Word Complexity is in the Eye of the Beholder

Sian Gooding
Dept of Computer Science and Technology
University of Cambridge
shg36@cam.ac.uk

Ekaterina Kochmar
Dept of Computer Science
University of Bath
ek762@bath.ac.uk

Seid Muhie Yimam
Language Technology group
Universität Hamburg, Germany
yimam@informatik.uni-hamburg.de

Chris Biemann
Language Technology group
Universität Hamburg, Germany
biemann@informatik.uni-hamburg.de

Abstract

Lexical complexity is a highly subjective notion, yet this factor is often neglected in lexical simplification and readability systems which use a "one-size-fits-all" approach. In this paper, we investigate which aspects contribute to the notion of lexical complexity in various groups of readers, focusing on native and non-native speakers of English, and how the notion of complexity changes depending on the proficiency level of a non-native reader. To facilitate reproducibility of our approach and foster further research into these aspects, we release a dataset of complex words annotated by readers with different backgrounds.

1 Introduction

Complex word identification (CWI) is the first step in a lexical simplification (LS) pipeline, concerned with identification of words in text that are in need of further simplification (Shardlow, 2013). For instance, in example (1) a CWI system might identify *engulfed* as a complex word, which would allow an LS system to replace it with a simpler alternative, e.g. *flooded*, in the next step (Paetzold and Specia, 2016a; Gooding and Kochmar, 2019b):

(1) Water *engulfed* Beringia.
↓
Water *flooded* Beringia.

It has been shown that accurate CWI can significantly reduce errors in simplification (Shardlow, 2014), thus improving the quality of an LS system output (Lee and Yeung, 2018). In addition, CWI has been shown to be an important component in readability assessment systems (Maddela and Xu, 2018) and in vocabulary acquisition modules of educational applications (Zaidi et al., 2020). However, an important aspect of CWI and LS that is often neglected is that text complexity is not an objective notion homogeneous across various target populations: what is challenging for a reader with a particular background (for example, a non-native reader at a lower level of language proficiency) would not necessarily be challenging for readers with other backgrounds (for example, more proficient readers) (Bingel, 2018). A number of factors may contribute to that, including the reader's age and level of language proficiency, among others (Paetzold and Specia, 2016c). LS systems often aim to address the needs of specific reader populations, such as children, non-native speakers, or readers with particular cognitive impairments. Thus, personalization in LS typically results in specialized simplification tools aimed at certain groups of readers (Carroll et al., 1998; Rello et al., 2013; Evans et al., 2014), with only a few systems addressing adaptation to the readers' needs in a more dynamic way (Bingel et al., 2018; Yimam and Biemann, 2018a,b; Scarton and Specia, 2018).

Despite CWI being one of the key steps in an LS pipeline in need of adaptation to readers' profiles, this is rarely addressed in practice (Lee and Yeung, 2018; Bingel, 2018). For instance, existing and widely used datasets on CWI present a homogeneous view on word complexity, merging annotations from various groups of readers (Paetzold and Specia, 2016c; Yimam et al., 2018). From the cognitive perspective, little is still known about the challenges that particular readers face when developing their reading skills and about the factors contributing to their vocabulary acquisition.

In this paper, we investigate factors focusing on the two key background aspects in the development of reading abilities: *whether a reader is a native speaker* of the language, and if not, *what is the reader's level of language proficiency*. We use the data from Yimam et al. (2017a), which contains English sentences where complex words are annotated by native and non-native speakers of English, spanning three different levels of language proficiency. We investigate which aspects contribute to the notion of lexical complexity for readers with

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4439–4449
June 6–11, 2021.

different backgrounds and how the notion of complexity changes depending on the proficiency levels of the non-native readers.

In our paper we make the following contributions:

- We show that the best models for predicting complexity are trained using the annotations of the target audience.
- We perform feature analysis by observing which correlate most with the notion of complexity for native and non-native audiences.
- We analyse the distribution of features for complex words across differing proficiency levels.
- Finally, we release a CWI dataset annotated by readers with different backgrounds.

2 Background

2.1 Models of Complex Word Identification

CWI was established as an essential step in LS in Shardlow (2013), which demonstrated that without this step, LS systems tend to over- or under-simplify, thus rendering the output less useful for the readers. Early approaches to this task considered simplification of all words (Devlin and Tait, 1998; Bott et al., 2012) and use of frequency-based thresholds (Zeng et al., 2005; Biran et al., 2011), however Shardlow (2013) shows that classification algorithms are more precise in identification of complex words than both these approaches. Recent shared tasks on CWI (Paetzold and Specia, 2016c; Yimam et al., 2018) helped it gain popularity in the NLP community as they provide researchers with shared data and benchmarks. Most systems participating in the shared tasks addressed CWI with classical machine learning algorithms, with the best-performing systems using ensemble-based approaches. Current state-of-the-art results on CWI are achieved by a sequence-labeling model of Gooding and Kochmar (2019a), however models of such type are less easily interpretable.

