Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5627–5636
Marseille, 20-25 June 2022

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

Gayatri Venugopal[◇], **Dhanya Pramod**[*], **Ravi Shekhar**[†]
[◇]Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India
[*]Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, India
[†]Cognitive Science Research Group, Queen Mary University of London, London, UK
gayatri.venugopal@sicsr.ac.in, dhanya@scit.edu, r.shekhar@qmul.ac.uk

Abstract

Text simplification is a method for improving the accessibility of text by converting complex sentences into simple sentences. Multiple studies have been done to create datasets for text simplification. However, most of these datasets focus on high-resource languages only. In this work, we proposed a complex word dataset for Hindi, a language largely ignored in text simplification literature. We used various Hindi knowledge annotators for annotation to capture the annotator's language knowledge. Our analysis shows a significant difference between native and non-native annotators' perception of word complexity. We also built an automatic complex word classifier using a soft voting approach based on the predictions from tree-based ensemble classifiers. These models behave differently for annotations made by different categories of users, such as native and non-native speakers. Our dataset and analysis will help simplify Hindi text depending on the user's language understanding. The dataset is available at `https://zenodo.org/record/5229160`.

Keywords: lexical simplification, complex word identification, dataset, classification, Hindi

1. Introduction

The significance of accessible technology has increased manifold today for creating an inclusive environment. Accessibility not only helps improve the quality of life of persons with impairments but also benefits individuals without impairments (Henry, 2006). We can associate accessibility with physical devices as well as with the content we consume in our day-to-day lives. We can achieve accessibility of textual content by simplifying it. Text simplification refers to the process of modifying text in such a way that it becomes easier for the reader to comprehend it. It is a sub-field of natural language processing, that has proven to be useful to children (De Belder and Moens, 2010), readers with language impairments (Caplan, 1992; Carroll et al., 1999; Carroll et al., 1998), readers with low literacy levels (Candido Jr et al., 2009), and non-native speakers (Paetzold and Specia, 2016c). Conceptual simplification, elaborative modification, syntactic simplification, and lexical simplification are different ways by which we can simplify text (Siddharthan, 2014).

In this work, 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 (Paetzold and Specia, 2017). The term 'complex' in this context does not indicate syntactic or morphological complexity. A complex word is defined as a word that is difficult to understand by the reader (Finnimore et al., 2019; Yimam et al., 2017a). Shardlow (2014) has demonstrated the steps in a lexical simplification pipeline, which is constituted of complex word identification, substitution generation, word sense disambiguation, and synonym ranking. However, there is no work related to lexical simplification in Hindi, which, according to Ethnologue[1], has the third-largest number of speakers in the world, after English and Mandarin Chinese. Moreover, Hindi is the official language of the government of India[2], and 43.63% people in the country are Hindi speakers, according to the last census conducted in India[3]. Since people's vocabulary varies according to their familiarity with the language, it is essential to produce and distribute content that all can understand. Manual simplification of content could be time-consuming and laborious; therefore, we need to use automatic text simplification approaches.

Soni et al. (2013) performed sentence simplification on Hindi text. They performed simplification by splitting the sentence into multiple sentences, which involved modification of the grammatical structure of the sentence. Mishra et al. (2014) also experimented with sentence splitting in order to improve the quality of translation from Hindi to English. Both these studies are instances of syntactic simplification of Hindi text. The area of text simplification, specifically lexical simplification, is unexplored for Hindi. Lexical simplification studies have gained popularity in various other languages such as English, French, Brazilian Portuguese, Spanish, to name a few (Paetzold, 2016; Lee and Yeung, 2018; Hmida et al., 2018; Aluísio and Gasperin, 2010; Štajner et al., 2019; Bott et al., 2012). However, certain resources used for simplification, such as the Simple English Wikipedia (Coster and Kauchak,

[1] `https://www.ethnologue.com/guides/ethnologue200`

[2] `https://rajbhasha.gov.in/en/official-language-resolution-1968`

[3] `https://censusindia.gov.in/2011census/Language-2011/Statement-4.pdf`

2011), simple word lists (Ogden, 1930), ranked synonym lexicons (Billami et al., 2018) etc. do not exist in Hindi. We also do not have an annotated dataset to differentiate between complex and simple words in Hindi text. The differences in languages, such as the presence of consonant conjuncts in Hindi words, imply a need to study the effect of such features on the lexical complexity of sentences. Therefore it is essential to create language-specific resources specifically for Hindi text simplification.

