessment and selecting treatment targets for various communication disorders. As well as these spoken words database have a significant role in the psycholinguistic, cognitive, applied psychology, applied educational researches, so there is a dearth of Indian study to establish the database for spoken words.

Aim and objectives

To establish the database for the frequency of occurrence of individual words occurs in an everyday situation in the Hindi language for Grade VI-Grade IX studying children.

1. To establish the database with respect to the gender-wise distribution
2. To create the database according to different grammatical categories.

Methods

Ethical approval

Before conducting the research and collecting the data, ethical approval was taken from the ethical committee of the Institute.

Participants

A total of five Hindi medium schools of Delhi were approached to participate in the study. Among five schools, only three schools showed interest and permitted data collection. Grade VI to Grade IX students of Hindi native speakers were taken in the present study. A total of 15 students from the age range of 11 years to 14 years (mean age: 13.2 years; standard deviation [SD]: 0.77) participated in the study. Among 15 subjects, 7 were girls (mean age: 13.42 years; SD: 0.78), and 8 were male (mean age: 13; SD: 0.75), as shown in Figure 1. As shown in Table 1, among 15 subjects, five students attended Grade IX, six students were studying in Grade VIII, three students attended grade VII, and only one student belongs to Grade VI.

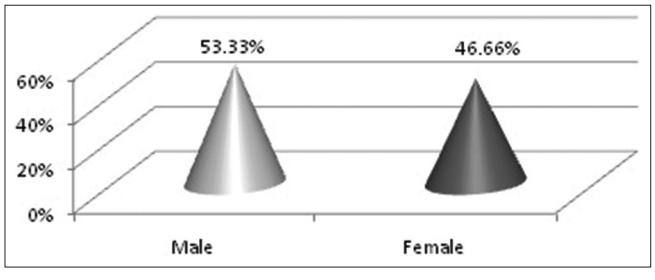

Figure 1: Graphical representation of the gender-wise distribution

All participants had Hindi as a first language with a Delhi region accent and belong to middle socioeconomic status. Subjects with a history of any neurological, psychological, speech and language problems, sensory issues were excluded from the present study.

Tools used

A digital recorder (SONY ICDUX560F) was used for recording the conversation sample.

Voice Note II tool[40] was used to convert speech samples into written text.

Text Analyzer[41] was used to calculate the frequency of words. It is an online computer utility tool. This tool is freely available online for students. This software supports NonEnglish languages like Hindi. This software helps to compute the most frequent phrases and frequencies of words. This software also counts the number of words, characters, sentences, syllables, and lexical density.

Spoken word corpus formation

Before collecting data, the aim and purpose of the study were explained to the school administration, guardians, and subjects, and written consent was taken from all the participants. To create a spoken word corpus, a digital voice recorder was given to subjects for 2 days and asks to record their voice during school hours. Subjects were instructed that don't reveal about the voice recorder to other friends and not to give any special speeches or read into the recorder. After 2 days of recording, the voice recorders were collected from the participants, and the speech samples were stored in password-protected computer. From the recorded samples, only the participants' speeches were extracted and converted into Hindi text using speech-to-text converter software (i.e. Voice Note II). During transcription, commonly used English words (e.g. ticket, bus,

Table 1: Demographic detail of subjects

SUBJECTS	GENDER	AGE	GRADE
1	FEMALE	14 yrs	IX
2	FEMALE	14 yrs	VIII
3	FEMALE	14 yrs	IX
4	FEMALE	13 yrs	VIII
5	FEMALE	13 yrs	IX
6	FEMALE	14 yrs	IX
7	FEMALE	12 yrs	VI
8	MALE	13 yrs	VII
9	MALE	14 yrs	VIII
10	MALE	13 yrs	VIII
11	MALE	12 yrs	VII
12	MALE	12 yrs	VII
13	MALE	13 yrs	VIII
14	MALE	14 yrs	IX
15	MALE	13 yrs	VIII

pen, etc.) were considered during the analysis. The recorded samples also contain nonverbal sounds made by participants and songs which were excluded from the database. After converting the sample using software, 10% of the sample was again retested manually using the trained ear and a respected correction was made. The Cronbach alpha test was used, and a reliability index of 0.81 was obtained.

Procedure

The transcribed file was uploaded in the computer software (i.e. Text Analyzer), and the frequency count was computed. The frequency count for grammatical categories was done manually. Inter-judgment and intra-judgment were done to check the reliability of the frequency count of grammatical categories and resolve the noun-verb and adjective-noun ambiguities.

