One Size Does Not Fit All:
The Case for Personalised Word Complexity Models

Sian Gooding
University of Cambridge
shg36@cam.ac.uk

Manuel Tragut
Google
mtragut@google.com

Abstract

Complex Word Identification (CWI) aims to detect words within a text that a reader may find difficult to understand. It has been shown that CWI systems can improve text simplification, readability prediction and vocabulary acquisition modelling. However, the difficulty of a word is a highly idiosyncratic notion that depends on a reader's first language, proficiency and reading experience. In this paper, we show that personal models are best when predicting word complexity for individual readers. We use a novel active learning framework that allows models to be tailored to individuals and release a dataset of complexity annotations and models as a benchmark for further research.[1]

1 Introduction

Illiteracy affects at least 773 million people globally.[2] For these individuals, reading information from unfamiliar sources or on unfamiliar topics can be extremely difficult. Furthermore, there are a reported 1.5 billion English-language learners worldwide, spanning all proficiency levels. Statistics such as these illustrate that the comprehension of written text across the world varies substantially. Even those who are deemed functionally literate have differing degrees of text comprehension. There are many characteristics of a written text that impact how difficult it is to read. However, one of the main aspects contributing to overall text difficulty is lexical complexity, i.e. word complexity (Dubay, 2004). Lexical content has been shown to be more important than syntactic features when predicting readability (Vajjala and Meurers, 2012) and a study performed by Nation (2006) found that a reader needs to be familiar with at least 95% of the words contained within a text for them to understand the content.

(1)   The *cantankerous* cat.
↓
The ***grumpy*** cat.

Complex word identification (CWI) focuses on identifying words in a text that may be difficult for a reader to understand and therefore may benefit from simplification (Shardlow, 2013a). For instance, in example (1) a CWI system may identify *cantankerous* as a complex word, then a lexical simplification system is able to substitute it with a simpler alternative, e.g. *grumpy* (Paetzold and Specia, 2016a; Gooding and Kochmar, 2019c).

Many downstream tasks benefit from the identification of complex words. For example, CWI has been shown to improve readability assessment systems (Maddela and Xu, 2018) as well as vocabulary acquisition modules in educational applications (Zaidi et al., 2020). Furthermore, CWI can significantly reduce errors in simplification systems (Shardlow, 2014) and result in higher quality simplifications (Lee and Yeung, 2018).

Previous work on modelling CWI, including two shared tasks, has relied on the collection of large corpora containing words that have been annotated for their complexity (Paetzold and Specia, 2016c; Yimam et al., 2018). However, annotating the difficulty of words is a subjective task, and previous data collection has yielded low levels of annotator agreement (Specia et al., 2012; Paetzold and Specia, 2016c). As a way of mitigating individual differences, such datasets typically present a homogeneous view on word complexity, by merging annotations across readers (Gooding et al., 2021). Further attempts to improve annotator agreement have included offering bonus incentives for annotators who select words matching other annotations (Yimam et al., 2017), as well as providing guidelines for annotators to mark words that they assume would be complex for audiences such as children and those with learning difficulties. Such incentives

---

[1]Dataset is available at: https://github.com/siangooding/personal_CWI
[2]https://en.unesco.org/themes/literacy

and instructions change the motivation of annotators and undermine the validity of the resulting complexity labels (Gooding, 2022).

It is evident that lexical complexity is a highly subjective and idiosyncratic notion. In this paper, we argue that contrary to the one-size-fits-all paradigm currently dominating many natural language processing tasks, the modelling of individual differences in word complexity is critical. Many of the downstream tasks that benefit from CWI, such as text simplification, would be more effective if tailored to the individual reader. In order to facilitate personalised CWI, and therefore personalised text simplification and readability systems, audience specific complexity annotations are required (Bingel et al., 2018). Whilst it has been shown that the concept of word complexity, and thus the level of agreement, is aligned between individuals sharing a common background (Gooding and Kochmar, 2018a; Gooding et al., 2021), we argue that the best CWI model for each individual is trained with them 'in the loop'.

In our paper, we use active learning to produce per-individual models of word complexity. We collect complexity annotations from 1,225 English as a second language (ESL) readers. Using a novel active learning framework, we show that individual models of word complexity can be trained without extensive annotation. Our experiments demonstrate that bespoke models outperform baselines and a state-of-the-art system when predicting word complexity for individuals, thereby emphasising the importance of viewing the task at a personal level. We present the following contributions:

- We are the first to train personalised word complexity models and show that training can be achieved in real-time using active learning.