We identified the following research objectives of our study:

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

We extracted sentences from a corpus of Hindi text consisting of novels and short stories spanning over a hundred years (Venugopal-Wairagade et al., 2020). One hundred native and non-native annotators annotated the sentences. We cleaned the data and thus acquired and created a dataset consisting of feature values of the annotated words and their synonyms. The dataset consists of a binary label, wherein 1 indicates complex and 0 indicates simple. It also contains each word's 'simplicity' value, ranging from 1 (complex) to 5 (simple). The dataset can be used further for classification-based approaches to complex word identification. The main challenge we faced was the subjective nature of the problem, as the complexity or incomprehensibility of a word is closely linked with the vocabulary of the reader and their familiarity with the word. Therefore we have attempted to include readers familiar with a diverse set of languages besides Hindi to replicate a real-world scenario wherein readers with varying vocabulary skills would consume the content. However, it must be noted that since there are common shared Sanskrit words in certain Indian languages, non-native Hindi speakers may be familiar with these shared words as well. Hence, it is possible that a non-native speaker of Hindi may be equally or more proficient in the language.

This paper reports the details and results of the annotation tasks conducted to collect data to create the dataset. We also present the results of tests conducted to determine the model's performance trained on the dataset for annotations obtained from different subsets of annotators. The main contributions of this work are listed below:

- A dataset for complex word identification in Hindi text
- Sense-based normalization of features. The feature values were normalized by considering only the target word and its synonyms instead of including the feature values of all the words in the dataset, thus avoiding the comparison of values of the target word with those of unrelated words.
- Performance analysis of classifier trained on the dataset, for different categories of readers

2. Related Work

This section consists of the details of lexical simplification datasets created in different languages. One of the earliest works on automatic lexical simplification datasets was SemEval-2007 Task 10: English Lexical Substitution Task (McCarthy and Navigli, 2009). In this task, native English-speaking annotators substituted a given target word with a simpler alternative. However, the focus of this task was not modeling the complex word identification process, and we believe that the absence of non-native speakers in the task may have introduced a bias. As part of SemEval-2012 Task 1: English Lexical Simplification(Specia et al., 2012), non-native English speakers annotated the corpus used in McCarthy and Navigli (2009). Here, the annotators were asked to rank the substitute words for the given set of target words present in the dataset used in SemEval-2007 Task 10: English Lexical Substitution Task. This task too focused on generating substitute candidates for a given target word. SemEval 2016 Task 11: Complex Word Identification focused on the problem of complex word identification (Paetzold and Specia, 2016b). The researchers annotated the dataset with the help of 400 non-native English-speaking annotators. The objective of the task was to annotate at the most one complex word in a given sentence. Here, 20 people annotated each sentence in the training set, whereas only one annotator annotated each sentence in the test set, which could have introduced bias in the test set. In the Complex Word Identification Shared Task 2018 (Yimam et al., 2018), native as well as non-native annotators participated in the study, and the languages in focus were English, German, Spanish and French. The dataset for this task was named CWIG3G2 and was created by Yimam et al. (2017b). They modeled complex word identification as a binary classification task and a probabilistic classification task, wherein the annotators assigned the probability of the given target word as complex. 134 native and 47 non-native annotators participated in the process. The organizers asked the annotators to assume that the target readers were children, readers with language impairments, or learners of the language, i.e., they focused on assumed complexity.

Maddela and Xu (2018) created a dataset for a neural readability ranking model. They chose the most frequent 15,000 English words from the Google 1T Ngram Corpus (Brants, 2006). 11 non-native English speakers annotated the words on a 6-point Likert scale

(Likert, 1932). In order to ensure the quality of annotations, they calculated the Pearson's correlation coefficient between an annotator's annotation and the average of the annotations given by the rest of the annotators and used only the correlated words. The final rating was the average of the ratings received by each word.

Shardlow et al. (2021) created two datasets - CompLex 1.0 and Complex 2.0. CompLex 1.0 consists of data annotated by their complexity level, whereas CompLex 2.0 is an improvement as it consists of more instances and more annotations for each instance. This dataset consists of English text from the Bible, Europarl (European Parliament proceedings), and Biomedical articles. They used a 5-point Likert scale depicting complexity to annotate each word. Similar to the process followed by Maddela and Xu (2018), they ensured the quality of annotations by calculating the correlation between the annotation assigned by an annotator to a word with the average annotations it received. They also calculated the correlation between the annotations received by the word and its frequency. The complexity value assigned to a word was a normalized average of its annotations. They assumed that there should be a correlation between an annotator's annotation and the average of the annotations made by the other annotators. However, since the problem of complex word identification is extremely subjective, we are unsure whether we should compare the annotations or expect a high inter-annotator agreement. They assigned a complexity value that was normalized across the annotations. However, since the annotators were not provided with a list of word synonyms, they could not comprehend it. Hence we are unsure whether the complexity value provided by an annotator was a judgment based on comparison with the familiarity of other unrelated words or a judgment based on the familiarity of other related, i.e., synonymous words.