Data analysis

The raw data obtained were analyzed using the software Text Analyzer version II. The frequency count was computed according to the gender and grammatical categories. The frequency count of the database was also computed according to the different grammatical categories (i.e. noun, verb, pronoun, adverbs, and adjectives). The lexical density was compared between the genders using an independent two-sample *t*-test on SPSS 20 version (Chicago, Illinois, USA) SPSS 20 version (Chicago, Illinois, USA) software. The confidence level was set at 0.05.

Inter-judge and intra- judge reliability

The sample of each conversation recording was subjected to inter-judge and intra-judge reliability measures. Three judges, including two experienced speech-language pathologists and a clinical linguist, served as judges for determining inter-judge reliability measures. For inter-judge reliability, 10% sample of the recording of each sample was transcribed by each of the three judges, and grammatical categories frequency count was also done. The recorded samples were played to the judges independently. They were not allowed to discuss the transcription of the sample before or after the task. For intra-judge reliability, 10% of each of the recordings were transcribed and analyzed by the investigator after transcribing all the samples completely. The statistical procedure, Cronbach alpha test, was used to assess the reliability index. The reliability index (alpha) of 0.87 was obtained for inter-judge, and 0.92 was obtained for intra-judge reliability for transcription, whereas 0.85 was obtained for inter-judge and 0.94 was obtained for intra-judge reliability for frequency count of grammatical categories.

Results and Discussion

Spoken word corpus

A total of 49,476 word counts accounted for the data collection from conversation samples of fifteen participants. The mean and SD of the frequency of occurrence of individual words were calculated as shown in Table 2. The mean spoken word frequency is 3298.467, and the SD is 282.40. The lexical density of the database is 6.3243. Lexical density is simply defined as the number of lexical words divided by the total number of words.[42] In short lexical density is a measure of how much informative a text is. Ure[43] reported that spoken text tends to have a lower lexical density than written ones, the same results also reported by Husson-Ettle *et al.*[44]

As shown in Figure 2, subject 8 had the highest number of words (3969) in his spoken sample; this is followed by subject nine and subject 14 with 3561 words. Subject 1 had the least number of words in the recorded sample. These individual differences were present due to the presence of a conscious level due to the presence of a voice recorder.

The most frequent words lie at the beginning of 20% of data. These words are the most frequently used words, whereas other words lie at the back of lemma storage. Present study result supported by Terzopoulos *et al*.[45] A study was done by Terzopoulos *et al*.[45] formulated the written word database of the Greek language. This database consists of 68, 692 words and reported that 20% of high-frequency words contain the words that majorly formed an individual's pre-dominant vocabulary. Same results were also reported in the literature by many studies such as Savrtvik;[46] Brown;[47] Coltheart[48] The above result may indicate that these words may lie at the beginning of the mental lexicon templates. The present study results further support the speech perception and speech production theories. The results of current studies support the Logogen model, template model, and other speech perception models. These models reported that very frequently used words lie in the templates located in the front of lemmas.

Formation of database based on gender-wise distribution

The database was also formed based on gender-wise distribution. The female recorded conversation sample consists of 21,357 words with a mean of 3051.14 and an SD of 48.85. In contrast, the male recorded sample consists of 28,119 words with a mean of 3514.88 and an SD of 206.79, as shown in Figure 3 and Table 3.

Female conversation samples consist of more words than male conversation samples. Females tend to use more whisper talk

Table 2: Mean and Standard deviation of spoken word corpus

Total no. of words	Mean of spoken words	Standard Deviation	Lexical Density
49,476	3298.467	282.40	6.324

Table 3: Mean and Standard deviation of gender-wise formed corpus

Gender	Total no. of words	Mean of spoken words	S.D. of spoken words	Lexical density
Male	28,119	3514.88	206.79	7.88
Female	21358	3051.14	48.85	9.21