2.2 Aspects of word complexity

The question of what contributes towards the notion of word complexity has been investigated before, for example in readability studies. *Word length* is commonly believed to correlate with text complexity and is included as a component in a wide range of readability formulas (Dale and Chall, 1948; Kincaid et al., 1975; Dubay, 2004). *Frequency*, another factor often considered in readability and text simplification approaches (Rudell, 1993; De Belder and Moens, 2010), was shown to correlate and cause word *familiarity*, which in its turn contributes to higher word recognition and lower reaction times (Connine et al., 1990; Morrel-Samuels and Krauss, 1992). Notably, word length and frequency have been widely used in CWI systems, and are reported to be good, cross-linguistic predictors of complexity (Bingel and Bjerva, 2018). Other factors considered important for word complexity include a variety of psycholinguistic properties, including word's *age of acquisition*, *concreteness*, and *imagability* (Carroll and White, 1973; Zevin and Seidenberg, 2002; Begg and Paivio, 1969). At the same time, not all factors are equally applicable to all groups of readers: for instance, while frequency may be an important factor for second language learners, other populations may be more affected by the length of a word or the occurrence of certain character combinations (Rudell, 1993; Rello et al., 2013). Yet, little is still known about the factors contributing to word complexity for native vs non-native readers as well as for non-native readers at different levels of language proficiency.

3 Data

The most comprehensive CWI dataset to date was released by Yimam et al. (2017a) and further used in the CWI shared task 2018 (Yimam et al., 2018). This dataset has been annotated for complex words across a number of languages, including English, German, and Spanish. In this paper, we use the English portion of the data with the information about annotators' backgrounds[1]. The dataset contains texts from 3 different sources: professionally written news articles (NEWS), amateurishly written news articles (WIKINEWS), and WIKIPEDIA articles. The annotation was performed using the Amazon Mechanical Turk platform, where a total of 20 annotators, 10 native speakers and 10 non-native speakers, were asked to mark words that they deemed complex for a given target readership, particularly *children*, *language learners*, and *people with reading impairments*. The workers were presented with text, consisting of 5 to 10 sentences (Figure 1), and were asked to select lexical items that they found complex (Figure 2). Workers use

[1] CWI Dataset with Language levels

their mouse pointer to highlight the complex units. The complex words or phrases included content words (e.g., nouns, verbs, adjectives, and adverbs) and phrases up to 50 characters in length. In this dataset, the complex units are considered if they are selected by at least one worker (Yimam et al., 2017a,b). Non-native speakers of English were asked to report their proficiency levels (beginner, intermediate, advanced). For our experiments, we concentrate on complex words only and disregard complex phrases. A break-down of proficiency labels for words (across all genres) is presented in Table 5, with label 1 denoting complex words and label 0 used for non-complex words. It is worth noting that the groups of annotators labelling portions of the dataset were not fixed. Within each group, the proficiency distribution varied, with some containing no annotators from a given class.

4 Method

We firstly show that when predicting word complexity, the needs of sub-groups differ and are best predicted using models targeting them specifically. We demonstrate that the best performing models for a sub-group are trained with the annotations of that group using a classical machine learning approach. Secondly, we analyse the correlation of features with the number of annotators who found the word complex for both native and non-native groups. Finally, we investigate how the distributions of features vary for words marked as complex across audiences.

4.1 Complexity Features

To gain fundamental insights into the performance across proficiency groups, we run experiments using the CAMB system by Gooding and Kochmar (2018) as it achieved the best results across all binary and two probabilistic tracks in the CWI 2018 shared task (Yimam et al., 2018). Furthermore, the code for this system has been made publicly available by the authors. The CAMB system relies on 27 features in total. Feature types include lexical, syntactic, frequency-based and other aspects of information about individual words, outlined below.

Lexical Features: For each target word, the word itself as well as the length and number of syllables (obtained using the *Datamuse* API) is included. Additionally, the number of senses, hypernyms and hyponyms are collected for the word lemma using WordNet (Fellbaum, 2005). Finally, the number of phonemes for the word are included sourced from the MCR Psycholinguistic Database (Wilson, 1988).

POS & Dependency Parse Relations: The target sentence is parsed using the NLPCore pipeline. Following this, the number of dependency relations are counted to produce a feature. The part-of-speech tag for the word is additionally included.