- We show that per-individual models outperform baseline and state-of-the-art approaches when predicting word complexity for individuals, most notably for lower proficiency levels.

- We demonstrate the benefits of bespoke CWI models in downstream tasks such as text simplification and proficiency prediction.

- We release a novel dataset containing 55,125 English word complexity annotations from 1,225 ESL participants, as well as all trained models, as a benchmark for further research.

2 Background

2.1 Complex Word Identification

The first dataset collected for the evaluation and benchmarking of CWI was the CW corpus (Shardlow, 2013b). The dataset contained 731 instances, collected using simplifications made by human editors when adapting content for Simple Wikipedia articles. This dataset was extended with further Wikipedia examples, including data collected by Horn et al. (2014), for the first CWI shared task. The resulting dataset contained a total of 9,200 sentences which were annotated by 400 non-native speakers for complex words. The shared task organised by Paetzold and Specia (2016b) required the prediction of binary word complexity for non-native speakers. However, as the dataset had low annotator agreement, the merged notion of complexity was difficult to model, and all systems resulted in low F-scores.

Following this, Yimam et al. (2017) collected a CWI dataset which was used in a second shared task (Yimam et al., 2018). The annotations for this data were collected using the Amazon Mechanical Turk platform. For each sentence, ten non-native and ten native English speakers were asked to mark words or phrases as complex. Words and phrases were annotated as complex (label 1) if at least one of the 20 annotators annotated them as such, and simple (label 0) otherwise. The winning system for the binary track was a feature based ensemble approach (Gooding and Kochmar, 2018b). The system used traditional features associated with word complexity such as the frequency, word length and psycholinguistic attributes of words. The performance on this dataset was subsequently improved using a neural sequence labelling approach by Gooding and Kochmar (2019a).

The prior datasets and systems largely focused on binary approaches to modelling CWI. Recognising the limitation of viewing word complexity in this way, Shardlow et al. (2021) arranged a shared task on lexical complexity prediction based on the Shardlow et al. (2020) dataset. The dataset contains judgements on word complexity using a Likert scale of 1–5, allowing an indication of the annotators' familiarity with the term. The framing of word complexity in a continuous fashion allows for more subjective approaches of CWI to be introduced. However, overall judgements of difficulty in this dataset are still produced by averaging the labels across multiple annotators. Whilst the data

gives an insight into complexity as a continuous measure, it does not provide insight into the individual differences across annotators.

Closest to our work, Gooding et al. (2021) make the case for audience specific CWI by showing that the best models for a given proficiency group are produced using annotations from the target audience. They use proficiency annotations from the Yimam et al. (2017) dataset and find that the features contributing to word complexity differ depending on the audience background. This work illustrates the importance of considering the target audience in CWI. However, the populations investigated still represent coarse groups, such as native vs non-native, or beginner, intermediate and advanced. Whilst this work is a step in the right direction for subjective complexity, there still remain many differences in complexity judgements even within such groups.

Finally, a similar task to CWI is that of vocabulary prediction, whereby the size and extent of a learner's vocabulary is estimated (Meara and Buxton, 1987; Laufer and Nation, 1999). Whilst word difficulty is subtly different to vocabulary size, as a learner may recognise a word but still find it difficult to understand, the areas are closely aligned. Empirical studies have found strong correlations between receptive vocabulary size tests and reading comprehension tests for ESL readers with learners from different proficiency levels (e.g. Laufer, 1992; Qian, 1999, 2002; Henriksen et al., 2004). In fact, vocabulary size has been found to predict as much as 72% of the variance in reading, leading many researchers to emphasise that receptive vocabulary size is the determinant factor for reading success in L2 (Staehr, 2008). Therefore, being able to predict a readers' vocabulary size, with a small number of samples, is very useful. The effectiveness of active learning for vocabulary prediction was investigated by Ehara et al. (2014), who showed that graph-based active learning improves vocabulary prediction for Japanese speakers. However, the active learning process is simulated and not performed in real-time with participants.

