

Frequency and Orthographical Similarity Ratings for Kannada Words

Muruli Manikanta

Research Scholar, Dept. of Studies in Psychology, University of Mysore, Manasagangotri, Mysore, Karnataka, India

Abstract: Reading is a complex cognitive process, and the role of script characteristics in reading is of great interest in psycholinguistic research. However, in a multilingual society such as India, researchers often find it difficult to find suitable material in local languages that could be used in research. This study focuses on Kannada, an alphasyllabic script language, to prepare a list of words with a similar frequency and high orthographic similarity that could be used in future studies. A total of 120 words were initially prepared by the researcher. Five native and proficient Kannada speakers participated as the raters. They rated the words on the frequency and orthographic similarity. Descriptive statistics and Pearson's correlation were used for data analysis. Results showed that the selected Kannada words were similar in frequency and orthographic similarity. The inter-rater reliability was significant. The list could be used for specific research that requires such stimuli.

Keywords: Frequency, Orthographical similarity, Inter-rater reliability, Word Reading

1. Introduction

Word frequency is a crucial linguistic characteristic that indicates how often a word is used in language. The word frequency refers to the relative occurrence or prevalence of specific words within a language or corpus, reflecting their usage or distribution patterns (Brysbaert & New, 2009). Evaluating word frequency helps researchers choose stimuli that accurately represent natural language and are more likely to trigger typical language-processing mechanisms (Jones et al. 2020). Orthographically similar words offer researchers a controlled experimental environment to examine the processing of visual and orthographic information. Orthographic similarity refers to the extent to which written words share visual characteristics, such as the shape, arrangement, or appearance of their constituent letters or characters (Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003). By manipulating specific features or positions of characters within sets of similar words, researchers can investigate how readers differentiate and process visually similar stimuli. This approach allows for a detailed analysis of the mechanisms involved in visual and orthographic processing (Gupta, 2019). The use of orthographically similar words enables researchers to study the intricate aspects of word recognition processes. By manipulating specific characters within pairs or sets of similar words, researchers can explore the influence of orthographic similarity on tasks such as lexical access, lexical decision-making, and word identification.

Unfortunately, many of the languages do not have such corpus providing the information on linguistic or orthographic features such as phonology, orthographic similarity, morphology, etc. In such a situation, researchers generally go by rating obtained by proficient native speakers of specific languages in question. This is the procedure followed in this study. If there are multiple raters, it is essential to consider inter-rater reliability when assessing stimulus words. It ensures consistent and dependable evaluation of various linguistic aspects associated with the stimuli. Inter-rater reliability guarantees that different evaluators consistently and reliably assess specific linguistic

features of stimulus words, such as word frequency, word length, or phonological characteristics (Costa & Santesteban, 2004; Levelt, Roelofs, & Meyer, 1999; Kutas & Hillyard, 1984; Marslen-Wilson, 1987). Availability of such reliable linguistic stimuli contributes to the rigor and reliability of psycholinguistic investigations.

2. Method

The researcher selected 120 Kannada words from Kannada textbooks and local newspapers. Another set of 120 words was made by changing the initial letter of words previously selected. Five experts (Native proficient Kannada Speakers) participated as Raters in the present study. For frequency rating, they were required to rate each word on a scale of 1 to 5, based on how frequently they see these words in print (1=Never, 2= Rarely, 3= Moderately, 4= High, 5=Very High). For the orthographic similarity, they were required to rate the words on the orthographic similarity on a five-point scale (1=very low, 2=low, 3=moderate, 4=high 5=very high).

3. Results

Descriptive statistics were followed to find out the average Frequency of words and the average rating for orthographical similarity. Pearson's Product-Moment correlation was employed for checking the inter-rater reliability (see Tables 1-4).

4. Analysis of Results

Table 1: Descriptive Statistics for Frequency

| | Mean | SD | N |
|---------|------|-----|-----|
| Rater 1 | 3.18 | .64 | 120 |
| Rater 2 | 3.63 | .81 | 120 |
| Rater 3 | 4.11 | .74 | 120 |
| Rater 4 | 3.35 | .61 | 120 |
| Rater 5 | 3.31 | .92 | 120 |
| Total | 3.51 | .74 | 120 |

Table 2: Inter-raters Correlation matrix for Frequency

| | Rater 1 | Rater 2 | Rater 3 | Rater 5 |
|---------|---------|---------|---------|---------|
| Rater 1 | - | | | |
| Rater 2 | .323* | - | | |
| Rater 3 | .244* | .428** | - | |
| Rater 4 | .720** | .504** | .777** | - |
| Rater 5 | .526** | .677** | .638** | .382* |

*Significant at 0.05 level, **significant at 0.01 level

Table 1 shows that the Mean Frequency of words was 3.51 with an SD of 0.74. Values below and above 1.0 SD could be considered as low and high Frequencies respectively. Table 2 indicates the correlations between inter-rater ratings. The results revealed that the correlation coefficient is significant ($p < 0.05$).