Besides these datasets, there are various other datasets that consist of target complex words and their simpler substitutions, such as LSeval (De Belder and Moens, 2012) LexMTurk (Horn et al., 2014), BenchLS (Paetzold and Specia, 2016a), CW Corpus (Shardlow, 2013) for English, SNOW E4 (Kajiwara and Yamamoto, 2015) for Japanese, and SIMPLEX-PB for Portuguese (Hartmann et al., 2018).

Our observations of the datasets and the description of the annotation processes carried out by the studies mentioned in this section are that the complex word was predetermined in many of the annotation tasks, and the annotator was required to rank its simpler substitutes. In a few cases, the annotators were native speakers or were asked to annotate by making assumptions about the target readers. Our work is motivated by the observation that there is no dataset for lexical simplification or complex word identification for the Hindi language, and a dataset with lexical features of words is not readily available. The objective of our study was to conduct an annotation task, analyze the annotations, and create a dataset that could be used for identifying complex words in a given Hindi text. The need for a new dataset is also justified owing to the differences in languages. Also, since no such study exists that correlates the frequency of a word with its complexity in Hindi texts, we cannot assume frequency to be a relevant feature.

3. Dataset Creation and Evaluation

We addressed the objectives mentioned in the introduction section by asking the following questions:

- Objective 1: To study the annotations obtained from annotation tasks conducted to identify complex words in a given Hindi sentence.
  - RQ1: Is there any difference between the native language of an annotator and the language that they were most comfortable reading?
  - RQ2: Is there any relationship between the language that an annotator was most comfortable with and the region's official language in which they spent the maximum number of years?
  - RQ3: Is the lemma of a word considered to be simpler than a morphological variant of the word?
- Objective 2: To create a dataset that can be used for automatic labeling of words as simple or complex
  - RQ4: Is there a significant difference between the values of features of words labeled as complex and that of words labeled as simple?
- Objective 3: To train a classifier on the dataset and subsequently test it using annotations obtained from annotators with varying levels of exposure to Hindi
  - RQ5: Is there a difference in the performance of the model w.r.t. test data created from annotations obtained from categories of users formed based on the following criteria?
    - ∗ Native language
    - ∗ Hindi being the language that they were most comfortable with
    - ∗ Years of academic training in Hindi
    - ∗ Self-identified gender

The following subsections contain a description of the tasks performed in order to find answers to the research questions.

3.1. Annotation Tasks

3.1.1. Annotators

Our first objective was to conduct an annotation task to create a labeled dataset and study the annotations obtained from these tasks. We recruited 100 annotators of Indian origin, with ages in the range of 18-30 years, who have studied Hindi as part of their school curriculum. Annotators consisted of university students as well as working professionals. However, we did not evaluate the annotators for language proficiency as we did not target a group based on language proficiency. Our objective was to understand how simplification varies according to varying exposure to the language. The average age of the annotators was 19.66, with a standard deviation of 2.775. 43 annotators identified themselves as females, whereas 57 annotators identified as males. The annotators also mentioned their native language. Based on this information, we divided the annotators into 20 groups, wherein each group consisted of 5 annotators. Groups 1-10 consisted of non-native annotators, whereas groups 11-20 consisted of native annotators.

3.1.2. Data for Annotation

We used anesthetics corpus (Venugopal-Wairagade et al., 2020) for our experimentation. The corpus consists of novels and stories available at `http://hindisamay.com`, an e-library that is developed and maintained by Mahatma Gandhi Antarrashtriya Hindi Vishwa Vidyalaya (translated to Mahatma Gandhi International Hindi University), Wardha, `http://premchand.co.in`, a website dedicated to the legendary Hindi novelist Premchand, and Bhandarkar Oriental Research Institute's Digital Library (`http://borilib.com`). We used 978 articles from these sources. We also scraped Twitter for sentences. We split the text into sentences and tokenized them further after removing English words, special characters, and Latin numbers. We created 20 sets of 100 sentences, wherein each set was presented to one group of annotators. There were 10 common sentences in each set extracted from Twitter. Our objective was to study the simplification requirements of words used in texts such as novels, short stories, and biographies; therefore, we did not consider texts from other domains such as history, law, technology, etc.

3.1.3. Annotation Tasks

We developed an online system for the annotation tasks. The annotation process was divided into two tasks, wherein Task 1 was used to obtain the words whose meanings the annotator could not understand, and Task 2 was used to obtain the complexity level of the word and its synonyms. Finally, we designed Task 2 in order to obtain a complexity value of the target word that can be compared with that of its potential substitutes.

Therefore each annotator annotated maximum 100 sentences in Task 1. The 5 annotators in a group were presented with the same set of sentences. They were required to highlight the word/s in a sentence whose meaning they could not understand. The annotators were not allowed to select multi-word expressions in Task 1. Therefore, an annotator did not need to annotate each sentence, i.e., if they understood the meanings of the words, they would not mark any word as complex.