2.2 Active Learning

Annotated data can be time consuming and expensive to obtain, notably in specialised domains where only experts can provide reliable labels (Konyushkova et al., 2017). By enabling a classifier to interactively query data points, active learning allows machine learning classifiers to achieve higher accuracies with fewer training instances. It is well-motivated in many modern machine learning problems where data may be abundant, but labels are scarce or expensive to acquire (Settles, 2009). For instance, active learning has been used in text classification (Hoi et al., 2006), information extraction (Settles and Craven, 2008) and financial applications (Gooding and Briscoe, 2019).

The task of personalised CWI is a good candidate for active learning for two reasons:

Labelling time Previous CWI datasets contain thousands of annotated instances to train word complexity models. It is unfeasible to collect this amount of data, in a reasonable time frame, for an individual reader.

Learner lifetime In complex word identification we can expect the needs of a user to change over their 'learner lifetime'. Whilst a model may be effective for a reader at one point in time, as their proficiency improves, the model will require retraining.

2.2.1 Uncertainty Sampling

In uncertainty sampling an active learner queries the instances it is least certain about how to label. This is implemented using an uncertainty measure and is one of the most commonly used active learning frameworks (Lewis and Gale, 1994; Settles, 2009). For many real-world learning problems, large collections of unlabelled data can be gathered easily. For instance, in our setting the unlabelled data represents all possible words. This motivates an uncertainty sampling approached called pool-based sampling (Lewis and Gale, 1994) which assumes that there is a small set of labelled data $\mathcal{L}$ and a large pool of unlabelled instances $\mathcal{U}$. Instances are drawn from this unlabelled pool for labelling in a greedy fashion, according to an informativeness measure calculated by applying the classifier and measuring the uncertainty.

In our paper, we employ active learning for the task of complex word identification. We train classifiers to identify whether a word would be considered difficult for a reader based on 23 word familiarity questions.

3 Experiment Design

In our experiment, participants are presented with vocabulary and asked whether they could confidently define the meaning. There are two stages to

the data collection; initially, word annotations are requested in tandem with a classifier as part of an active learning process. During this stage, participants are shown a total of 23 words to annotate.[3] These annotations are used to train a personal word complexity model. Following this, annotations are collected for a test set, per participant, which can be used to validate the trained classifier. The test set contains 22 words across CEFR levels and is presented in a randomised order. Annotators are not aware of the switch from training to test data collection. The study was hosted online and implemented using the Flask framework for Python (Grinberg, 2018).

3.1 Participants

We collected annotations via the Prolific platform from 1,255 ESL readers spanning 57 different first languages.[4] Participants were required to fill out a demographic questionnaire which is available to view in our data repository. The questionnaire was informed using guidelines on judging reading ability (Acheson et al., 2008) and included the self-reported English proficiency, native language, hours spent reading per week and highest level of formal education. Demographic statistics are available in Appendix A.

3.2 Materials

We created a test set containing 22 words spanning different levels of the Common European Framework of Reference for Languages (CEFR) (North and Piccardo, 2020) shown in Figure 1. The items were selected by sampling those with the fewest dictionary entries to minimise word ambiguity.

To bootstrap the active learning process, we train an initial logistic regression classifier with 150 instances from the CWI dataset of Yimam et al. (2017). The features used are those found to correlate most with non-native judgements of word complexity according to Gooding et al. (2021), namely the word length, word frequency and psycholinguistic properties of the words such as the familiarity, concreteness, and imagability, which have been shown as important when predicting word complexity (Carroll and White, 1973; Begg and Paivio, 1969; Zevin and Seidenberg, 2002). This classifier is then tuned to the individual during the active learning process.

3.2.1 Pre-clustering

A typical approach for collecting annotations, in an active learning setting, is to select samples close to the classification boundary. However, it has been shown that taking into account the prior data distribution can improve performance (Settles, 2009). One way to achieve this, is by modelling the input distribution using clustering, and propagating label information to instances in the same cluster (Nguyen and Smeulders, 2004). In our experiment, we use the Yimam et al. (2017) dataset as our unlabelled pool and perform clustering on the 7,476 uniquely labelled words prior to active learning.

We use agglomerative clustering, a hierarchical clustering approach which performs clustering in a bottom up fashion (Gowda and Krishna, 1978). Each word initially belongs to its own cluster and the algorithm successively merges these to form larger groups. This type of unsupervised clustering has been used across multiple natural language processing tasks including word clustering, co-reference resolution and word sense disambiguation (Chen and Ji, 2010). The features used for clustering are the same as employed for the task of the complexity prediction.