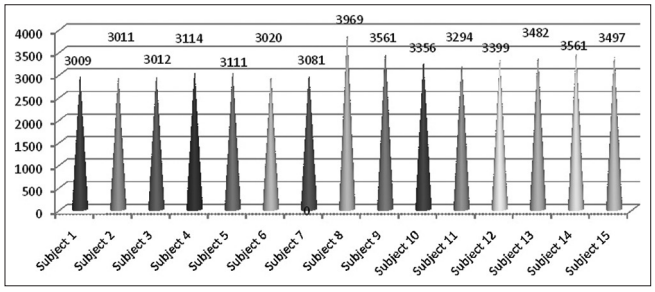

Figure 2: Graphical representation of a word spoken by each participant

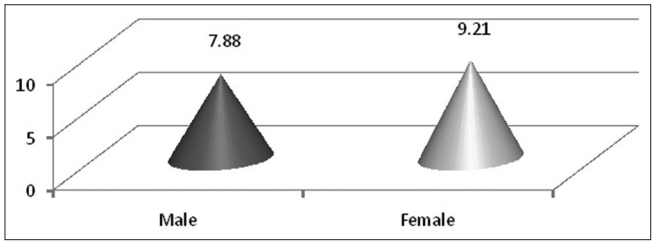

Figure 4: Graphical representation of lexical density across gender

because most of the data was not recorded adequately and had more distortions. Thomson *et al*.[37] reported the same statement in their study that females use more whisper talk than males. Mulac *et al*.[38] studied 96 school children taken from 4th, 8th, and 12th grades and reported that men used more words overall than females. They also reported that men took more turns in conversation and had more circumlocution compare to females. The lexical density of the male database (7.88) was lesser than a female database (9.21), and a significant difference ($P = 0.005$) was found on the *t*-test, as shown in Figure 4. The same results were also reported by Berman and Verhoeven.[49] A study was done by Johansson[50] gave further support to the result of the current study. He computed the children's lexical density according to the age-wise and gender-wise distribution and reported that females had more lexical density than males. The author further reported that lexical density increase as age increases.

The result further provides an idea about the use of language among both genders. As a result, indicate that females asked more question compared to the males. This result is further supported by the study done by Xia,[51] Mehl and Penebaker,[52] and Mulac *et al*.[38]

Formation of a database based on different grammatical categories

As shown in Figure 5, the database contained a higher number of verbs followed by nouns, pronouns, adjectives, and adverbs.

Li and Fang[39] studied the word frequency count of the English language for different grammatical categories from the age range of 1.8 years to 2 years. They reported that children used more Nouns (3080) which was followed by verbs (682), whereas children least used pronouns.[52] The sequence of different grammatical categories was as follows: Noun > Verbs > Adjectives > Adverb > Pronoun, whereas the present study reveals that children studying from grade VI

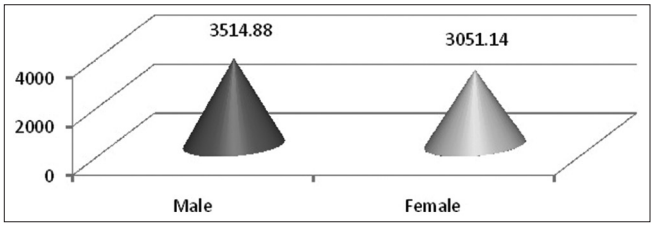

Figure 3: Mean of the frequency of occurrence of words in both the gender

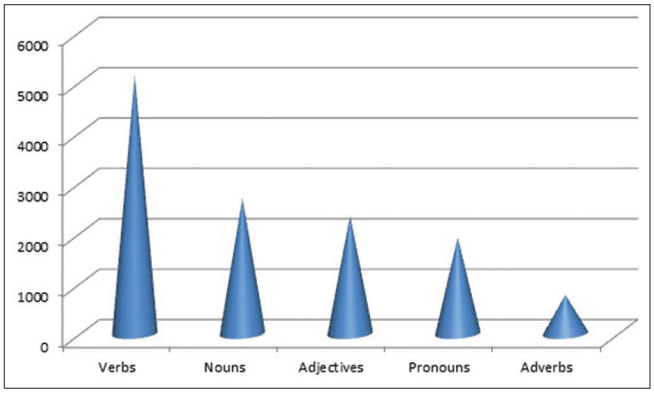

Figure 5: Graphical representation of word frequency based on grammatical categories

to grade IX used verbs more frequently than nouns and least frequently used adverbs. This difference may occur due to the age difference as both studies have taken different age groups. The other hypothesis we can form is that maybe it is a property of the Hindi language; however, we could not find any study in support of the same.

Results of the present study go hand in hand with a study done by Horovits and Newman.[53] They studied the 5-min spoken samples of High school children and reported that subjects used verbs more frequently than other grammatical categories, followed by nouns.

