VALIDATING DEEP LEARNING GENERATED PSEUDOWORDS IN HINDI

BEHAVIOURAL RESEARCH AND EXPERIMENTAL DESIGN
PROJECT REPORT

Megha Bose

Mukund Choudhary

December 2021

Contents

1 Introduction

1.1 Background

One of the understudied languages in psycholinguistics is Hindi. Like other Indian languages its phonology is different from that of languages which already boast of years of psycholinguistic research and data. As a result, neurologists and linguists alike cannot conduct trials and behavioral experiments for the native Hindi diaspora.

Psycholinguistic databases consist of sequences in the native language words, pseudowords, nonwords etc. with some features like abstractness, number of phones etc. Here are the definitions of the same:

- **Words**: Actual meaningful tokens in a language's dictionary.
- **Pseudowords**: Tokens that sound like words, but don't have a meaning assigned to them yet.
- **Non words**: Tokens which are neither a part of the lexicon nor sound like words in the language.

So far Hindi did not have a psycholinguistic database so far, but recently *Shabd* [1], a psycholinguistic database for Hindi was released. For semantic disambiguation tasks, one needs all three categories in the database. Thus there is a need to include pseudowords too.

1.2 Motivation

As shown above, pseudowords are of immense use in psycholinguistic batteries used to diagnose reading and language disabilities like Aphasia, Dyslexia, etc. They are useful in measuring how intact a patient's semantic/lexical judgement abilities are. For instance, if they can correctly separate all pseudowords from words, then don't have an aphasia which affected their semantic abilities.

Till now, the clinicians have relied on ad-hoc approaches to form pseudowords for use. But these might not conform to an individual's understanding of the sounds in the native dialect of Hindi or can be unfair towards actual words that clinicians didn't know of. Hence it is crucial to have a dependable method that can be used to generate Hindi Pseudowords. One of the methods currently in process at the Cognitive Science Lab is a Deep Learning method to do so and to test its validity/the actual pseudowordiness of the tokens generated, we present the following pilot study.

1.3 Objectives

In this project we aim to design and conduct experiments to validate pseudowords generated by a deep learning model which understands a language's phonology, to validate their closeness to being an actual Hindi word according to native Hindi speakers. This is crucial for enhancing the reliability of this model and then expanding to other languages in the Indian subcontinent.

This validation behavioral experiment is a proof of concept that such experiments can be conducted and what factors one needs to think about while designing such an experiment. We aim to run the entire experiment and generate relevant statistics to show what the entire pipeline for an actual expeirment would look like and also reflect on our gaps so that the main experiment can avoid them.

1.3.1 Hypothesis

Following the above motives, we decided upon the alternate **hypothesis** of the research problem:

H_A: The native speakers will find pseudowords generated by the model to be close to an actual word, thus our model is generating pseudowords which are fluent.

This means that if most participants eliminated the possibility of a lot of samples being a word right away with a high confidence, then the model did not understand the native phonology.

2 Method

2.1 Overview

In order to test the validity of our model generated pseudowords, we hypothesized that repurposing the exisiting Lexical Decision Task framework would be a safe way. However we realised that Lexical Decision tasks have a binary answer to each sequence being a word or not. So we presented the stimuli with a rating system for their Wordiness instead, i.e. how likely people find them to be an actual Hindi word.

Since this was a pilot study, we were unsure of how familiar randomly sampled words would be to a native speaker today and we wanted to make sure that there is an implicit check of nativeness as well, we included a second but similar scale of Familiarity for these tokens.

The stimuli were divided into 20 tokens each in the categories of pseudowords, low frequency, and high frequency words. This division was done keeping in mind that pseudowords have no use frequency while words do, which could lead to a drastic difference in responses. To protect from the effect and to study if it exists, we included low frequency words and separated them from more common words as a set. The other control variable was of length (10 long and 10 short stimuli in each category), as this was the only common variable across the three sets (since pseudowords don't have a semantic or syntactic category associated with them). Finally, the words, their length, and frequency values were all taken from Shabd [1]. The pseudowords were sampled from a dataset of randomly generated output from the DL model and by controlling only their lengths (which we kept close to the other words). A detailed description of the stimuli considered for the experiment can be found in 2.3.