We set the number of clusters to 7 in an attempt to align with the CEFR categories chosen in our test set. One disadvantage of hierarchical clustering methods, is that they have large storage requirements, and so can be computationally intensive to compute (Gan et al., 2020). However, as we precalculate and cache clusters the time complexity is that of base informativeness measures, e.g. uncertainty sampling (Settles and Craven, 2008). The time complexity is a necessary consideration, as we are conducting active learning interactively with oracles in real-time.

An example of words from the resulting clusters is displayed in Figure 2a. Words of length 5 were sampled randomly and have been colour coded according to the corresponding CEFR level.[5]

3.2.2 Cluster Evaluation

We evaluate our clusters using two techniques; the first employs EFLLEX which is a graded lexical resource for learners of English as a foreign language. The dataset, collected by Dürlich and François (2018), contains 15,280 entries from 1,971 graded texts. The frequency of the lemma at each level of
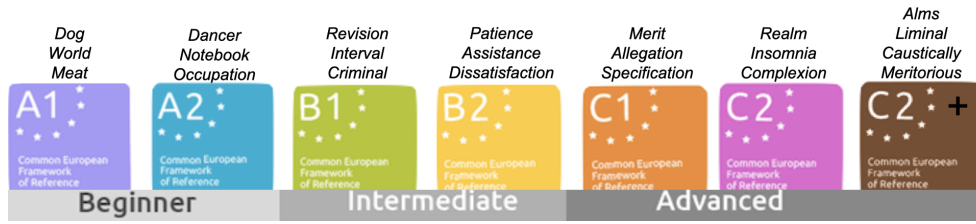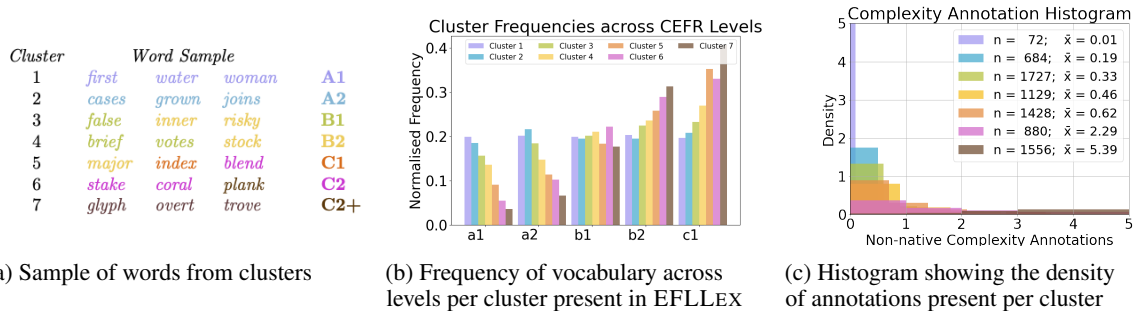

Figure 1: Test set

(a) Sample of words from clusters

(b) Frequency of vocabulary across levels per cluster present in EFLLEX

(c) Histogram showing the density of annotations present per cluster

Figure 2: Clustering statistics

the CEFR (C2 excepted) is provided. A total of 5,983 entries were present in both EFLLEX and our clusters. We plot the average frequencies of words in each cluster across CEFR levels in Figure 2b. Generally, the plot shows that there are clear differences in the levels of vocabulary across clusters. When we consider A1 texts, the cluster with the highest average frequency of words is Cluster 1. There is a linear decrease in the frequency of words at this level as the cluster level increases. When we consider the proportion of higher level vocabulary (i.e. B1 to C2), a clear trend emerges where Clusters 5–7 contain words that appear much more frequently at this level. Cluster 7, contains words at the highest level of CEFR for instance *glyph*, and is shown to contain words that occur much more commonly at C1 and B2 level.