Task 2 consisted of a list of words containing the lemma of the word highlighted as complex in Task 1, and all its synonyms extracted from the Hindi WordNet (Bhattacharyya et al., 2008). The annotators were required to rate each of these words using emojis. However, we did not use relative rating to avoid confusion and cognitive load if a word has many synonyms. There were 5 emojis on the rating scale, ranging from angry to happy, where angry indicated complex words and happy indicated simple words. A word with a rating less than or equal to 3 was considered complex. We included this task as it seemed intuitive to compare a word with its synonyms instead of with unrelated words.

3.2. Dataset

The distribution of annotations obtained from Task 1, can be seen in Figure 1. We observed that out of the 4,599 annotated words, the number of words for which all the annotators in a group of 5 annotators agreed was 109, i.e., lower than 5%, whereas the number of words for which there was no agreement was 2321, i.e., approximately 50%, indicating a low inter-annotator agreement.

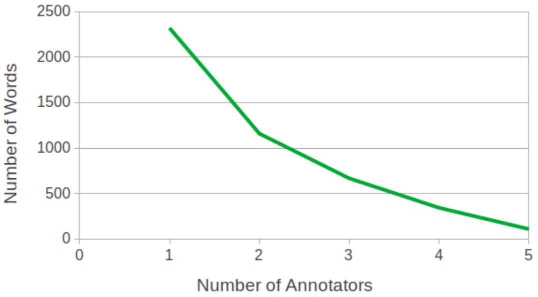

Figure 1: Distribution of 4599 words with respect to the number of annotators who agreed on a word being complex in Task 1

Figure 2 contains the description of the data obtained from Task 2. The annotation software presented the annotators with the word that they annotated as complex in task 1, along with the word's synonyms. They were required to rank the complexity of each word in this list. Since these ranked words also consisted of numbers, we cleaned the data by removing digits. We observed that annotators ranked the same word multiple times. Therefore we moved the duplicate instances by keeping only the first instance. We then selected only those words ranked by at least two annotators, as the label of a word was decided based on the average rank assigned to the word. Since the dataset consisted of corpus frequency as one of the features, we chose only those words that were present in our corpus. The annotators ranked 68,107 words. 52.1% words were assigned a rank of three or less, thus indicating that the words were complex, and 47.9% were assigned a rank greater than 3, indicating that the words were simple. After removing digits, duplicate instances, and words not present in the corpus, the dataset consisted of 7,321 words.

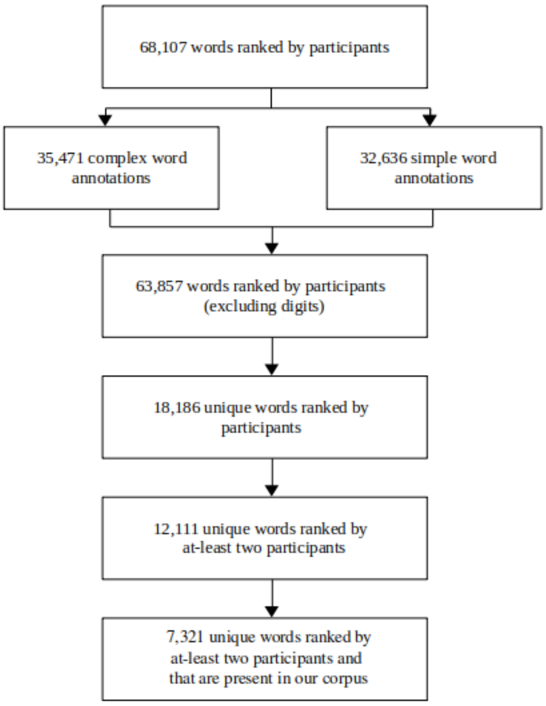
68,107 words ranked by participants
35,471 complex word annotations
32,636 simple word annotations
63,857 words ranked by participants (excluding digits)
18,186 unique words ranked by participants
12,111 unique words ranked by at-least two participants
7,321 unique words ranked by at-least two participants and that are present in our corpus

Figure 2: Process of filtering data for creating the dataset

We assigned a label of 1 to a sample if the average rank assigned to the word in Task 2, was less than or equal to 3, and 0, otherwise. The dataset contained 4,365 simple words and 2,956 complex words.