2.2 Participants

The survey-cum-experiment form was scripted on Psytoolkit [2, 3] and was floated by **snowballing** through Hindi native/bilingual friends and family in IIIT-Hyderabad and around our circles. They were instructed by showing a demo of the experiment and we stayed on the online meet for any doubts.

Forty-two individuals participated in this experiment, however some participants were not native enough for the experiment and were not considered for analysis, here are the stats for the final set of participants (A detailed description of the pruning can be found in 2.6.1.):

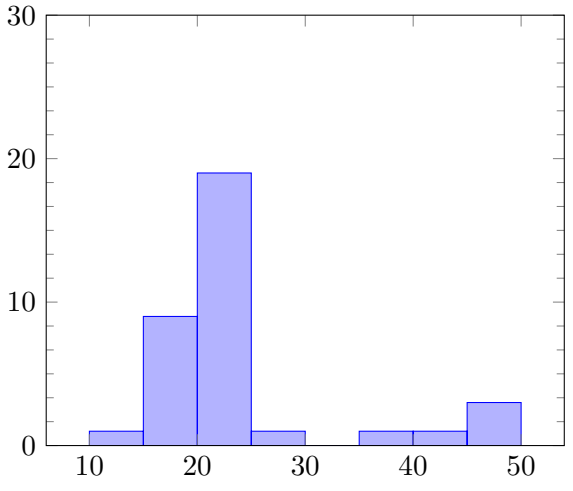

Table 1: Use of Hindi - 1: English, 1.5: Either, 2: Hindi

Question	Mean
To recite the days of a week	1.05
To name the months in a year	1.05
To count upto thirty	1.33
To name common fruits / vegetables	1.63
To help someone write an informal message	1.50
To help someone write a formal mail	1.01
To understand conversations	1.63
To read popular and academic texts	1.07
To argue with someone when angry	1.58
To watch a movie / series	1.39
To talk with parents / grandparents	1.91
To talk with siblings / friends	1.78

Table 2: Self Proficiency Rating (1-5)

Question	Mean
Speaking (Hindi, yourself)	4.4
Listening (someone else in Hindi)	4.58
Reading (a text in Hindi)	4.0
Writing (something in Hindi)	3.80

The **hindi nativeness** is calculated by taking the average of the above Hindi self proficiency ratings.

Table 3: Participant summary

Measure	Min	Max	Mean	SD
Age	14	47	23.28	8.51
Hindi Nativeness (1-5)	2	5	4.21	0.79

2.3 Design

2.3.1 Overview

The whole process was divided into three major sections:

1. **Survey**: The survey described risks, confidentiality, consent etc. followed by age and confirmation on vision abilities and calm environment.

2. **Self Proficiency Report**: This part asked participants to report their proficiency in Hindi and English, and on the Language Use Questionnaire as used in [4].
3. **Task**: The experiment asked the participants to rate Devanagari strings according to the rating system described briefly above and in detail over at 2.3.
4. **Feedback**: Finally they were asked to fill in Feedback if they had something to convey.

2.3.2 Variables

Independent Variables The independent variables (IVs) in our experiment are as follows (most of these were taken directly from Shabd [1]):

- **Token Length (Long/Short)** - number of phonemes in the token. (threshold: 5.0)
- **Word Frequency (High/Low)** - number of times a word appears in the language. (threshold: 3.50)
- **Status (Word/Pseudoword)** - whether the token is a word taken from Shabd database of a potential pseudoword generated by our deep learning model.

Dependent Variables The dependent variables (DVs) in our experiment are as follows:

- **Wordiness Score** - score on a scale from 1-7 to rate how likely a token is an actual word in Hindi, score of 1 being least likely to be a word and score of 7 being most likely to be a word.
- **Familiarity Score** - score on a scale from 1-7 to rate how familiar a token is to the participant, score of 1 being least familiar and score of 7 being most familiar.
- **Wordiness Reaction Time** - time taken to rate wordiness per token.
- **Familiarity Reaction Time** - time taken to rate familiarity per token.