Table 3: Descriptive Statistics: Ratings for orthographical similarity

| | Mean | SD | N |
|---------|------|------|-----|
| Rater 1 | 4.03 | 0.77 | 120 |
| Rater 2 | 3.51 | 0.69 | 120 |
| Rater 3 | 3.09 | 0.62 | 120 |
| Rater 4 | 3.71 | 0.67 | 120 |
| Rater 5 | 3.85 | 0.73 | 120 |
| Total | 3.63 | 0.70 | 120 |

Table 4: Correlation Matrix: Inter-Rater Reliability of ratings for Orthographical Similarity

| | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---------|---------|---------|---------|---------|
| Rater 1 | - | | | |
| Rater 2 | .384* | - | | |
| Rater 3 | .261* | .594** | - | |
| Rater 4 | .366* | .536** | .713** | - |
| Rater 5 | .554** | .849** | .584** | .695** |

*Significant at 0.05 level, **significant at 0.01 level

Table 3 indicates the Mean and SD of ratings for orthographic similarity. The Mean orthographic similarity of words for words was 3.63 with an SD of 0.70. Values above 4.33 (Mean + 1 SD) may be considered as words with high orthographic similarity by a researcher. Table 4 presents the inter-rater agreement. The results revealed a significant correlation coefficient ($p < 0.05$) suggesting a fair agreement between the raters.

5. Discussion and Conclusion

The present study was aimed at preparing a set of Kannada word lists for future research as such stimuli are not readily available for researchers working in local Indian languages. The study has been successful in providing a list of Kannada word pairs with measures of frequency and orthographic similarity. This list could be used by other researchers, if found suitable for their research questions, in future research. The results have also shown high inter-rater reliability. Using such controlled stimuli ensures the consistency and accuracy of stimulus selection (Paivio, Yuille, & Madigan, 1968), enhances replicability (Simmons, Nelson, & Simonsohn, 2011), and improves the interpretation of research results (Vasishth, Mertzen, Jäger, & Gelman, 2008).

6. Implications and future suggestions

The findings of this study have significant implications for future research on language processing, acquisition, word recognition, etc. The present study offers valuable insights for future researchers regarding the significance of validating procedures for stimuli before incorporating them into a study. Subsequent investigations can delve deeper into the psychometric properties of different types of stimulus words through thorough analysis.

References

- [1] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977-990.
- [2] Johns, B. T., Jamieson, R. K., Crump, M. J., Jones, M. N., & Mewhort, D. J. K. (2020). Production without rules: Using an instance memory model to exploit structure in natural language. *Journal of Memory and Language*, 115, 104165.
- [3] Costa, A., & Santesteban, M. (2004). Bilingual word perception and production. In J. Kroll & A. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 417-439). Oxford University Press.
- [4] Gupta, P. (2019). Orthographic processing in Hindi: Insights from masked priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 45(1), 33-48.
- [5] Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, 307(5947), 161-163.
- [6] Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22(1), 1-38.
- [7] Marslen-Wilson, W. (1987). Functional parallelism in spoken word-recognition. *Cognition*, 25(1-2), 71-102.
- [8] Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology Monographs: General and Applied*, 76(1, Pt.2), 1-25.
- [9] Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359-1366.
- [10] Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2008). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, 99, 100033
- [11] Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology*, 86(3), 169-193.

Appendix

List of Kannada Words (Average rating for frequency)