List-Based Features: A set of binary features are used that indicate the presence of the target word in a given list. The source of each list is outlined below:

- *SubIMDB*: using the SubIMDB corpus (Paetzold and Specia, 2016b), the word frequencies are calculated from the '*Movies and Series for Children*' section. The top 1,000 most frequent words are then included.
- *Simple Wikipedia (SimpWiki)*: a list of the top 6,368 words contained in the Simple Wikipedia (Coster and Kauchak, 2011).
- *Ogden's Basic English*: the top 1,000 words from Ogden's Basic English list (Ogden, 1968).
- *Cambridge Advanced Learner's Dictionary (CALD)*:[2] the entries contained in the Cambridge Advanced Learner's Dictionary.

Word Frequency: The frequency of the target word is estimated using the Google dataset of n-grams (Goldberg and Orwant, 2013). Additionally, the Thorndike-Lorge written frequency derived from Thorndike and Lorge (1944) is obtained from the MCR Psycholinguistic Database (Wilson, 1988).

Psycholinguistic Features: Finally, the following features are extracted from the MCR Psycholinguistic Database (Wilson, 1988):

- *Word familiarity rating (FAM)*
- *Imagability rating (IMG)*, representing the ease of associating the word with an image.
- *Concreteness rating (CNC)* represents the degree to which the word refers to a tangible entity, based on the norms of Gilhooly and Logie (1980).
- The *number of categories (*KFCAT*)* and *samples (*KFSMP*)* are derived from Kučera and Francis (1967).
- *Age of acquisition (*AOA*)* is based on the norms of Gilhooly and Logie (1980)