Summary and Conclusion

The current database provides information about the high frequency and low-frequency words used by children studying in Grade VI to grade IX. A spoken word database was also established based on gender and different grammatical categories. This database provides evidence in support of many speech perception and speech production theories and models. Gender-wise creation of database provides the fact that there is a difference in the vocabulary used by males and females. This database will be helpful in psycholinguistic and cognitive experiments; however, the present corpus included data from the middle socioeconomic group and contained fewer words. The present study is the preliminary database future study demands and requires a large word database.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Tognini Bonelli E. Corpus Linguistics at Work. Amsterdam: John Benjamins; 2001.
2. Nation P, Waring R. Vocabulary size, text coverage and word lists. In: Schmitt N, McCarthy M, editors. Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press; 1987. p. 6-19.
3. Carroll JB. Diversity of vocabulary and the harmonic series law of word-frequency distribution. Psychol Record 1938;2:379-86.
4. Foss DJ. Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. J Verbal Learning Verbal Behav 1969;8:457-62.
5. Rubenstein H, Garfield L, Millikan JA. Homographic entries in the internal lexicon. J Verbal Learning Verbal Behav 1970;9:487-94.
6. Whaley CP. Word-nonword classification time. J Verbal Learning Verbal Behav 1978;17:143-54.
7. Riches NG, Tomasello M, Conti-Ramsden G. Verb learning in children with SLI: Frequency and spacing effects. J Speech Lang Hear Res 2005;48:1397-411.
8. Brown R. Linguistic determinism and parts of speech. J Abnorm Soc Psychol 1957;55:1-5.
9. Marinellie SA, Johnson CJ. Adjective definitions and the influence of word frequency. J Speech Lang Hear Res 2003;46:1061-76.
10. Marinellie SA, Chan YL. The effect of word frequency on noun and verb definitions: A developmental study. J Speech Lang Hear Res 2006;49:1001-21.
11. Becker CA. Allocation of attention during visual word recognition. J Exp Psychol Hum Percept Perform 1976;2:556-66.
12. Nation P. Learning Vocabulary in another Language. Cambridge: Cambridge University Press; 2001.
13. Mason JM. The role of orthographic, phonological and word frequency variables on word-non word decision. Am Educ Res J 1976;13:199-206.
14. Monsell S. The nature and locus of word frequency effects in reading. In D. Bresner & G. W. Humphreys (Eds.), Basic processes in reading: Visual word recognition. New York: Taylor & Francis Group 1991: pp. 148-97.
15. Van Heuven WJ, Mandera P, Keuleers E, Brysbaert M. Subtlex-pl: Subtitle-based word frequency estimates for Polish. Behav Res Methods 2014;47:471-83.
16. Balota DA, Yap MJ, Hutchison KA, Cortese M, Kessler B, Loftis B, *et al.* The English lexicon project. Behav Res Methods 2007;39:445-59.
17. Keuleers E, Brysbaert M, New B. SUBTLEX–NL: A new measure for Dutch word frequency based on film subtitles. Behav Res Methods 2010;42:643-50.
18. Duyck W, Vanderelst D, Desmet T, Hartsuiker RJ. The frequency effect in second-language visual word recognition. Psychon Bull Rev 2008;15:850-5.
19. Stemberger JP, Mac-Whinney B. Frequency and the lexical storage of regularly inflected forms. Mem Cogn 1986;14:17-26.
20. Shepard RN. Recognition memory for words, sentences, and pictures. Journal of Verbal Learning and Verbal Behavior, 1967;6:156-63.
21. Steyvers M, Malmberg KJ. The effect of normative context variability on recognition memory. J Exp Psychol Learn Mem Cognit 2003;29:760-6.
22. Jescheniak JD, Levelt WJ. Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. J Exp Psychol Learn Mem Cognit 1994;20:824-43.
23. Solanki P, Sinha A, Verma H. Substituted- letter and transposed- letter effect in a masked priming paradigm in Hindi developing readers with and without dyslexia. Indian Linguist 2020;81:117-27.
24. Census of India; 2011. Available from: https://censusindia.gov.in/2011census/C-16.html. [Last accessed on 2021 Mar 10].
25. Vaid J, Gupta A. Exploring word recognition in a semi-alphabetic script: The case of Devanagari. Brain Lang 2002;81:679-90.
