CWID-hi: A Dataset for Complex Word Identification in Hindi Text

Gayatri Venugopal[1], Dhanya Pramod[2], Ravi Shekhar[3]
[1]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University)
[2]Symbiosis Centre for Information Technology, Symbiosis International (Deemed University)
[3]Cognitive Science Research Group, Queen Mary University of London

Motivation and Objectives

We focus on lexical simplification of Hindi text, i.e., the process of identifying complex words in a given text and substituting them with their simpler synonyms based on the context of the target complex word. This area is unexplored for Hindi with respect to methods and resources.

Objectives:

- To study the annotations obtained from annotation tasks conducted to identify complex words in a given Hindi sentence
- To create a dataset of simple and complex words
- To build a model on the dataset and subsequently test it using annotations obtained from annotators with varying levels of exposure to Hindi

Annotators, Annotation Data and Tasks

100 Resident Indians in the age group of **18-30 years**, who have **studied Hindi** as part of their curriculum

Average Age: 19.66
Standard Deviation: 2.78
43 Females
57 Males

Native	Non-Native
10 Groups consisting of 5 annotators each	**10 Groups consisting of 5 annotators each**

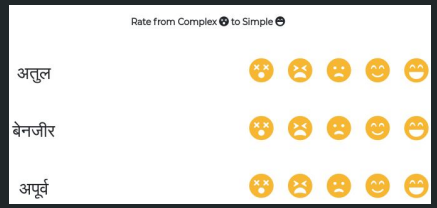
Annotation Data

Aesthetics Corpus (90 x 20) 1800 sentences	Twitter 10 sentences

Task 1 (Complex Word Identification)

कतर एयरवेज ने एयर इंडिया के अधिग्रहण से जुड़ी किसी भी बातचीत से इनकार किया ।

Task 2 (Rate the complexity of the selected word and its synonyms)



Observations

- Only 28% annotators chose their native language as the most comfortable language to read. 93.05% of the remaining annotators were most comfortable reading English text.
- Only 30% of the native Hindi speakers chose Hindi as the most comfortable language to read.
- 66% native annotators chose English as the
- 66% native annotators were comfortable with English whereas 24% annotators chose the language of the region they resided in, for the maximum duration.

Dataset Creation

35,471 complex word annotations | **32,636 simple word annotations**

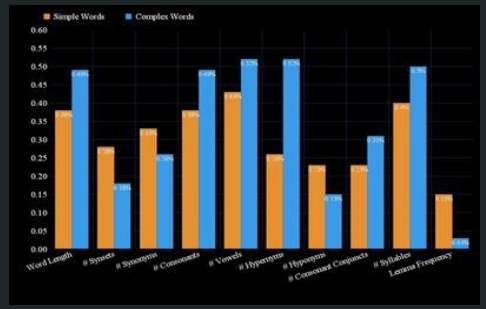
63,857 words (excluding digits)

18,186 unique words

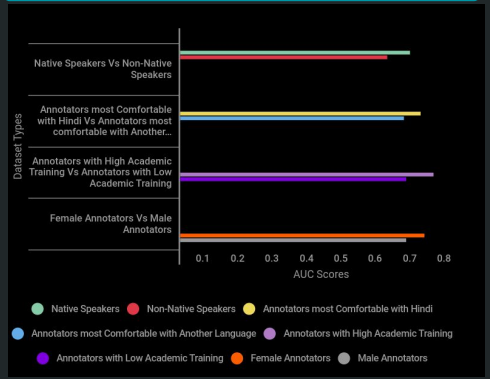
12,111 unique words ranked by at-least 2 participants

7,321 unique words present in the corpora

We performed sense-based normalisation, i.e., the values of features of a word was compared with those of its synonyms while performing normalisation. Values of features of complex words, such as length were slightly larger than that of words labelled as simple, and values of frequency were smaller in complex words as compared to those in simple words.
The mean of the feature values of complex and simple words is shown here.



Dataset Evaluation



Link to the dataset:
https://zenodo.org/record/5229160

The study is funded by Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University) under gr
1591.
RS is supported by European Union Horizon, 2020 research and innovation programme under grant 825153 (EMBEDDIA).