3.3. Classifier

We built a classifier that was trained on 5,111 records containing 40% complex words and 60% simple words. The classifier used a soft voting method that derived its prediction from multiple models such as random forest, ada boost, extra trees, XG boost, and gradient boosting. In order to answer RQ5, we created separate test sets that contained words ranked (in Task 2) by different categories of annotators. The categories were:

- native annotators
- non-native annotators
- annotators who were more comfortable with Hindi as compared to other known languages
- annotators who were comfortable with other known languages than with Hindi
- annotators who had high academic training
- annotators who had low academic training
- male annotators
- female annotators

Here, the years of academic training ranged from 1 to 16. We encoded it as a binary variable, which could be either high or low. A value less than 9 was considered low, and a value greater than eight was considered high.

4. Results and Discussion

This section contains the observations made with respect to the data collection as part of the annotation study, and the performance of a classifier on different types of datasets.
This section is divided into four subsections:

- Annotation tasks analysis that addresses RQs 1, 2 and 3
- Dataset, that contains an analysis of the values of features used in the dataset, which would lead us to the answer to RQ4
- Classifier Evaluation, that consists of the results of the tests performed on specialised subsets of the dataset, thus addressing RQ5
- Other Observations

4.1. Annotation Tasks

RQ1: Is there any difference between the native language of an annotator and the language that they were most comfortable reading?
72% annotators were more comfortable reading text of a language that was not their native language. Out of these, 93.05% annotators chose English as the language that they were most comfortable reading owing to the prevalence of the language in schools, universities and daily routine. We also observed that 70% of the annotators whose native language was Hindi, did not choose Hindi as the language they were most comfortable reading.
RQ2: Is there any relationship between the language that an annotator was most comfortable with and the official language of the region in which they spent the maximum number of years?
66% annotators chose English as the language they were most comfortable reading. However, 24% annotators chose the official language of the region they resided in, for the maximum number of years, which indicates that the region in which a person spends the maximum amount of time may have an influence on their language preferences.
RQ3: Is the lemma of a word considered to be simpler than a morphological variant of the word?

8,744 words were marked as complex in Task 1. The number of words that were marked as complex in Task 1 where the word was not in its lemmatised form, but was ranked simple in Task 2, where the word was present in its lemmatised form, was 2,424. A low proportion of 27.72% indicates that the word being its lemmatised form may not have had a significant effect on its perceived complexity.

We found that 70% of the 50 native speakers specified a language other than Hindi as the language they were most comfortable with. English, is considered the global lingua franca (Smokotin et al., 2014; House, 2014), was a common preference among many annotators. Since Hindi or a non-Hindi native language is overpowered by the use of English, specifically in an academic setting, many annotators were more comfortable in English. Around one-fourth of the annotators chose the region's official language as the language they were most comfortable with, although that was not their native language. Therefore even though we had native Hindi speakers as annotators, we could not assume that they were proficient in Hindi. Similarly, some annotators were non-native speakers who have studied Hindi for over eight years. With respect to annotations, the number of sentences annotated by non-native annotators was more. However, the average number of sentences annotated by a native annotator was more than the average number of sentences annotated by a non-native annotator. The average number of annotations across native as well as non-native speakers is comparable, which once again indicates the subjective and hard nature of the problem. Owing to these factors, we obtained a low inter-annotator agreement, which is also visible in Figure 1. Previous studies (Paetzold and Specia, 2016b) have also mentioned this observation wherein the non-native annotators had diverse linguistic backgrounds.

We observed no relationship between the annotators' number of years of academic training in Hindi and the number of words they annotated. However, regular readers of Hindi annotated fewer words than annotators who did not read Hindi content very often. Therefore we infer that reading habits may be a strong factor in identifying familiarity with the language instead of academic training.

We also found that in a context agnostic scenario, the complexity of a word does not change significantly when it is presented in its lemmatized form. Therefore we cannot assume that the morphological variation of a word should be considered as different from the word itself in such a scenario. However, the role of morphological variation in identifying complex words in a context-sensitive scenario needs to be explored further.

4.2. Dataset

RQ4: Is there a significant difference between the values of features of words labeled as complex and that of words labeled as simple?

Table 4.2 contains the mean and standard deviation (S.D.) of the features of the words labeled as simple and the words labeled as complex. Since the data were not normally distributed, we performed the Mann-Whitney U test (Mann and Whitney, 1947) to determine whether there is a significant difference among the values of each feature in both the categories, i.e., simple words and complex words. We observed that the differences in the values of all the features were statistically significant, with a p-value of 0.00. Since the differences were significant, this indicates that the features that were included as part of the dataset can be used to differentiate between simple and complex words.