Inferred Dependent Variables We also built on top of DVs to infer other useful statistics:

- **Accuracy** - For the word tokens, we find out if a participant rated words lowly on the wordiness scale consistently for a lot of tokens. This is done as an implicit check of nativeness of the participant.
- **Confidence** - A participant's reaction time for ratings is also taken into account to judge whether they were too slow in responding to well known words or too quick to judge pseudowords. Only the extreme ends were taken into account. This was used in combination with Accuracy to detect outliers/non-native participants.

Controlled Variables The controlled variables used in our experiment are as follows:

- **Breaks** - We kept consistent but equal gaps between presenting the wordiness and familiarity scales for a particular stimuli (200 ms) and a higher but equal gaps between presenting two different stimuli (400 ms).
- **Timeout limit** - maximum time before the next stimulus is displayed (10000 ms). In case of timeout the response collected was -1 and we replaced that data point with the mean response of the participant for stimuli of that category.
- **Display details** - position of the token and the scales, colour scheme, etc.

Since it is a pilot, our experiment is a **controlled within-subject (repeated measures)** experiment where researchers have complete control over independent variables and the participants. The researchers are known to each other and all the participants rate all the words and pseudowords considered.

The experiment can be viewed in 2 ways:

- operationalized using a **2X2 factorial design** with the IVs of token length and the boolean of whether the token is a word or a pseudoword.

- operationalized using a **3X2 factorial design** with the IVs of token length and the boolean of whether the token is a high frequency word, low frequency word or a pseudoword.

	Short	*Long*
Word	2 x 2 design	
Pseudoword		

	Short	*Long*
High Freq. Word	3 x 2 design	
Low Freq. Word		
Pseudoword		

Figure 1: Both kinds of Designs

2.4 Task

A quick summary of above, the task had three sets of randomised stimuli with **20 tokens** each: *pseudowords*, *low frequency words*, and *high frequency words*. Each set had **10 words each of short and long lengths**, all presented in Devanagari script.

The participants had to click on one of the seven circles to **rate** the stimuli on a scale of 1 - 7 based on their wordiness and familiarity (Figure 2). Each slide timed out with a default option in case the participant didn't answer right away. The stimuli was presented one by one with a gap of 400 ms and were positioned at (0, -100) with respect to top of the screen. They were presented in black font on yellow background. The wordiness scale was positioned at (0, 200) with respect to to the bottom of the screen with the 7 circles ranging from "could be a word" to "cannot be a word" and after a gap of 200 ms from the last rating, a blue 7 pointer scale was shown from "not familiar" to "familiar".

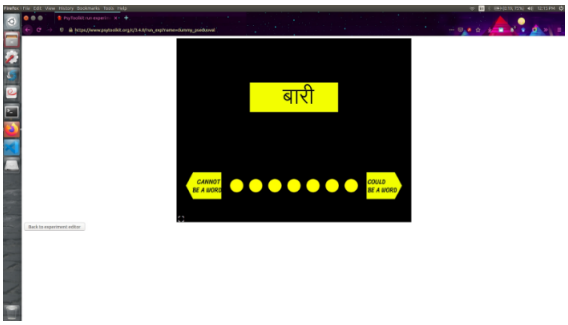

Figure 2: Stimulus

2.5 Procedure

The participants were shown the demo, asked for nativeness and consent verbally before being sent the psytoolkit survey link. After this, they were shown the policies, risks involved etc and on acceptance they were asked for demographics, vision information, and nativeness self reports.

They were then asked to be calm and ready for a continuous sitting to attempt the main experiment and they were shown the stimuli to rate as described above.