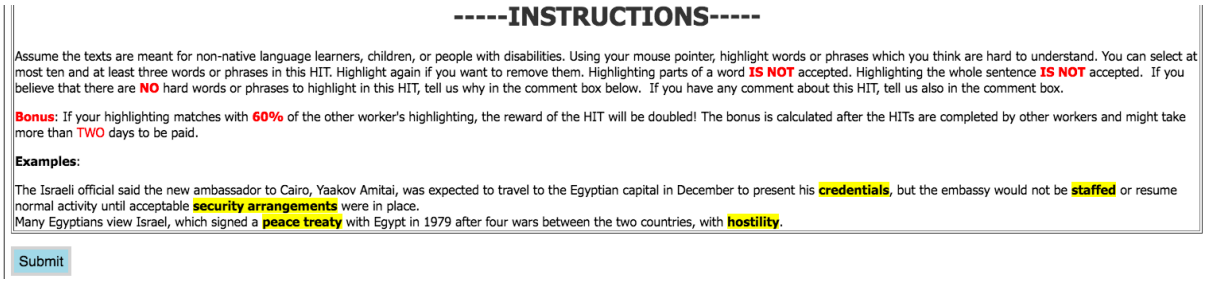


Figure 1: Complex word identification instruction with examples

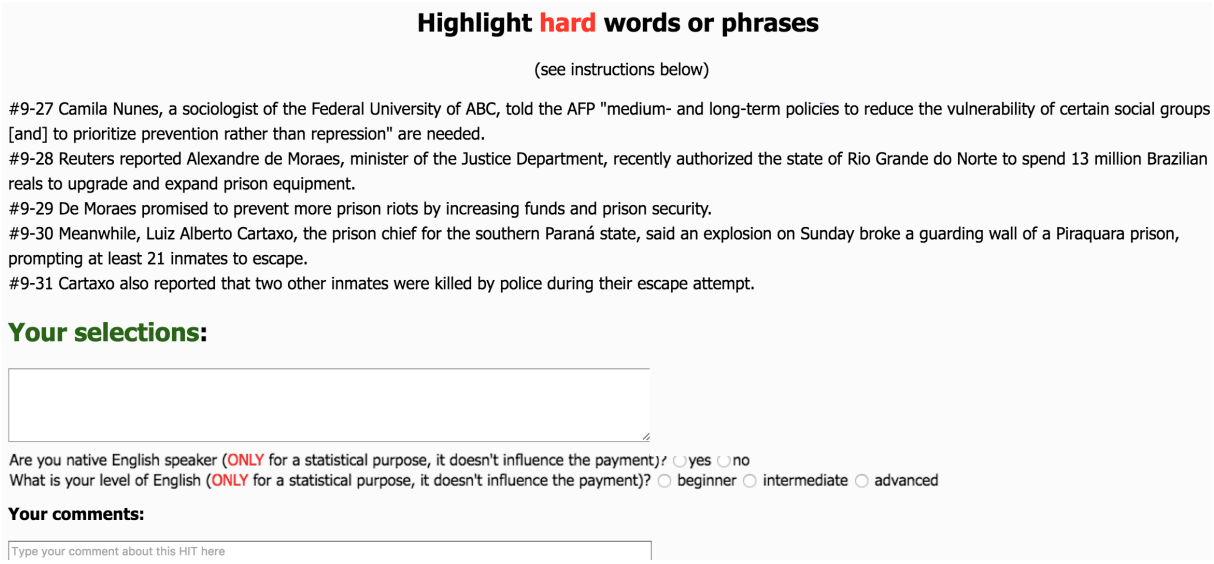


Figure 2: Complex word identification annotation interface

4.2 Experimental Framework

The CAMB system uses the `sklearn` machine learning framework[3] and achieves best results using an ensemble of algorithms. In our experiments, we use the `logistic regression` classifier as this was the best performing classifier for proficiency prediction due to the reduced number of annotations. As shown in Table 5, the number of annotations for each subgroup varies and the ratio of non-complex to complex words is highly skewed. For the data in our experiments, we firstly convert all proficiency annotations to a binary format, where if at least one annotator has marked the word as complex the word is given a binary label of 1. For our initial experiments the aim is to see if the needs of a proficiency group are best predicted by that target group. In order to make a fair comparison, we control for the number of binary annotations by restricting all groups to the same amount of labels as in the beginner class $(2,263)$. The annotations are ordered by the highest class agreement and the top $2,263$ values are selected. Additionally, we remove 20% of non-complex labels, where no proficiency groups had marked the word as complex, to re-balance the class distribution to that of the original binary shared task. This resulted in a dataset containing $9,828$ non-complex words and $4,423$ words marked with at least one proficiency annotation. Stratified 5-fold cross-validation was used resulting in a test size of $2,850$ and total training size of $11,400$ per fold.

5 Results

In all experiments, 5-fold stratified cross validation is performed and the average scores across folds presented. Table 1 shows the results of training the system using the annotations of one proficiency subgroup and the subsequent model performance across subgroups. Columns represent the training annotations used and the rows represent the results on the respective test sets. As a result of the small training size, the overall F1-SCORE achieved across classes is low. For instance, when all avail-

[2] Publicly available here
[3] `http://scikit-learn.org/stable/`

	Training Data								
	Beginner			*Intermediate*			*Advanced*		
Test	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Beginner	0.649	0.245	**0.356**	0.425	0.289	0.344	0.433	0.270	0.333
Intermediate	0.529	0.201	0.291	0.669	0.452	**0.538**	0.596	0.423	0.494
Advanced	0.513	0.196	0.283	0.594	0.398	0.477	0.659	0.476	**0.552**

Table 1: Results of models trained and tested with differing proficiency labels

	Training Data					
	Native			*Non-native*		
Test	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Native	0.794	0.801	**0.797**	0.761	0.796	0.773
Non-native	0.766	0.730	0.748	0.785	0.792	**0.788**

Table 2: Results of models trained and tested with native and non-native annotations

	Native		
Test	Precision	Recall	F1-Score
Beginner	0.232	0.789	0.359
Intermediate	0.539	0.794	0.642
Advanced	0.623	0.803	**0.702**

Table 3: Results of model trained with native annotations across non-native proficency

	Binary Labels	
	1	0
Beginner	2,263	27,433
Intermediate	5,203	24,493
Advanced	5,849	23,847

Table 5: Binary label distribution for words per proficiency class, *1* is complex and *0* is simple.

	Prec	Rec	F1-Score
$Beginner_{(2263)}$	0.62	0.22	0.33
$Intermediate_{(5203)}$	0.80	0.80	0.80
$Advanced_{(5849)}$	0.76	0.77	0.78

Table 4: Results showing Precision, Recall and F1-Score using all sub-group annotations

able labels are used for *intermediate* and *advanced* classes an F1-Score of over 75% is achieved as shown in Table 4. However, the results are still highly informative, as we observe that in all cases the best F1-Score is obtained when the original sub-group annotations are used. This finding supports the case that the needs of such sub-groups differ and are best predicted using models targeting them specifically. The Precision, Recall and F1-Score across all categories are best when the model is trained using the annotations of the target subgroup. The only exception is Recall for *beginner*, where the *intermediate* and *advanced* models perform the best (results underlined). However, it is worth noting that if an intermediate or advanced learner considers a word to be complex, it is highly likely that a beginner will too. This observation is further supported by the finding that whilst the *advanced* and *intermediate* models perform adequately on the beginner test set, the *beginner* model performs very poorly when predicting the needs of intermediate or advanced users. The *advanced* and *intermediate* models achieve higher F1-Scores than the *beginner* model. These results support the case that beginner word acquisition is more idiosyncratic than at an intermediate or advanced level where the concept of word complexity converges.