Features	Mean		S.D.	
	Simple	Complex	Simple	Complex
Word Length	0.38	0.49	0.27	0.27
#Synsets	0.28	0.18	0.35	0.28
#Synonyms	0.33	0.26	0.32	0.29
#Consonants	0.38	0.49	0.3	0.29
#Vowels	0.43	0.52	0.29	0.29
#Hypernyms	0.26	0.52	0.33	0.29
#Hyponyms	0.23	0.15	0.34	0.29
#Consonant Conjuncts	0.23	0.31	0.35	0.37
#Syllables	0.40	0.50	0.31	0.30
Lemma Frequency	0.15	0.03	0.27	0.12

Table 1: Mean and standard deviation of the values of features of simple and complex words in the dataset

While creating the dataset, we used minmax normalisation based on the feature values of the target word's senses. Therefore the dataset is expected to closely model the real world scenario wherein a word is compared with related words as opposed to unrelated words. We calculated the average values of features and compared the values of words labelled as complex and those labelled as simple. We observed that values of features of complex words, such as length and others were slightly larger than that of words labelled as simple, and vice-versa for frequency, thus aligning with the observations from previous studies (Kauchak, 2016; Quijada and Medero, 2016).

4.3. Classifier Evaluation

RQ5: Is there a difference in the performance of the model w.r.t. test data created from annotations obtained from categories of users formed based on the following criteria?

- Native language
- Hindi being the language that they were most comfortable with
- Years of academic training in Hindi
- Self-identified gender

The performance of the classifier on different specialised datasets have been reported in Table 2.

Type of Dataset	AUC Score	F1 Score	% Complex Words	% Simple Words
Native Speakers	0.668	0.528	55.88	44.12
Non-Native Speakers	0.601	0.449	57.26	42.74
Annotators most comfortable with Hindi	0.699	0.581	54.59	45.41
Annotators most comfortable with another language	0.650	0.548	54.27	45.73
Annotators with high academic training	0.737	0.677	52.15	47.85
Annotators with low academic training	0.656	0.525	55.54	44.46
Female Annotators	0.709	0.611	53.71	46.29
Male Annotators	0.657	0.563	53.83	46.17

Table 2: Performance of the classifier on different types of datasets and the distribution of complex and simple words in each type of dataset

We evaluated the model using AUC and F1 scores. Though the model was not heavily biased towards an annotator category, we noticed that the scores were better for native speakers, annotators who were more comfortable with Hindi over other languages, annotators with high academic training, and female annotators. Although we divided the annotators into equal proportions of native and non-native speakers, we noticed that the native annotators' test set achieved slightly better performance than the non-native annotators' test set. This could be related to another observation that the inter-annotator agreement of the native speakers' group (0.193) was slightly more than that of the non-native annotators' group (0.179). The agreement values were obtained by calculating Krippendorff's alpha for the ten sentences that were assigned to each annotator. (Krippendorff, 2011) Therefore, the predictions of the ranks assigned by native annotators are more accurate than those of the ranks assigned by non-native annotators. We found similar results for the other groups, i.e., the group of annotators most comfortable with Hindi had a slightly better agreement coefficient (0.381) than the group of annotators who did not choose Hindi as their most comfortable language (0.158). Similarly, the agreement coefficient value of the annotators who were highly trained in Hindi (0.207) was better than the annotators who were not highly trained in Hindi (0.145). However, although the model performed better for annotations by female annotators, the difference between the inter-annotator agreement values of female annotators (0.189) and male annotators (0.187) was not significant. Therefore we cannot attribute the difference in the performance of models to the inter-annotator agreement values.

4.4. Other Observations

This subsection lists significant observations with respect to annotations that have not been mentioned in the previous subsections.

- The inter-annotator agreement of the native and non-native groups, calculated using Krippendorff's alpha (Krippendorff, 2011), were 0.2421 and 0.1143, leading to an average of 0.1782. This value was expected to be low due to the annotators' diverse backgrounds and vocabulary.
- 5,213 sentences were annotated, non-native annotators annotated 2,768 sentences, and native readers annotated 2,445 sentences. On average, 55.36 sentences were annotated by non-native readers, and native readers annotated 98.82 sentences. This contradicts the assumption that native readers would find the text more comprehensible than non-native readers.
- The total number of words annotated by the native annotators was 3,911, whereas that annotated by non-native annotators was 4,645. Although the minimum number of annotations belongs to the group with native annotators, the range of annotations is comparable across both types of groups – 215 for the native annotators and 195 for the non-native annotators. The average number of annotations made by non-native annotators was 84.82, whereas that made by native annotators was 86.3, suggesting similar vocabulary limitations among both groups.
- There was no correlation between the total number of words annotated in task 1 and the age of the annotators (r = -0.121).
- The average number of annotations made by annotators who identified themselves as females was 86.233, whereas those who identified themselves as males, was 85.228.
- Annotators who read content in Hindi daily annotated an average of 78.67 words, whereas annotators who did not read content in Hindi daily annotated an average of 86.61 words.
- There was no significant relationship between the number of years spent studying in Hindi as part of the school curriculum and the number of words annotated as complex in task 1 (r = 0.203).