On completion of the main task, they were asked for feedback and then redirected to the homepage of duckduckgo. During the attempt, the researcher stayed on an online meeting with them for clarifications of any kind. A sequential image of the same can be seen below:

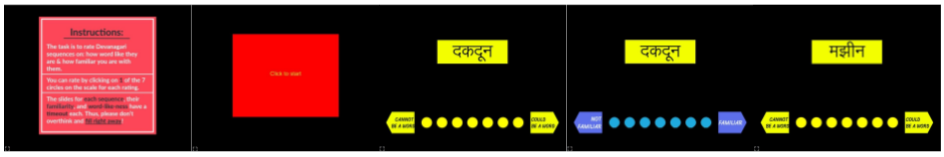

Figure 3: Sample Sequence of slides

2.6 Measure of the Performance

Since this is majorly a validation experiment, we did not have a baseline to measure performances to compare responses against. We also did not have an experimentally proven expected response to measure and find out, thus in this section we present how we judged their nativeness and found outliers who didn't match our conditions for a good quality performance on the basis of their word scores (not including pseudowords) and their reaction time scores.

2.6.1 Outliers/Non-natives

To find the outliers from the data who were non-native or were timed out for most responses etc. we used the following measures' intersection of outliers:

- **Reaction Time extremes**: We found the intersection of outliers from standard Inter Quartile Range elimination on the RTs collected from familiarity, wordiness, and overall.
- **Inaccurate Word ratings**: We found participants who consistently rated words (not pseudowords) below their mean ratings for a large amount of words too.

We then find out the worst performing outliers from both lists and eliminate them. The statistics for the same can be found below at 3.1.

3 Results

3.1 Description of the scope of the analysis

We ensured that we could draw meaningful statistics and graphs by making sure outlying and erratic participants were not considered and by replacing timed out default values (-1) by means.

The mean replacements had to be done for 58 out of 5040 datapoints. We replaced these default values which were marked as -1 by psytoolkit when the participants timed out for the particular stimuli's scale, by means of the participants' responses for the stimuli in that subcategory.

Finally, from the Outlier detection as detailed above, we found 6 participants who were outliers with respect to RT (intersection of standard IQR on RTs by wordiness, familiarity, and overall) and had a high (above or around 50%) number of word (not pseudoword) ratings below mean per word.

We calculate the summary statistics for the measures and observe the trends. We can see that our study checks the construct of how close a token is to an actual Hindi Word through wordiness and familiarity ratings. In order to look into how the ratings correlate with each other for different types of tokens, we use Spearman's rank correlation coefficient (ρ).

3.2 Data representation

3.2.1 Summary Statistics

The following are the summary statistics from the final 36 participants, organised according to the divisions by design as mentioned above:

Table 4: Summary Statistics and Correlation

	Token Type	Wordiness Rating				Familiarity Rating				ρ	
		Mean		SD		Mean		SD			
Pseudoword (n=20)	pslong (n=10)	3.94	4.16	1.88	1.87	2.25	2.47	1.51	1.65	0.74	0.95
	psshort (n=10)	4.39		1.84		2.7		1.75			
Word (n=40)	lflong (n=10)	6.48	6.4	1.16	1.29	5.65	5.79	2.07	1.99	0.89	
	lfshort (n=10)	6.2		1.5		5.52		2.24			
	hflong (n=10)	6.43		1.27		5.87		1.86			
	hfshort (n=10)	6.5		1.19		6.13		1.71			

Table 5: Pseudoword vs High Frequency Words vs Low Frequency Words

	Wordiness Rating		Familiarity Rating		
	Mean	SD	Mean	SD	ρ
Pseudoword (n=20)	4.16	1.87	2.47	1.65	0.74
Low Frequency (n=20)	6.34	1.35	5.59	2.16	0.91
High Frequency (n=20)	6.47	1.24	6	1.79	0.87