26. Sahoo K, Vidyasagar V. Kannada WordNet – A Lexical Database. Paper Presented in IEEE Xplore; 2003. Available from: https://www.researchgate.net/publication/4058884. [Last accessed 2020 Jun 25].
27. IIT Bombay; 2003. Available from: http://www.cfilt.iitb.ac.in/wordnet/webhwn/. [Last accessed 2020 Jun 25].
28. Ganesan. Tamil Corpus Generation and Text Analysis; 2003. Available from: http://www.infitt.org/ti2009/papers/ganesan_m_final.pdf. [Last accessed on 2020 Jun 08].
29. Shrishirmal PP, Deshmukh RR, Waghmare V, Borade N, Janse PV, Janvale GB. Development of Marathi Language Speech Database from Marathwada Region. Shangai China: Paper presented in 18th Oriental COCOSDA; 2015.
30. The Hindi-Urdu Treebank Project; 2012. Available from: http://verbs.colorado.edu/hindiurdu/. [Last accessed on 2021 Mar 14].
31. Tillmann GH. Eight main differences between collections of written and spoken language data. FIPKM 1997;35:139-43.
32. Redeker G. On differences spoken and written languages. Discourse Process 1984;7:43-55.
33. Piaget J, Boring EG, Werner H, Langfeld HS, Yerkes RM, editors. "Jean Piaget.", A History of Psychology in Autobiography. Vol. IV. Worcester: Clark University Press; 1957. p. 237-56.
34. National Institute of Environmental Health Sciences; 2020. Available from: https://kids.niehs.nih.gov/. [Last accessed on 2021 Mar 10].
35. Rosselli M, Ardila A, Matute E, Velez-Uribe I. Language development across the life span: A neuropsychological/neuroimaging perspective. Hindwai Neurosci J 2014;2014:21.
36. Gibson KR, Petersen AC. Brain Maturation and Cognitive Development. Oxford, UK: Blackwell; 2010.
37. Thomson R, Murachver T. Predicting gender from electronic discourse. Br J Soc Psychol 2001;40:193-208.
38. Mulac A, Studley LB, Blau S. The gender-linked effect in primary and secondary students' impromptu essays. Sex Roles 1990;23:439-69.
39. Li H. & Fang AC. Word frequency of the CHILDES corpus: Another perspective of child language features. ICAME Journal 2011;35:95-116.
40. Voice Note II. Available from: https://voicenote.in/live/. [Last accessed 2019 Aug 21].
41. Text Analyzer. Available from: https://www.online-utility.org/text/analyzer.jsp. [Last accessed 2019 Aug 21].
42. Malvern D, Richards B, Chipere N, Duran P. Lexical Diversity and Language Development: Quantification and Assessment. New York: Palgrave Macmillan; 2013.
43. Ure J. Lexical density and register differentiation. In: Perren GE, Trim JL, editors. Applications of Linguistics. Selected Papers of the Second International Congress of Applied Linguistics, Cambridge. Cambridge: Cambridge University Press; 1969. p. 443-52.
44. Hudson-Ettle D, Krohne B, Schmied J. International Corpus of English – East Africa; 1999. Available from: http: www.tu-chemnitz.de/phil/english/real/eafrica. [Last accessed on 2019 Dec 18].
45. Terzopoulos AR, Duncan LG, Wilson MA, Niolaki GZ, Masterson J. HelexKids: A word frequency database for Greek and Cypriot primary school children. Behav Res Methods 2016;49:83-96.
46. Svartvik J. The London Corpus of Spoken English: Description and Research. Lund Studies in English 82. Lund: Lund University Press. 1990.
47. Brown GD. A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. Behav Res Methods Instrum Comput 1984;16:502-32.
48. Coltheart M. MRC Psycholinguistic Database User Manual: Version 1. [Available from Professor Coltheart, Birkbeck College, London WC1, U.K.]; 1981b.
49. Berman RA, Verhoeven L. Cross-linguistic perspectives on the development of text-production abilities: Speech and writing. Writ Lang Lit 2002;5:1-44.
50. Johansson V. Lexical diversity and lexical density in speech and writing: A developmental perspective. In: Dept. of Linguistics and Phonetics Working Papers. Vol. 53. Lund: Lund University; 2008. p. 61 79.
51. Xia X. Gender differences in using language. Theory Pract Lang Stud 2013;3:1485-9.
52. Mehl MR, Pennebaker JW. The sounds of social life: A psychometric analysis of students' daily social environments and natural conversations. J Pers Soc Psychol 2003;84:857-70.
53. Horowitz MW, Newman FS. Spoken and written expression: An experimental analysis. J Abnorm Psychol 1964;68:640-7.