The study's objective was to create a resource for complex word identification in Hindi text. As the first step, we aimed to conduct an annotation task to create a dataset containing words labeled as complex or simple, based on lexical and semantic features and the word's frequency. We hired native and non-native annotators, where native annotator refers to a person who learned to speak the language of the place where he or she was born as a child rather than learning it as a foreign language. We found that the problem of complex word identification is hard owing to the varying exposure of individuals to the language. We also found that the years of academic training in a language may not be relied upon solely in determining an individual's familiarity with words in the language, although regular reading could be a significant factor in improving the vocabulary. We also noticed these variations in the classifier's performance that was tested using different specialized datasets. We also observed that the values of features included in the dataset were significantly different for simple and complex words, thus justifying the inclusion of the features in the dataset.

5. Conclusion

We present a dataset for context-agnostic complex word identification in Hindi. The dataset contained 7321 words and was annotated by 100 annotators. Our goal was to create the dataset and understand different users' perceptions of complex words. Our analysis shows that in an Indian context, where there are numerous regional languages, the native language of a person may not be their language of preference. The preferred language may depend on the language spoken in the region they spent the maximum number of years in, not solely on the years of academic training in a particular language. We observed a notable difference among the feature values of a simple and complex word, indicating the need to explore the role of these features. We also designed a model based on normalized features relative to the target word and its synonyms. The predictions of the model were derived from tree-based ensemble models. Our model analysis shows that though there is no heavy bias towards a specific category of users, there are notable differences among different categories. We found that categories of users who were more familiar with the language with respect to its use in their daily lives had slightly better agreement values, and therefore the test sets created by collating the annotations of these annotators gave a slightly better performance as compared to the annotations obtained from their counterparts in each category.

The dataset is freely available at `https://zenodo.org/record/5229160` and can be used further to create a lexical simplification system that would display recommendations of simpler words given a complex word.

6. Acknowledgements

This study is funded by Symbiosis Centre for Research and Innovation, Symbiosis International (Deemed University), and received approval from the Independent Ethics Committee, Symbiosis International (Deemed University). We would like to thank the organizations that provided the corpora that were used for fetching feature values of words present in the dataset (OPU, nd), (Wikipedia contributors, 2019), (ILCI Consortium, JNU, 2015), (EILMT Consortia, CDAC Pune, 2017), (ILCI Consortium, JNU, 2016), (ILCI-II, JNU, 2017), (Kunchukuttan et al., 2018), (Ramamoorthy et al., 2019), (Kakwani et al., 2020), along with the Stanford NLP group that released the StanfordNLP package for Python which was used to fetch accurate lemmas of the words in Hindi (Qi et al., 2019). The "Linguistic Resources" obtained from TDIL have been developed made available by TDIL, MeitY, Government of India. The resources from CFILT, IIT-Bombay, have been downloaded from http://www.cfilt.iitb.ac.in/Downloads.html. We are grateful to Dr. Matthew Shardlow and Dr. Fernando Alva-Munchego for providing valuable feedback. RS has been supported by the European Union Horizon 2020 research and innovation programme under grant 825153 (EMBEDDIA).

7. References

Aluísio, S. and Gasperin, C. (2010). Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

Billami, M. B., François, T., and Gala, N. (2018). Resyf: a french lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581.

Bott, S., Rello, L., Drndarević, B., and Saggion, H. (2012). Can spanish be simpler? lexsis: Lexical simplification for spanish. In *Proceedings of COLING 2012*, pages 357–374.

Brants, T. (2006). Web 1t 5-gram version 1. *http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp? catalogId= LDC2006T13.*

Candido Jr, A., Maziero, E. G., Specia, L., Gasperin, C., Pardo, T., and Aluisio, S. (2009). Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42.

Caplan, D. (1992). *Language: Structure, processing, and disorders.* The MIT Press.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artifi-*

cial Intelligence and Assistive Technology, pages 7–10. Citeseer.

Carroll, J. A., Minnen, G., Pearce, D., Canning, Y., Devlin, S., and Tait, J. (1999). Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270.

Coster, W. and Kauchak, D. (2011). Simple english wikipedia: a new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669.

De Belder, J. and Moens, M.-F. (2010). Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

De Belder, J. and Moens, M.-F. (2012). A dataset for the evaluation of lexical simplification. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 426–437. Springer.

Finnimore, P., Fritzsch, E., King, D., Sneyd, A., Rehman, A. U., Alva-Manchego, F., and Vlachos, A. (2019). Strong baselines for complex word identification across multiple languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 970–977.

Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). Simplex-pb: A lexical simplification database and benchmark for portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.

Henry, S. L. (2006). Understanding web accessibility. In *Web Accessibility*, pages 1–51. Springer.