The second way we evaluate clusters is using the annotations of the Yimam et al. (2017) dataset. As we use this data for our unlabelled pool, we have the complexity annotations for each item from both native and non-native annotators. Our work focuses on complexity prediction for ESL readers and so we evaluate clusters using the non-native annotations. This results in a complexity score per word ranging from 1–10 which represents how many non-native annotators considered the word to be complex. A histogram showing the density of non-native complexity annotations present for words in each cluster is shown in Figure 2c. The average of these scores per cluster ($\bar{x}$) is shown in Figure 2c,
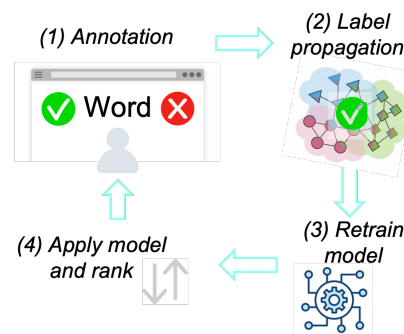

Figure 3: Figure showing the stages of our active learning framework

as well as the number of words present in each cluster (n). Clusters are colour coded according to the same key in Figure 2c. Cluster 7 contains the highest average complexity labels (5.39) and Cluster 1 the lowest (0.01). As the cluster levels increase, we see a shift in the density of complexity annotations present, supporting the notion that these clusters contain vocabulary at differing levels.

3.3 Active Learning Loop

Figure 3 shows the four stages that take place during the active learning process. Each of these steps is described below.

(1) Annotation In the initial step, the annotator is presented with a word form and asked whether they could confidently define the meaning. Whilst the context of a word impacts its meaning and com-

plexity, we use word forms as this allows for a pool-based active learning approach. Furthermore, the use of word forms alone has been shown as an effective measure of proficiency and vocabulary prediction (Lemhöfer and Broersma, 2012; Ehara et al., 2014). The user interface showing the instructions for this experiment are provided in Appendix B. When the annotator selects an answer this provides a binary annotation for the word.

(2) Label propagation The chosen label is then propagated to the closest 150 instances from the clustered pool.

(3) Re-train model We train a log-linear model (i.e. logistic regression) per individual, for the following reasons: the model can provide a class ownership probability which is required for active learning, log-linear models are flexible and interpretable, finally the algorithm is efficient and able to train in $\mathcal{O}(nd)$ time, where n is the number of samples and d the feature dimensions (Gujarati and Porter, 2009; Bulso et al., 2019). As each model is retrained every time an annotation is provided, the speed of training was a critical consideration.

(4) Apply model and re-rank Once the model has been re-trained, it is applied to the unlabelled pool so that the samples can be ranked according to the model uncertainty. The uncertainty sampling technique we use is entropy (Shannon, 1948). In the following notation x^* denotes the most informative instance from an unlabelled vocabulary pool:

$$x_H^* = \underset{x}{\operatorname{argmax}} - \sum_i P_\theta(y_i|x) \log P_\theta(y_i|x)$$

where y_i ranges over the possible labellings. The instance from the pool that the model is least certain about is then presented to the annotator for the next annotation step.

4 Results

In our experiment, each participant produces a gold test set containing annotations for words across CEFR levels. This varies for each individual, as word complexity is subjective. For instance, consider the test set samples shown in Table 1; words marked with a 1 have been annotated as difficult by participants. Participant a, who is self-reported to be intermediate, has found three words in this sample difficult, whereas participant b, self-reported at a near native level, only reports one.

	World	Realm	Merit	Alms
P_a	0	1	1	1
P_b	0	0	0	1

Table 1: Test set samples from participants P_a and P_b

We calculate the F-score and Kappa agreement across all annotations to evaluate the performance of bespoke models. The F-score represents the harmonic mean between the precision and recall of the classifier, whereas the Kappa coefficient measures pairwise agreement, correcting for expected chance agreement (Cohen, 1960).

We additionally run an experiment to evaluate the effectiveness of active learning compared to randomly sampling vocabulary. In the experiment, we require annotators to label a sample of 22 random words, as well as 22 words sampled using active learning. As this doubles the annotation effort we perform this with a subset of 100 annotators. The results are shown in Table 3. We compare with models trained using random sampling as well as random sampling with cluster-based label propagation. The best results are achieved using the active learning framework.

We evaluate the performance of the active learning models and compare to the following baselines. (1) Group Average: we calculate the average of the test set complexity annotations across participants, if a word has been labelled as complex by more than 10% of the target group it is labelled as difficult. This is a competitive baseline as we use the direct annotations of the target audience on the test sets. (2) Seq-CWI: we compare to a state-of-the-art system for complex word identification (Gooding and Kochmar, 2019a). However, we note that this is an unfair comparison as the system is one-size-fits-all and not audience specific. (3) Frequency: relies on frequency thresholding, where all words below a given frequency are marked as complex. (4) All Simple: all words are marked as simple.