Table 2 additionally shows that the complex annotations of a subgroup are the best predictors for that class. We observe that the best results for the native group occur when trained with native only annotations and the same holds for the non-native class.

We perform experiments by training with native complexity annotations and observe the performance across non-native proficiency groups. The results of these are shown in Table 3, and as there is a larger training set the scores are higher than those in Table 1. We see that the native annotations perform best when predicting the advanced non-native word complexities. However, this is not the case for the beginner class. We also observe a pattern in native annotations being preferential for higher

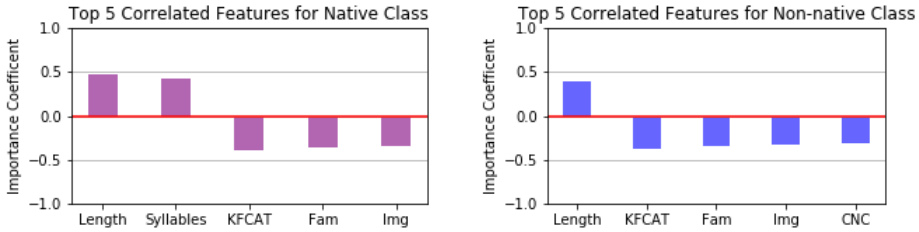


Figure 3: Graphs showing the top 5 correlated features against the absolute number of annotations for the native and non-native classes all values are significant ($N = 17250; p < .001$)

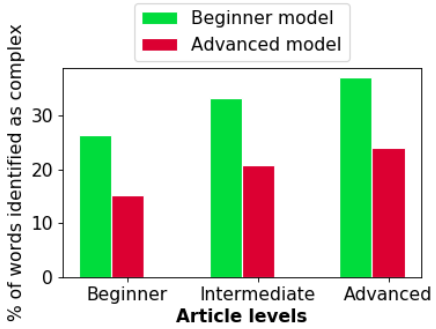


Figure 4: The average percentage of complex words as identified by CWI models trained with advanced and beginner annotations on the Newsela dataset

proficiency levels.

Newsela results

(2) His frequent use of ***prepositions*** suggests he was ***rigorously*** educated in *grammar*.

(3) The way he wrote shows he was very educated in *grammar*.

We apply our beginner and advanced CWI models on an additional dataset, Newsela.[4] Newsela contains articles which are rewritten by professional editors at differing levels of simplicity with each grade level as defined by the Common Core Standards (Porter et al., 2011). We take the highest, intermediate and lowest level of each article and perform CWI using the models trained with all advanced and beginner annotations. Our aim is to see if these models are able to differentiate between levels as CWI has been shown to be an important component in readability assessment systems (Maddela and Xu, 2018). In Figure 4, we see that the model trained with the annotations from beginners identifies a higher percentage of words as complex across levels when compared to the advanced model. Additionally, both models identify more complex words in the advanced texts than in the intermediate or beginner. These results show that models trained for specific audiences can result in a different concept of complexity. For instance, examples 2 and 3 show a sentence from an advanced and simplified article. Words in bold are identified as complex by the **advanced** model and italicised if found complex by the *beginner* model. We see that in the higher level sentence (2), two words are identified as difficult by both models and one word is identified as complex by only the beginner model. In the lower level article, the words identified as complex by both models have been simplified. This results in only one word being identified as complex by the model tailored for beginners. We know that text begins to be accessible for non-native readers if they are familiar with at least 90% of word content (Nation, 2006). Therefore, being able to model text understanding across audiences relies on audience specific models of word complexity as demonstrated in our example.

Feature Correlations

As the absolute number of native and non-native annotators remained constant across annotations (i.e. 10), we explore the feature correlations for these subgroups. For instance, the word *vowed* in a given context has been marked as complex by 10 non-native and 1 native annotator. This indicates that the word might be more challenging for a non-native audience than for native in the given context.

Figure 3 shows the highest correlated features for the native and non-native groups, all of which are significant ($p < .001$). Overall, the correlations for the native class are higher than for non-native which is likely due to a more united perspective of complexity. This follows as individuals with a similar first language or educational background are more likely to annotate the same words as complex (Specia et al., 2012).