| A | B | A | B | A | B | A | B |
|---------|-----|---------|-----|--------|-----|----------|-----|
| ಒತ್ತಡ | 4 | ಸ್ಥಳ | 3 | ಮದನ | 3.4 | ದಟ್ಟಣೆ | 3.4 |
| ಘಳಕು | 3.6 | ಧ್ವನಿ | 3 | ಪೂರಕ | 3.8 | ಪುಸ್ತಕ | 3.2 |
| ಪದವಿ | 3 | ಸ್ಥಳೀಯ | 3.4 | ಗಾಯನ | 3.4 | ಮದ್ದನ | 3.8 |
| ಪೀಠಿಕೆ | 3.2 | ವ್ಯಾಸಂಗ | 2.8 | ನರಕ | 4 | ಪತ್ರಿಕೆ | 4 |
| ವಸತಿ | 3 | ಧ್ವಂಸ | 3.4 | ಚರಿತೆ | 3.4 | ಸುಮ್ಮನೆ | 4 |
| ಚರತೆ | 3.6 | ಪ್ರಕಾಶ | 3.4 | ವಸತಿ | 3.8 | ಕಮ್ಮಾರ | 4.2 |
| ಬರಹ | 3.8 | ನೈರುತ್ಯ | 3.8 | ಪವನ | 3.6 | ಕೀರ್ತನೆ | 4 |
| ನರಕ | 3.8 | ಚೈತ್ಯ | 4 | ಬೆಳಕು | 3.8 | ಕನ್ನಡಿ | 3.8 |
| ನಿಮಿಷ | 3.4 | ಸ್ತ್ರಾವ | 3.4 | ಸಮಯ | 4 | ಕತ್ತಲು | 3.6 |
| ನೌಕರ | 4.2 | ಭ್ರಮೆ | 3.2 | ನಯನ | 3.4 | ನೆತ್ತರು | 3.4 |
| ರಚನೆ | 3 | ಸ್ವರಣೆ | 3.4 | ಗಮನ | 3.8 | ಧಾರ್ಮಿಕ | 3.6 |
| ಸರಳ | 3.8 | ಗ್ರಹ | 3.6 | ಸರಾಗ | 3.6 | ಪಲ್ಲವಿ | 4 |
| ಬಂಗಾರ | 3.6 | ನ್ಯಾಯ | 4 | ಬಾಗಿಲು | 3.6 | ಚಾಗ್ರತೆ | 3.6 |
| ಫಲಕ | 3.4 | ತ್ರಾಸ | 4 | ಕುರುಹು | 4 | ಪ್ರಕೃತಿ | 3.8 |
| ಕಡಲು | 4 | ಸ್ವಾರಕ | 3.2 | ಕನಸು | 3.8 | ಒದ್ದಾಟ | 4 |
| ಸರಳ | 3.8 | ಚ್ಚಾಲೆ | 4 | ಕಠಿಣ | 3.8 | ನಾಣ್ಣಡಿ | 4 |
| ಮದುರ | 3.8 | ಕ್ರಮ | 3.6 | ಚರಕ | 4 | ನೆಮ್ಮದಿ | 4.2 |
| ಪಾತಾಳ | 3.2 | ಸ್ವರ | 3.6 | ಸಮರ | 4.2 | ನಿವೃತ್ತಿ | 3.6 |
| ಚಂದನ | 3.8 | ಧ್ಯಾನ | 3.6 | ನುರಿತ | 3.8 | ರಕ್ಷಣೆ | 4.4 |
| ನೆಗಿಲು | 3.6 | ಗೃಹ | 3.4 | ಜಿರಲೆ | 3.6 | ಮತ್ತರ | 3.2 |
| ನೆಮ್ಮದಿ | 3.8 | ಧ್ವಂಸ | 4.2 | ನಾಯಕ | 3.8 | ಘಟ್ಟನೆ | 3.8 |
| ನಡತೆ | 3.6 | ಜೈಷೋಟಿ | 4 | ಸಾಮಾನು | 3.8 | ಮಾನ್ಯತೆ | 4 |
| ಚೆಲುವು | 4 | ಪ್ರಕೃತಿ | 3.8 | ಮುನಿಸು | 3.6 | ಸಕ್ರಮ | 4 |
| ಮರಣ | 3.8 | ಪ್ರಧಾನ | 3.2 | ವಸಂತ | 3.6 | ನತ್ವರ | 4.6 |
| ಕನಸು | 3.6 | ಚ್ಚಾಪಕ | 4 | ನಿಮಿಷ | 3.6 | ಮನ್ನಣೆ | 4 |
| ಕೋಗಿಲೆ | 3.8 | ನೈರುತ್ಯ | 3.8 | ಸಮರ | 4 | ಅಮೃತ | 3.8 |
| ಜನನ | 3.6 | ವೈರಿ | 3.6 | ತಮಟೆ | 3.6 | ಕಿನ್ನರಿ | 3.8 |
| ಹಲವು | 3.4 | ಕ್ರಿಯೆ | 3.6 | ಗುಮಾನಿ | 3.8 | ಸನ್ಮಾನ | 3.6 |
| ಕವನ | 3.4 | ದೈವ | 3.4 | ದಹನ | 3 | ಅಪ್ಪರೆ | 2.4 |
| ಗಾಯಕ | 3.2 | ಸ್ವಾರಕ | 2.8 | ರಗಲೆ | 2.6 | ಕರುಣೆ | 3.2 |