Table 2 shows the results of the active learning models compared to baselines. We group results according to the proficiency of participants. The beginner results are shown for completeness, but should be interpreted with caution due to the smaller test set size (176). Kappa values in the range of 0.40–0.60 represent moderate agreement (Cyr and Francis, 1992), although it has been argued that scores above 0.45 represent substantial agreement (Munoz and Bangdiwala, 1997).

Model	Proficiency level					
	All	Beginner	Intermediate	Advanced	Near Native	Native
			F-score			
Active Learning	**0.751**	**0.763**	**0.720**	**0.752**	**0.764**	**0.761**
Group Average	0.669	0.608	0.624	0.679	0.726	0.754
Seq-CWI	0.677	0.620	0.688	0.689	0.660	0.642
Frequency	0.534	0.461	0.529	0.541	0.530	0.514
All Simple	0.441	0.450	0.436	0.437	0.447	0.454
			Kappa			
Active Learning	**0.503**	**0.525**	**0.441**	**0.504**	**0.529**	**0.524**
Group Average	0.379	0.296	0.304	0.395	0.473	0.523
Seq-CWI	0.395	0.300	0.409	0.414	0.371	0.347
Frequency	0.189	0.083	0.173	0.194	0.196	0.182
Test size	26,136	176	4,180	12,122	7480	2,156

Table 2: Model results grouped by annotator proficiency

Model	*F-score*	*Kappa*
Random	0.605	0.218
Cluster	0.649	0.361
AL	0.740	0.482

Table 3: Results showing differing sampling techniques applied to 2,200 annotations from 100 participants

When we consider the active learning models, we see that the F-score calculated across all test annotations is 0.751, and the Kappa agreement is 0.503. The active learning models outperform the competitive group-average baseline which supports the hypothesis that individual models are preferable to aggregate group-level approaches. The best scores for all proficiency groups are achieved using the active learning framework. However, for the native annotations, the difference between the group-based approach and active learning models is small. When considering the group-average baseline we see that both the F-score and Kappa values increase in line with proficiency. This supports the notion that the concept of word complexity stabilises for higher proficiency levels (Krstić et al., 2018; Gooding et al., 2021). As such, aggregate models may work well for highly proficiency audiences and individual models best for lower levels where word knowledge is more variable.

4.1 Analysis

Each annotator produces a bespoke model based on their personal perception of word difficulty. We tested whether there was a significant difference in complexity predictions of models depending on the annotator's self-reported proficiency. To investigate this, we selected 100 models from near native, advanced and intermediate annotators. We control for the number of classifiers due to having an unequal number of annotators across levels.

These models are applied to the EFLLEX dataset, removing any vocabulary that may have been seen during train time. This results in complexity predictions for approximately 10,000 words per model. The average number of words found complex by these classifiers is shown in Table 5. The intermediate classifiers, on average, identify more complex words across all vocabulary levels, whereas the near native classifiers predict fewer words as difficult across all levels compared to both the intermediate and advanced classifiers. We calculate the significance using Satterthwaite's method (Kuznetsova et al., 2017), applied to a mixed-effects model that treats the first language of annotators and the trained model as random effects.[6] We find that the difference between model predictions depending on the proficiency of the annotator is highly significant ($p < 1 \times 10^{-10}$).

4.2 Downstream Tasks

Being able to distil the vocabulary experience of an individual, using a limited number of annotations, is extremely useful for many applications.

---

[6]Using R formula notation, the model is: model_predictions $\sim$ annotator_proficiency + (1 | annotator_L1) + (1 | model)

Near Native:	0.34	0.32
Advanced:	0.51	0.47
Intermediate:	0.59	0.59

(1) The	current	cuts	in	public	expenditure	will	exacerbate	this	situation.
				↓					
(2) The	current	cuts	in	public	*spending*	will	exacerbate	this	situation.
(3) The	current	cuts	in	public	*spending*	will	*worsen*	this	situation.

Table 4: Example of how simplification output changes depending on audience specific CWI models. The complexity probability for audiences is the result of aggregating output from all trained models across proficiencies. A word is considered complex if the probability is above 0.50.

	A1	A2	B1	B2	C1
Intermediate	333	881	1208	2109	2474
Advanced	277	754	1016	1773	2062
Near Native	232	643	866	1505	1735

Table 5: The average number of words identified as complex across CEFR levels. Results are averaged from a sample of 100 classifiers from each proficiency group.