For both classes, the feature with the highest

[4] https://newsela.com

correlation is that of word length: the positive correlation shows that the longer the word, the more likely it will belong to the complex class. Following this, for the native class we see that the number of syllables is second. Whilst the length of a word and the number of syllables are highly correlated (0.64), it is interesting to note that the number of syllables correlates more highly with the native notion of complexity than for non-native. This may be explained by the fact that syllable and phoneme awareness plays an independent role in the processing of text (Engen and Høien, 2002). This impact is especially pronounced in lower skilled readers, where due to a reduced vocabulary set, the development of precise phonological representations are not yet formed (Elbro, 1996).

For the non-native class, the second highest correlated feature is *KFCAT* which represents the number of categories of text in which the word was present as given in the norms of Kučera and Francis (1967). The negative correlation shows that the more categories of text a word appears in, the less likely it is to be considered complex. This measure can also be considered as the specificity of the word. For instance, we see that the word *grounds* is found across a wide range of text categories and is rarely considered complex. Whereas words like *altimeter* and *aneroid*, which are highly specific to a particular domain, are considered complex in all contexts by both native and non-native readers. The number of categories that a word occurs in is correlated with the word's frequency (0.35). However, when you control for the word frequency, the effect of this correlation is even higher: -0.40 and -0.41 for non-native and native respectively. Therefore, the narrower the scope of application for a word the more likely it will be considered difficult.

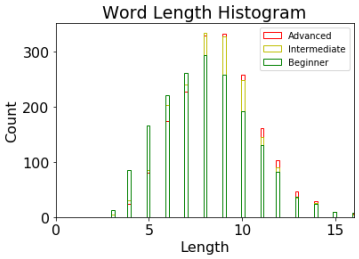
Finally, we see that psycholinguistic measures such as the word familiarity and imagability are highly correlated with both the native and non-native absolute number of annotations. When considering imagability, the larger the *img* score the higher the imagability, for instance '*dog*' has a high *img* factor whereas '*decision*' has a low score as it cannot be easily associated with an image. The negative correlation shows that the higher the score the less likely the word is considered complex. Intuitively, it makes sense why this feature would be influential in determining word complexity. In fact, research on children's reading has shown that words high in imagability are easier to read than words low in imagability (Coltheart et al., 1988). It has been suggested that this occurs because low imagability words are acquired later in life than high imagability words. Finally, concreteness is one of the top five features correlated with the non-native annotations. It has been found that the higher the concreteness of a word, the more likely it is to be comprehensible (Sadoski et al., 2000).

Feature Distributions

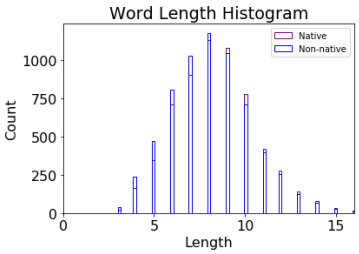
Word length and frequency have been widely used in CWI systems and are reported to be good cross-linguistic predictors of complexity (Bingel et al., 2018). Additionally, psycholinguistic properties are considered important in word complexity estimation (Carroll and White, 1973). When investigating the feature importance for our binary models in Section 5, we find that the features with the highest importance across models are word length, frequency and imagability. We investigate whether the distribution of the feature values is dependent on the intended audience.

Figure 5 contains two histograms presenting binned word lengths across proficiency classes. Words that have been marked as complex are grouped into 20 bins and the distribution of lengths plotted. We observe that beginners mark more shorter words as complex than either the intermediate or advanced class do. Generally, the distribution of lengths shifts to the right as proficiency increases. This same pattern is observed for the native and non-native classes, where non-native annotators are more likely to mark shorter words as complex than native.

Figure 6 contains histograms presenting the binned frequencies for complex words (20 bins). For frequencies, we observe a clear difference between the beginner and intermediate/advanced classes. The beginner sub group has marked many more low frequency words as complex. For the advanced class, the range (difference in largest and smallest frequency value) is 259 whereas for beginners the range is 569. Furthermore, the mean frequency values show that the advanced and intermediate classes, on average, are more likely to consider words with lower frequencies to be complex (15.09 and 16.22) whereas for beginners the mean is higher (22.63). As the advanced and intermediate classes have a narrower spread and lower mean, it is likely frequency based thresholding techniques would work well for these groups.