A-Kannada Words, B-Average rating for frequencies

List of Kannada Words (Average rating for Orthographical Similarity)

| A | B | C | A | B | C | A | B | C | A | B | C |
|---------|---------|-----|---------|---------|-----|--------|--------|-----|----------|----------|-----|
| ಒತ್ತಡ | ಬತ್ತಡ | 3.2 | ಸ್ಥಳ | ಸ್ಥಳ | 4 | ಮದನ | ಗದನ | 3.6 | ದಟ್ಟಣೆ | ಗಟ್ಟಣೆ | 4.4 |
| ಛಳಕು | ಛಳಕು | 3.6 | ಧ್ವನಿ | ಧ್ವನಿ | 4 | ಷೂರಕ | ಸೂರಕ | 3.6 | ಪುಸ್ತಕ | ವುಸ್ತಕ | 4 |
| ಷದವಿ | ಷದವಿ | 4.4 | ಸ್ಥಳೀಯ | ಸ್ಥಳೀಯ | 3.4 | ಗಾಯನ | ದಾಯನ | 3.2 | ಮಜ್ಜನ | ಡಜ್ಜನ | 3.4 |
| ಪೀಠಿಕೆ | ಪೀಠಿಕೆ | 3.6 | ವ್ಯಾಸಂಗ | ವ್ಯಾಸಂಗ | 4 | ನರಕ | ವರಕ | 3.6 | ಪತ್ರಿಕೆ | ವತ್ರಿಕೆ | 4 |
| ವಸತಿ | ಭಸತಿ | 4.4 | ಧ್ವಂಸ | ಧ್ವಂಸ | 3.4 | ಚರಿತೆ | ಜರಿತೆ | 3.2 | ಸುಮ್ಮನೆ | ದುಮ್ಮನೆ | 3.4 |
| ಚರತೆ | ಜರತೆ | 4 | ಪ್ರಕಾಶ | ಪ್ರಕಾಶ | 3.6 | ವಸತಿ | ಯಸತಿ | 3.6 | ಕಮ್ಮಾರ | ಗಮ್ಮಾರ | 3 |
| ಬರಹ | ಒರಹ | 3.2 | ನೈರುತ್ಯ | ನೈರುತ್ಯ | 3.4 | ಪವನ | ಷವನ | 3.6 | ಕೀರ್ತನೆ | ಗೀರ್ತನೆ | 2.6 |
| ನರಕ | ಸರಕ | 3.4 | ಜ್ಯೇಷ್ಠ | ಜ್ಯೇಷ್ಠ | 3.4 | ಬೆಳಕು | ದೆಳಕು | 3.6 | ಕನ್ನಡಿ | ಗನ್ನಡಿ | 3.4 |
| ನಿಮಿಷ | ಸಿಮಿಷ | 4.4 | ಸ್ವಾವ | ಸ್ವಾವ | 3.8 | ಸಮಯ | ದಮಯ | 3.4 | ಕತ್ತಲು | ಗತ್ತಲು | 3.4 |
| ನೌಕರ | ಸೌಕರ | 3.8 | ಭ್ರಮೆ | ಭ್ರಮೆ | 3.8 | ನಯನ | ಸಯನ | 3.8 | ನೆತ್ತರು | ಸೆತ್ತರು | 3.8 |
| ರಚನೆ | ಲಚನೆ | 4.2 | ಸ್ವರಣೆ | ಸ್ವರಣೆ | 3.6 | ಗಮನ | ದಮನ | 3.4 | ಧಾರ್ಮಿಕ | ಯಾರ್ಮಿಕ | 3.4 |
| ಸರಳ | ನರಳ | 3.6 | ಗ್ರಹ | ಗ್ರಹ | 3.6 | ಸರಾಗ | ನರಾಗ | 3.8 | ಪಲ್ಲವಿ | ಮಲ್ಲವಿ | 3 |
| ಬಂಗಾರ | ಒಂಗಾರ | 3.2 | ನ್ಯಾಯ | ನ್ಯಾಯ | 3.2 | ಬಾಗಿಲು | ದಾಗಿಲು | 3.4 | ಚಾಗ್ರತೆ | ಚಾಗ್ರತೆ | 4 |
| ಫಲಕ | ಫಲಕ | 3.6 | ತ್ಯಾಸ | ತ್ಯಾಸ | 3.6 | ಕುರುಹು | ಗುರುಹು | 3.8 | ಪ್ರಕೃತಿ | ಗ್ರಕೃತಿ | 3 |