For instance, when judging the proficiency of a reader. Using our trained classifiers, we run a preliminary experiment to see if it is possible to predict the proficiency of the annotator using their model predictions. We train a linear model to predict whether the annotator self-reported at an intermediate, advanced or near native level. We use the total number of C1 words identified as complex as a feature and test using 5-fold stratified cross-validation. By training a simple model with this feature alone, we can obtain a precision of 0.70. We release the trained models with our paper to encourage further work on the applications and evaluation of personal CWI models for proficiency prediction.

Finally, a further application that benefits from individualised models of word complexity is text simplification. To illustrate, we provide an example in Table 4 which was produced using an open source lexical simplification system by Gooding and Kochmar (2019c). The system is designed to accommodate differing models of word complexity so that simplification output can be personalised. In this example, we use the mean probability that a word is complex by averaging the output from our classifier groups. Sentence (1) shows the result of using near native CWI models, no words are considered complex and so no simplifications are performed. The advanced models produce a probability larger than 0.50 for the word *expenditure*

and it is therefore simplified. Finally, the intermediate models result in two words being identified as complex which are subsequently simplified. This example provides further motivation for why considering word complexity in a subjective manner is beneficial.

5 Conclusions and Future Work

We show that with active learning it is feasible to build individual models of word complexity. Our models are able to be trained efficiently, in real-time, using only 23 word annotations per individual. Models trained in this fashion are able to accommodate the subjective nature of word complexity, thereby facilitating personalised readability and simplification systems. We release our dataset containing 55,125 word complexity annotations, participant metadata and models, providing a new benchmark for the task of human-centered word complexity prediction.

Our work is a first step towards the development of personal complexity models. In future work, we aim to more extensively investigate the relationship between a learner's perceived and actual knowledge of words, as it has been reported that there is a misalignment between what a learner thinks they know, and what they are actually able to define (Laufer, 1997). Additionally, the context of a word impacts its perceived difficulty (Gooding and Kochmar, 2019b; Strohmaier et al., 2020); considering token-level complexity as opposed to type-level complexity is another important avenue for future work.

References

Daniel J Acheson, Justine B Wells, and Maryellen C MacDonald. 2008. New and updated tests of print exposure and reading abilities in college students. *Behavior research methods*, 40(1):278–289.

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, 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.

Nicola Bulso, Matteo Marsili, and Yasser Roudi. 2019. On the complexity of logistic regression models. *Neural computation*, 31(8):1592–1623.

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.

Zheng Chen and Heng Ji. 2010. Graph-based clustering for computational linguistics: A survey. In *Proceedings of TextGraphs-5-2010 Workshop on Graph-based Methods for Natural Language Processing*, pages 1–9.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Louis Cyr and Kennon Francis. 1992. Measures of clinical agreement for nominal and categorical data: the kappa coefficient. *Computers in biology and medicine*, 22(4):239–246.

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

Luise Dürlich and Thomas François. 2018. Efllex: A graded lexical resource for learners of english as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014. Formalizing word sampling for vocabulary prediction as graph-based active learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1374–1384.

Guojun Gan, Chaoqun Ma, and Jianhong Wu. 2020. *Data clustering: theory, algorithms, and applications*. SIAM.

Sian Gooding. 2022. On the ethical considerations of text simplification.

Sian Gooding and Ted Briscoe. 2019. Active learning for financial investment reports. In *Proceedings of the Second Financial Narrative Processing Workshop (FNP 2019)*, pages 25–32, Turku, Finland. Linköping University Electronic Press.

Sian Gooding and Ekaterina Kochmar. 2018a. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana.

Sian Gooding and Ekaterina Kochmar. 2018b. 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 Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153.

Sian Gooding and Ekaterina Kochmar. 2019b. 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. 2019c. 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.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449.

K Chidananda Gowda and G Krishna. 1978. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern recognition*, 10(2):105–112.

Miguel Grinberg. 2018. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.".

DN Gujarati and DC Porter. 2009. How to measure elasticity: The log-linear model. *Basic Econometrics, McGraw-Hill/Irwin, New York*, pages 159–162.