3.2.2 Statistics Visualization

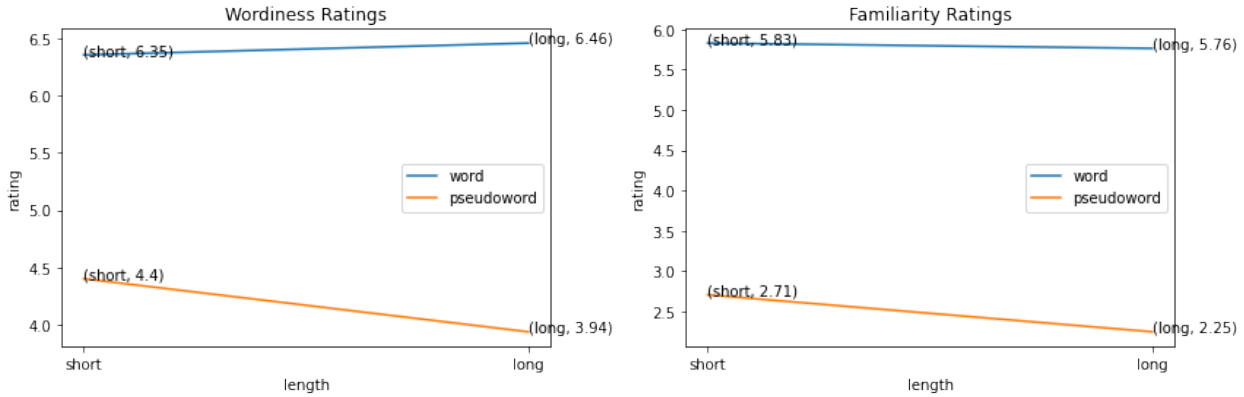

Figure 4: Ratings Graphs for 2x2 design

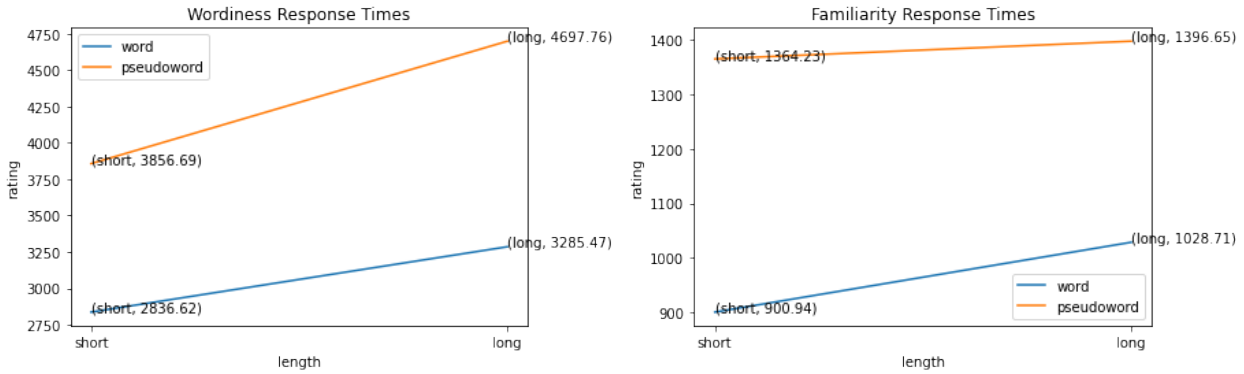

Figure 5: Response Times Graphs for 2x2 design

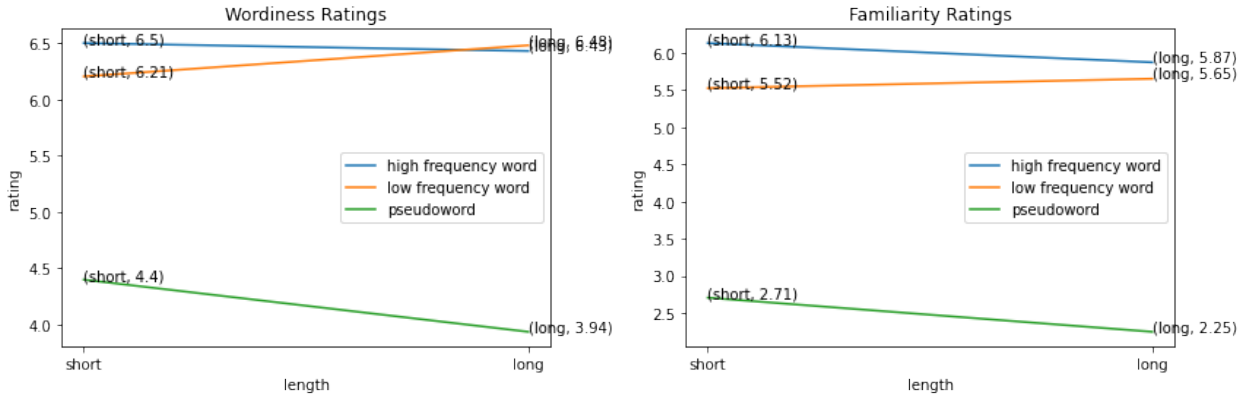

Figure 6: Ratings Graphs for 3x2 design

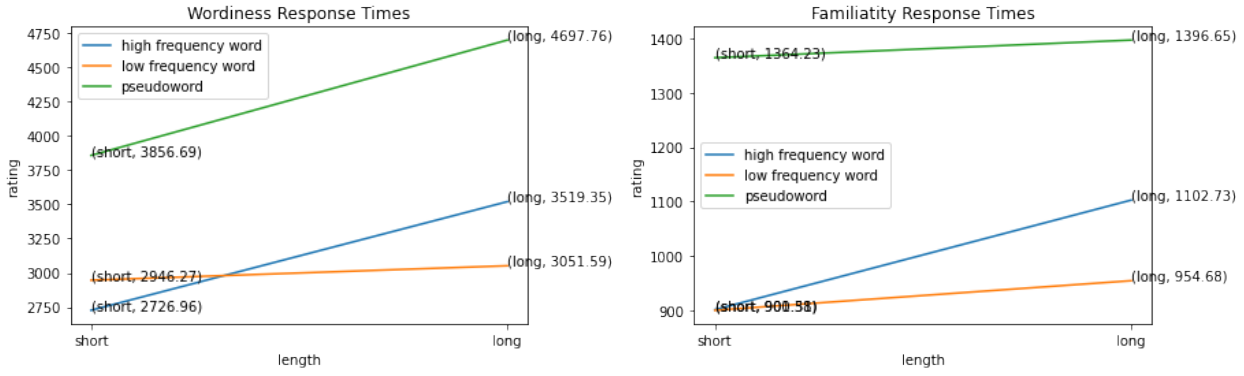

Figure 7: Response Times Graphs for 3x2 design

Main Effects From the 2x2 design graphs, we can see that there exists main effect of whether a token is word/pseudoword in all the measures. Words, in general, have higher rating scores compared to pseudowords. On the other hand, as expected, pseudowords have higher response time values in general as compared to words.

However, main effect of token length is seen for familiarity rating, familiarity response time and wordiness response time but it is significant only for wordiness response time.

From the 3x2 design graphs, we see that main effect of token type exists in familiarity ratings and familiarity response times (almost). Main effect of token length is seen in both the response time graphs.

A general trend that we can see from all the response time graphs is that, as expected, the response times are higher for longer tokens.

Interaction From the 2x2 design graphs as well as the 3x2 design graphs, we can see that interaction between the IVs can be seen in all the four measures.

4 Discussion

We can see from Table 4 that the mean ratings of pseudowords came out to be 4 while that of words is 6.41. Since the scale ranged from 1-7 with 1 being the score given if the participant completely rejected the possibility of the token shown to be a Hindi word, the results we got imply that the participants deny pseudowords nor completely accept them, and took reasonable time as well.

This is further strengthened by the trends observed in means and standard deviations. We can see from the Table 5 that the Mean wordiness ratings for the categories grow in magnitude as they become more common/frequent and participants are more sure of their answers as well (indicated by the lessening standard deviation).

Overall it indicates that in the natural progression of wordiness and familiarity of the stimuli presented, pseudowords closely follow words and are not floored down to the extreme values completely.

The experiment as a notable chaos in the longer stimuli segments. This is reflected in the trends for only long words in means and standard deviations and can also be seen in the main effect graphs. The results get dilute for these set of stimuli and we hypothesise this could be because of the DL model or readability issues for the given timeout for participants.

We also see a strong correlation between wordiness and familiarity for all categories and divisions all are above 0.7 and one value above 0.9 can also be seen. This increases the reliability of the pseudowords generated being close to a Hindi word for the participants. Not only do they think it could be a word, it also *appears familiar* to them as much as it seems word-like.