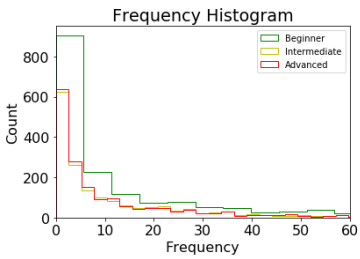


(a)

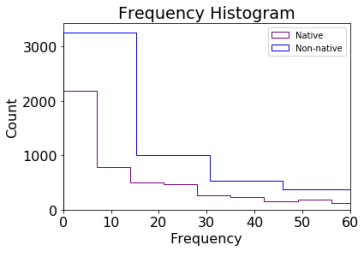


(b)

Figure 5: Word length histograms with 20 bins

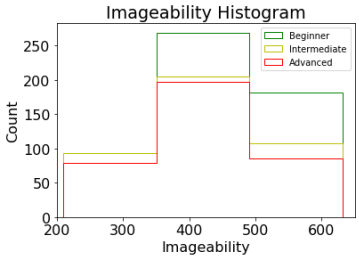


(a)

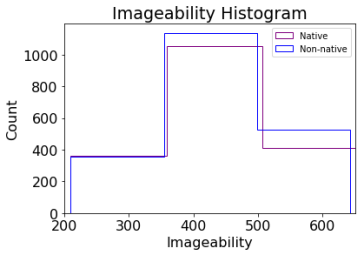


(b)

Figure 6: Frequency histograms with 20 bins



(a)



(b)

Figure 7: Imagability histograms with 3 bins

When we consider the native and non-native frequency distributions, we notice the same pattern emerging between classes. The non-native class has many more low frequency words annotated as complex and the relationship between native and non-native closely resembles the one between advanced and beginner. Word frequency provides signal on the likelihood of an individual being exposed to the word. However, the actual likelihood of exposure will depend on whether an individual is a native or non-native speaker as well as their experience of the language.

Finally, in Figure 7 we group imagability ratings into 3 bins representing high, medium and low scores. We see that for the advanced and intermediate classes most complex annotations fall in the middle range. However, for the beginner class there are still many high imagability words that are deemed as complex. It is worth noting, that the coverage of imagability is limited and therefore results should be considered more cautiously. Regarding the native and non-native imagability, we again see that the non-native class has slightly more higher imagability words marked as complex.

To conclude, the relative relationships between beginner and advanced feature distributions very closely mirror the relationship between native and non-native. There is a clear trend across features based on the proficiency and experience the reader. Furthermore, the feature profiles of advanced non-native speakers are more similar to that of a native speaker. As far as we are aware, this is the first work exploring how the thresholds of features vary across audiences for complexity. Investigating this is insightful, as there are numerous threshold based approaches to CWI (Zeng et al., 2005; Elhadad, 2006; Biran et al., 2011), therefore understanding how these thresholds differ for audiences can produce more informed techniques.

6 Conclusions

Textual complexity is a subjective phenomenon that is dependent on the intended audience. We show that when considering lexical complexity, the best performing CWI models for a target proficiency level are trained with the labels of that sub-group. We investigate which features correlate most with the absolute number of native and non-native annotations as well as observe how the distributions of classic complexity features are dependent on the intended audience. We find strong similarities between the notion of word complexity for advanced non-native readers and native readers. Finally, we release a dataset for CWI with proficiency subgroup annotations. In future work we plan to collect additional annotations across classes, especially concentrating on beginners. We would also like to investigate how effective informed-thresholding techniques for CWI are compared to high resource systems.

Acknowledgments

This work has been done while the second author was a Senior Research Associate at the University of Cambridge. We thank Cambridge English for supporting this research via the ALTA Institute. We are also grateful to the anonymous reviewers for their valuable feedback.

References

Ian Begg and Allan Paivio. 1969. Concreteness and imagery in sentence meaning. *Journal of Verbal Learning and Verbal Behavior*, 8(6):821–827.

Joachim Bingel. 2018. *Personalized and Adaptive Text Simplification*. Ph.D. thesis, University of Copenhagen.

Joachim Bingel and Johannes Bjerva. 2018. Cross-lingual complex word identification with multitask learning. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 166–174, New Orleans, LA, USA. Association for Computational Linguistics.

Joachim Bingel, Gustavo Paetzold, and Anders Søgaard. 2018. Lexi: A tool for adaptive, personalized text simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 245–258, Santa Fe, NM, USA. Association for Computational Linguistics.

Or Biran, Samuel Brody, and Noemie Elhadad. 2011. Putting it Simply: a Context-Aware Approach to Lexical Simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 496–501, Portland, OR, USA. Association for Computational Linguistics.

Stefan Bott, Luz Rello, Biljana Drndarevic, and Horacio Saggion. 2012. Can Spanish Be Simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of COLING 2012: Technical Papers*, pages 357–374, Mumbai, India. COLING.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, Madison, WI, USA.