BIRGIT Henriksen, DORTE Albrechtsen, and KIRSTEN Haastrup. 2004. The relationship between vocabulary size and reading comprehension in the l2. *Angles on the English-speaking World*, 4(1):129–140.

Steven CH Hoi, Rong Jin, Jianke Zhu, and Michael R Lyu. 2006. Batch mode active learning and its application to medical image classification. In *Proceedings of the 23rd international conference on Machine learning*, pages 417–424.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a lexical simplifier using Wikipedia. In *Proceedings of the 52nd ACL*, pages 458–463.

Ksenia Konyushkova, Raphael Sznitman, and Pascal Fua. 2017. Learning active learning from data. In *Advances in Neural Information Processing Systems*, pages 4225–4235.

Ksenija Krstić, Anđela Šoškić, VanjaKović, and Kenneth Holmqvist. 2018. All good readers are the same, but every low-skilled reader is different: an eye-tracking study using pisa data. *European Journal of Psychology of Education*, 33(3):521–541.

Alexandra Kuznetsova, Per B Brockhoff, Rune HB Christensen, et al. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13):1–26.

Batia Laufer. 1992. How much lexis is necessary for reading comprehension? In *Vocabulary and applied linguistics*, pages 126–132. Springer.

Batia Laufer. 1997. The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. *Second language vocabulary acquisition: A rationale for pedagogy*, 1:20–34.

Batia Laufer and Paul Nation. 1999. A vocabulary-size test of controlled productive ability. *Language testing*, 16(1):33–51.

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

Kristin Lemhöfer and Mirjam Broersma. 2012. Introducing lextale: A quick and valid lexical test for advanced learners of english. *Behavior research methods*, 44(2):325–343.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

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.

Paul Meara and Barbara Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language testing*, 4(2):142–154.

Sergio R Munoz and Shrikant I Bangdiwala. 1997. Interpretation of kappa and b statistics measures of agreement. *Journal of Applied Statistics*, 24(1):105–112.

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.

Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 79.

Brian North and Enrica Piccardo. 2020. *Companion volume COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES: LEARNING, TEACHING, ASSESSMENT Companion volume Language Policy Programme Education Policy Division Education Department Council of Europe*.

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. 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.

Gustavo Paetzold and Lucia Specia. 2016c. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974, San Diego, CA, USA. Association for Computational Linguistics.

David Qian. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian modern language review*, 56(2):282–308.

David D Qian. 2002. Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language learning*, 52(3):513–536.

F. E. Satterthwaite. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Burr Settles. 2009. Active learning literature survey.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Matthew Shardlow. 2013a. 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. 2013b. The cw corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics (ACL).

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.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. Semeval-2012 task 1: English lexical simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 347–355. Association for Computational Linguistics.

Lars Stenius Staehr. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2):139–152.

David Strohmaier, Sian Gooding, Shiva Taslimipoor, and Ekaterina Kochmar. 2020. SeCoDa: Sense complexity dataset. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5962–5967, Marseille, France. European Language Resources Association.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Seid Muhie Yimam, Sanja Štajner, Martin Riedl, and Chris Biemann. 2017. 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*.

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.

A Participant demographics

The demographic information of participants is shown in Table 6. A more detailed breakdown including first language statistics is available on our data repository.

B Experiment interface

The experiment interface is shown in Figure 4. Annotators were paid 15 cents per minute in accordance with Prolific guidelines. Annotators were informed that the study and their subsequent annotations were for research purposes. Data is fully anonymised and both this and the models will be released under the Affero General Public License 3 (AGPL-3.0). All participant instructions as well as the demographic questionnaire are included in our GitHub data repository.

Proficiency (%)		Education (%)		Age range (%)		Hours reading English per week (%)	
Native	8.29	Graduate	31.16	18 - 24	39.24	0 - 10	40.40
Near native	30.50	Undergraduate	29.83	25 - 34	22.17	10 - 20	22.83
Advanced	45.51	High School	33.37	35 - 44	14.52	20 - 30	14.95
Intermediate	15.14	Vocational Training	5.42	45 - 54	11.64	30 - 40	9.82
Beginner	0.66	No formal education	0.22	55+	9.53	40 +	11.98

Table 6: Background statistics for the 1225 participants

Instructions

Below you will be shown **67** words.

Please click the 'yes' and 'no' buttons to answer the following question:

Can you confidently define this word?

initially

| Yes | No |

Figure 4: Experiment interface