Finally, the pilot also shows that the experimental setup can be replicated on a larger scale to validate pseudowords generated by Machine learned methods and be put to use for low resource languages.

4.1 Some linguistic observations

We could also see that there were issues from the words and their frequencies we obtained from the Shabd psycholinguistic database. We saw that some words which were classified as high frequency were not actually so and a few which were noted as low frequency were percieved as pseudowords by the participants. This also skewed our results a bit.

We saw that words which were shorter/had lesser phonemes had slightly stronger results than the longer words. This could be a result of expectation of meaningful morphemes towards the end of long words by the participants. Thus the DL model should be designed to account for length based generation and should be supplied that as a factor too.

4.2 How the current results contribute to knowledge in the field?

There has been a need of research for psycholinguistic studies in Indian Languages. However we don't have the appropriate lexicons to do such research on. The Deep Learning model takes a step in this direction, but also needs an appropriate validation for its outputs. This experiment show that it is a valid method and can be used again or on different Indian languages.

This pilot also demonstrates that a validation experiment for a domain as tricky as pseudowords can be designed well. A full scale experiment fixing the mistakes as we describe below, could be the solution to a lowly researched domain as well.

5 Conclusions

5.1 Major findings of the current project/work

We found that this design is a good framework to build upon, for a pseudoword validation experiment. We can gauge the pseudowordiness to a comfortable degree and also store them as a new psycholinguistic feature for clinicians and other researchers to explore. We also found that wordiness can be impacted by a lot more than just phonology, e.g. factors like borrowed words, knowledge of the world etc. Statistically, we see that pseudowords are also being correctly treated as lower frequency words by the participants under this design. Thus we conclude that the DL model can be validated by such a design from the pilot findings we have here.

As a minor result, we also saw pseudowords being rated just below the word range.

5.2 Important implications of the current findings

We can see that a feasible pseudoword validation experiment could be made similarly and conducted fairly with native Hindi speakers. It is also implied (albeit on a small scale) that the pseudoword generation is valid. This is because the participants have not classified them as the lowest category, but towards the middle on the wordiness scale.

Finally, we now know some changes that need to be done to make this a full fledged experiment:

- **Wording** - Changing the wording from "Could be a word" to "Sounds like a word" could make a major impact on the perception of participants. We saw this happen during the pre-pilot where it was "Is a word" and also during feedback from the demo, where some participants were confused to an extent
- **Nonwords** - In our design, a natural extension and a strong reference point could have been made by including nonwords as a set of stimuli as well. This is also important linguistically as it would make it easier to find patterns which are absent in nonwords but present in both pseudowords and words.

5.3 Limitations of the current work

We were limited by the following:

1. **Quality**: Participants in this pilot study (as seen by the data on the self report and word accuracies) were mostly bilingual and were not native "enough". This skewed some of our data and we also had to remove 6 responses which had an exceptionally outlying RT in combination with inaccurate word ratings.
2. **Quantity**: Since this was a limited course project/pilot, we could not compensate the participants for their participation and could not extend the experiment to more than 15-20 mins. Which also limited our stimuli that we could present thus diminishing the power of the experiment to an extent.

References

[1] Ark Verma, Vivek Sikarwar, Himanshu Yadav, Ranjith Jaganathan, and Pawan Kumar. Shabd: A psycholinguistic database for hindi. *Behavior Research Methods*, pages 1–15, 08 2021. doi:10.3758/s13428-021-01625-2.

[2] Gijsbert Stoet. Psytoolkit: A software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4):1096–1104, Nov 2010. ISSN 1554-3528. doi:10.3758/BRM.42.4.1096. URL https://doi.org/10.3758/BRM.42.4.1096.

[3] Gijsbert Stoet. Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1):24–31, 2017. doi:10.1177/0098628316677643. URL https://doi.org/10.1177/0098628316677643.

[4] D Vasanta, Suvarna Alladi, J. Sireesha, and Bapi Surampudi. Language choice and language use patterns among telugu-hindi/urdu-english speakers in hyderabad, india. *International Conference on Language, Society, and Culture in Asian Contexts (LSCAC) Proceedings*, pages 57–67, 01 2010.