Hmida, F., Billami, M., François, T., and Gala, N. (2018). Assisted lexical simplification for french native children with reading difficulties. In *The Workshop of Automatic Text Adaptation, 11th International Conference on Natural Language Generation*.

Horn, C., Manduca, C., and Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 458–463.

House, J. (2014). English as a global lingua franca: A threat to multilingual communication and translation? *Language Teaching*, 47(3):363–376.

Kajiwara, T. and Yamamoto, K. (2015). Evaluation dataset and system for japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40.

Kauchak, D. (2016). Pomona at semeval-2016 task 11: Predicting word complexity based on corpus frequency. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1047–1051.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Lee, J. S. and Yeung, C. Y. (2018). Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Maddela, M. and Xu, W. (2018). A word-complexity lexicon and a neural readability ranking model for lexical simplification. *arXiv preprint arXiv:1810.05754*.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

McCarthy, D. and Navigli, R. (2009). The english lexical substitution task. *Language resources and evaluation*, 43(2):139–159.

Mishra, K., Soni, A., Sharma, R., and Sharma, D. M. (2014). Exploring the effects of sentence simplification on hindi to english machine translation system. In *Proceedings of the Workshop on Automatic Text Simplification-Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 21–29.

Ogden, C. K. (1930). Basic english: A general introduction with rules and grammar.

Paetzold, G. and Specia, L. (2016a). Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080.

Paetzold, G. and Specia, L. (2016b). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California, June. Association for Computational Linguistics.

Paetzold, G. and Specia, L. (2016c). Understanding the lexical simplification needs of non-native speakers of english. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727.

Paetzold, G. H. and Specia, L. (2017). A survey on lexical simplification. *Journal of Artificial Intelligence Research*, 60:549–593.

Paetzold, G. H. (2016). *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2019). Universal dependency parsing from scratch. *arXiv preprint arXiv:1901.10457*.

Quijada, M. and Medero, J. (2016). Hmc at semeval-2016 task 11: Identifying complex words using depth-limited decision trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1034–1037.

Shardlow, M., Evans, R., and Zampieri, M. (2021).

Predicting lexical complexity in english texts. *arXiv preprint arXiv:2102.08773*.

Shardlow, M. (2013). The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77.

Shardlow, M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Siddharthan, A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Smokotin, V. M., Alekseyenko, A. S., and Petrova, G. I. (2014). The phenomenon of linguistic globalization: English as the global lingua franca (eglf). *Procedia-Social and Behavioral Sciences*, 154:509–513.

Soni, A., Jain, S., and Sharma, D. M. (2013). Exploring verb frames for sentence simplification in hindi. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1082–1086.

Specia, L., Jauhar, S. K., and Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Štajner, S., Saggion, H., and Ponzetto, S. P. (2019). Improving lexical coverage of text simplification systems for spanish. *Expert Systems with Applications*, 118:80–91.

Venugopal-Wairagade, G., Saini, J. R., and Pramod, D. (2020). Novel language resources for hindi: An aesthetics text corpus and a comprehensive stop lemma list. *arXiv preprint arXiv:2002.00171*.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017a). Cwig3g2-complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017b). Multilingual and cross-lingual complex word identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 813–822.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G. H., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

8. Language Resource References

Pushpak Bhattacharyya and Prabhakar Pande and Laxmi Lupu. (2008). *Hindi WordNet*.

EILMT Consortia, CDAC Pune. (2017). *Hindi-English Agriculture Entertainment Text Corpus ILCI-II*.

ILCI Consortium, JNU. (2015). *Hindi-English Health Text Corpus-ILCI*.

ILCI Consortium, JNU. (2016). *English-Hindi Tourism Text Corpus – EILMT*.

ILCI-II, JNU. (2017). *Hindi Monolingual Text Corpus ILCI-II*.

Divyanshu Kakwani and Anoop Kunchukuttan and Satish Golla and Gokul N.C. and Avik Bhattacharyya and Mitesh M. Khapra and Pratyush Kumar. (2020). *IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages*.

Kunchukuttan, A. and Mehta, P. and Bhattacharyya, P. (2018). *The IIT Bombay English-Hindi Parallel Corpus*.

(n.d.). *The Open Parallel Corpus*.

Ramamoorthy, L. and Narayan Choudhary and Jitendra Kumar Singh and Richa and Anjali Sinha and Dheeraj Kumar Mishra and Arimardan Kumar Tripathi and Aditi Debsharma and Satyaendra Kumar Awasthi and Madhupriya Pathak. (2019). *A Gold Standard Hindi Raw Text Corpus*. Central Institute of Indian Languages, Mysore.

Wikipedia contributors. (2019). *Wikimedia Downloads*. Wikimedia.