| ಕಡಲು | ಸಡಲು | 3.4 | ಸ್ವಾರಕ | ಸ್ವಾರಕ | 2.8 | ಕನಸು | ಗನಸು | 3.4 | ಒದ್ದಾಟ | ಗದ್ದಾಟ | 3.6 |
| ಸರಳ | ದರಳ | 3.4 | ಚ್ಚಾಲೆ | ಚ್ಚಾಲೆ | 3.6 | ಕಠಿಣ | ಗಠಿಣ | 3.2 | ನಾಣ್ಣಡಿ | ಸಾಣ್ಣಡಿ | 3.4 |
| ಮದುರ | ಗದುರ | 3.4 | ಕ್ರಮ | ಗ್ರಮ | 3.6 | ಚರಕ | ಜರಕ | 3.8 | ನೆಮ್ಮದಿ | ಕೆಮ್ಮದಿ | 3.4 |
| ಪಾತಾಳ | ಷಾತಾಳ | 3.4 | ಸ್ವರ | ಸ್ವರ | 3.6 | ಸಮರ | ದಮರ | 3.2 | ನಿವೃತ್ತಿ | ದಿವೃತ್ತಿ | 3.4 |
| ಚಂದನ | ಬಂದನ | 3.2 | ಧ್ಯಾನ | ಧ್ಯಾನ | 3.4 | ನುರಿತ | ದುರಿತ | 3.4 | ರಕ್ಷಣೆ | ಗಕ್ಷಣೆ | 3.2 |
| ನೇಗಿಲು | ಸೇಗಿಲು | 3.6 | ಗೃಹ | ದೃಹ | 3.8 | ಜಿರಲೆ | ಗಿರಲೆ | 3.2 | ಮತ್ತರ | ಡತ್ತರ | 3.4 |
| ನೆಮ್ಮದಿ | ಸೆಮ್ಮದಿ | 3.8 | ಧ್ವಂಸ | ಘ್ವಂಸ | 3.2 | ನಾಯಕ | ಮಾಯಕ | 3.4 | ಛಟ್ಟನೆ | ಧಟ್ಟನೆ | 3.4 |
| ನಡತೆ | ಸಡತೆ | 3.8 | ವೈಷೋಟಿ | ವೈಷೋಟಿ | 3.8 | ಸಾಮಾನು | ಗಾಮಾನು | 3.4 | ಮಾನ್ಯತೆ | ನಾನ್ಯತೆ | 3.4 |
| ಚೆಲುವು | ಜೆಲುವು | 3.6 | ಪ್ರಕೃತಿ | ಪ್ರಕೃತಿ | 3.4 | ಮುನಿಸು | ಯುನಿಸು | 3 | ಸಕ್ರಮ | ದಕ್ರಮ | 2.8 |
| ಮರಣ | ದರಣ | 3.8 | ಪ್ರಧಾನ | ಪ್ರಧಾನ | 3.8 | ವಸಂತ | ಯಸಂತ | 3 | ನಪ್ಪರ | ದಪ್ಪರ | 3.4 |
| ಕನಸು | ಗನಸು | 3.6 | ಚ್ಚಾಪಕ | ಗ್ಚಾಪಕ | 3 | ನಿಮಿಷ | ಗಿಮಿಷ | 3.4 | ಮನ್ನಣೆ | ಯನ್ನಣೆ | 3.2 |
| ಕೋಗಿಲೆ | ಗೋಗಿಲೆ | 3.8 | ನೈರುತ್ಯ | ನೈರುತ್ಯ | 3.2 | ಸಮರ | ನಮರ | 3 | ಅಮೃತ | ಹಮೃತ | 3.2 |
| ಜನನ | ಚನನ | 3.4 | ವೈರಿ | ವೈರಿ | 3.8 | ತಮಟೆ | ಕಮಟೆ | 3.6 | ಕಿನ್ನರಿ | ಗಿನ್ನರಿ | 3.2 |
| ಹಲವು | ಯಲವು | 3.6 | ಕ್ರಿಯೆ | ಗ್ರಿಯೆ | 3.4 | ಗುಮಾನಿ | ದುಮಾನಿ | 3 | ಸನ್ಮಾನ | ಗನ್ಮಾನ | 3.6 |
| ಕವನ | ಗವನ | 3.4 | ದೈವ | ದೈವ | 3.8 | ದಹನ | ಗಹನ | 3.8 | ಅಪ್ಪರೆ | ಹಪ್ಪರೆ | 4.4 |
| ಗಾಯಕ | ದಾಯಕ | 4.4 | ಸ್ವಾರಕ | ಸ್ವಾರಕ | 3.8 | ಕರುಣೆ | ಗರುಣೆ | 4.6 | ರಗಳೆ | ಲಗಳೆ | 3.8 |

A-Normal Words, B-Orthographical Similar Words, C-Average rating for Orthographical Similarity