John B Carroll and Margaret N White. 1973. Word frequency and age of acquisition as determiners of picture-naming latency. *The Quarterly Journal of Experimental Psychology*, 25(1):85–95.

Veronika Coltheart, Veronica J Laxon, and Corriene Keating. 1988. Effects of word imageability and age of acquisition on children's reading. *British journal of psychology*, 79(1):1–12.

Cynthia M Connine, John Mullennix, Eve Shernoff, and Jennifer Yelen. 1990. Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(6):1084.

William Coster and David Kauchak. 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, Portland, OR, USA. Association for Computational Linguini.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26, Geneva, Switzerland. ACM; New York.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases*, pages 161–173.

William H Dubay. 2004. The Principles of Readability. *Costa Mesa, CA: Impact Information*.

Carsten Elbro. 1996. Early linguistic abilities and reading development: A review and a hypothesis. *Reading and Writing*, 8(6):453–485.

Noemie Elhadad. 2006. Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms. In *AMIA Annual Symposium Proceedings*, pages 239–243, Washington, DC, USA.

Liv Engen and Torleiv Høien. 2002. Phonological skills and reading comprehension. *Reading and Writing*, 15(7-8):613–631.

Richard Evans, Constantin Orăsan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 131–140, Gothenburg, Sweden.

Christiane Fellbaum. 2005. *Encyclopedia of Language and Linguistics, Second Edition*, chapter WordNet and wordnets. Oxford: Elsevier.

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, volume 1,

pages 241–247, Atlanta, GA, USA. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, LA, USA. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019a. Complex Word Identification as a Sequence Labelling Task. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019b. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4855–4865, Hong Kong, China. Association for Computational Linguistics.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Henry Kučera and Winthrop Nelson Francis. 1967. *Computational analysis of present-day American English*. Dartmouth Publishing Group.

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Palmer Morrel-Samuels and Robert M Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(3):615.

I. S. Paul Nation. 2006. How Large a Vocabulary Is Needed For Reading and Listening? *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*, 63(1):59–82.

Charles Kay Ogden. 1968. *Basic English: international second language*. Harcourt, Brace & World.

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

Gustavo Paetzold and Lucia Specia. 2016b. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan.

Gustavo Paetzold and Lucia Specia. 2016c. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, CA, USA. Association for Computational Linguistics.

Andrew Porter, Jennifer McMaken, Jun Hwang, and Rui Yang. 2011. Common Core Standards: The New U.S. Intended Curriculum. *Educational Researcher*, 40:103–116.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, pages 203–219, Cape Town, South Africa.

Alan P Rudell. 1993. Frequency of word usage and perceived word difficulty: Ratings of kučera and francis words. *Behavior Research Methods, Instruments, & Computers*, 25(4):455–463.

Mark Sadoski, Ernest T Goetz, and Maximo Rodriguez. 2000. Engaging texts: Effects of concreteness on comprehensibility, interest, and recall in four text types. *Journal of Educational Psychology*, 92(1):85.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718, Melbourne, Australia. Association for Computational Linguistics.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *Proceedings of the Student Research Workshop at the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2014. Out in the open: Finding and categorising errors in the lexical simplification pipeline. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1583–1590, Reykjavik, Iceland. ELRA.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 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. Association for Computational Linguistics.

Edward L Thorndike and Irving Lorge. 1944. The teacher's wordbook of 30,000 words. New York: Columbia University, Teachers College.

Michael Wilson. 1988. The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments and Computers*, pages 6–11.

Seid Muhie Yimam and Chris Biemann. 2018a. Demonstrating Par4Sem - A Semantic Writing Aid with Adaptive Paraphrasing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 48–53, Brussels, Belgium. Association for Computational Linguistics.

Seid Muhie Yimam and Chris Biemann. 2018b. Par4Sim–Adaptive Paraphrasing for Text Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 331–342, Santa Fe, NM, USA. COLING.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, LA, USA. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 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*, pages 401–407, Taipei, Taiwan. Association for Computational Linguistics.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 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, Varna, Bulgaria. INCOMA Ltd.

Ahmed Zaidi, Andrew Caines, Russell Moore, Paula Buttery, and Andrew Rice. 2020. Adaptive forgetting curves for spaced repetition language learning. *arXiv preprint arXiv:2004.11327*.

Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A Text Corpora-Based Estimation of the Familiarity of Health Terminology. In *Biological and Medical Data Analysis. ISBMDA 2005. Lecture Notes in Computer Science Vol. 3745*, ISBMDA 2005, Aveiro, Portugal. Springer.

Jason D Zevin and Mark S Seidenberg. 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and language*, 47(1):1–29.