Language Resources and Evaluation
https://doi.org/10.1007/s10579-022-09588-2

ORIGINAL PAPER

Predicting lexical complexity in English texts: the Complex 2.0 dataset

Matthew Shardlow[1] · **Richard Evans**[2] · **Marcos Zampieri**[3]

Accepted: 7 March 2022

Abstract

Identifying words which may cause difficulty for a reader is an essential step in most lexical text simplification systems prior to lexical substitution and can also be used for assessing the readability of a text. This task is commonly referred to as complex word identification (CWI) and is often modelled as a supervised classification problem. For training such systems, annotated datasets in which words and sometimes multi-word expressions are labelled regarding complexity are required. In this paper we analyze previous work carried out in this task and investigate the properties of CWI datasets for English. We develop a protocol for the annotation of lexical complexity and use this to annotate a new dataset, CompLex 2.0. We present experiments using both new and old datasets to investigate the nature of lexical complexity. We found that a Likert-scale annotation protocol provides an objective setting that is superior for identifying the complexity of words compared to a binary annotation protocol. We release a new dataset using our new protocol to promote the task of Lexical Complexity Prediction.

Keywords Complex word identification · Lexical complexity · Text simplification

✉ Matthew Shardlow
m.shardlow@mmu.ac.uk

Richard Evans
r.j.evans@wlv.ac.uk

Marcos Zampieri
marcos.zampieri@rit.edu

[1] Manchester Metropolitan University, Manchester, UK

[2] University of Wolverhampton, Wolverhampton, UK

[3] Rochester Institute of Technology, Rochester, USA

Published online: 23 March 2022

1 Introduction

Predicting lexical complexity can enable systems to better guide a user to an appropriate text, or tailor it to their needs. The task of automatically identifying which words are likely to be considered complex by a given target population is known as Complex Word Identification (CWI) and it constitutes an important step in most lexical simplification pipelines (Paetzold & Specia, 2017).

The topic has gained significant attention in the last few years, particularly for English—which is also the focus of our study. A number of studies have been published on predicting complexity of both single words and multi-word expressions (MWEs) including two recent competitions organized on the topic, CWI 2016 and CWI 2018, discussed in detail in Sect. 2. The first shared task on CWI was organized at SemEval in 2016 (Paetzold & Specia, 2016a) providing participants with an English dataset in which words in context were annotated as non-complex (0) or complex (1) by a pool of human annotators. The goal was to predict this binary value for the target words in the test set. A post-competition analysis of the CWI 2016 results (Zampieri et al., 2017) examined the performance of the participating systems and evidenced how challenging CWI 2016 was with respect to the distribution (more testing than training instances) and annotation type.

The second edition of the CWI shared task was organized in 2018 at the BEA workshop (Yimam et al., 2018). CWI 2018 featured multilingual (English, Spanish, German, and French) and multi-domain datasets (Yimam et al., 2017). Unlike in CWI 2016, predictions were evaluated not only in a binary classification setting but also in terms of probabilistic classification in which systems were asked to assign the probability of the given target word in its particular context being complex. Although CWI 2018 provided an element of regression, the continuous complexity value of each word was calculated as the proportion of annotators that found a word complex. For example, if 5 out of 10 annotators labeled a word as complex then the word was given a score of 0.5. This measure relies on an aggregation of absolute binary judgments of complexity to give a continuous value.

Instead of using binary judgments, the CompLex dataset uses Likert Scale judgments (Shardlow et al., 2020), for which the specification is discussed in depth in Sect. 4. CompLex is a multi-domain English dataset annotated with a 5-point Likert scale (1-5) corresponding to the annotators comprehension and familiarity with the words in which 1 represents *very easy* and 5 represents *very difficult*. The CompLex dataset was used as the official dataset of SemEval-2021 Task 1: Lexical Complexity Prediction (LCP) (Shardlow et al., 2021). The goal of LCP 2021 is to predict this complexity score for each target word in context in the test set.

In this paper, we investigate properties of multiple annotated English lexical complexity datasets such as the aforementioned CWI datasets and others from the literature (Maddela & Xu, 2018). We investigate the types of features that make words complex. We analyse the shortcomings of the previous CWI datasets and use this to motivate the specification of a new type of CWI dataset, focusing not on complex-word identification (CWI), but instead on lexical complexity prediction (LCP), that is CWI in a continuous-label setting. We further develop a dataset based on adding

additional annotations to the existing CompLex 1.0 to create our new dataset, CompLex 2.0, and use this to provide experiments into the nature of lexical complexity.

The main contributions of this paper are:

- A concise yet comprehensive survey of the two editions of the CWI shared tasks organized in 2016 and 2018;
- An investigation into the types of features that correlate with lexical complexity;
- A qualitative analysis of the CWI–2016 (Paetzold and Specia 2016a), CWI–2018 (Yimam et al., 2018) and Maddela–2018 (Maddela and Xu 2018) datasets, highlighting issues with the annotation protocols that were used;
- The specification of a new annotation protocol for the CWI task;
- An implementation of our specification, describing the annotation of a new dataset for CWI (CompLex 1.0 and 2.0);
- Experiments comparing the features affecting lexical complexity in our dataset, as compared to others;
- Experiments using our dataset, demonstrating the effects of genre on CWI.

The remainder of this paper is organized as follows. Section 2 provides an overview of the previous CWI shared tasks. Section 3 provides a preliminary investigation into the types of features that correlate with complexity labels in previous CWI datasets. Section 4 firstly discusses the datasets that have previously been used for CWI, highlighting issues in their annotation protocols in Sect. 4.1, and then proposes a new protocol for constructing CWI datasets in Sect. 4.5. Section 5 reports on the construction of a new dataset following the specification previously laid out. Section 6 compares the annotations in our new dataset to those of previous datasets by developing a categorical annotation scheme. Section 7 shows further experiments demonstrating how our new corpus can be used to investigate the nature of lexical complexity. Finally, a discussion of our main thesis and conclusions of our work are presented in Sects. 8 and 9 respectively.

We have previously published the CompLex 1.0 data as a workshop paper (Shardlow et al., 2020). The CompLex 2.0 data was also described in the SemEval task description paper (Shardlow et al., 2021). In this paper, we seek to build upon these prior works to give an in depth and rounded treatment to the lexical complexity problem.

2 Related work

There have been various studies which have both created datasets and explored computational models for CWI, particularly focusing on English texts (Shardlow, 2013b, a; Gooding and Kochmar, 2019; Finnimore et al., 2019). These studies have addressed CWI as a stand-alone task or as part of lexical simplification pipelines.

Given the direct application of CWI to lexical simplification systems, where the goal is to decide whether or not a word needs to be substituted for a simpler one, the clear majority of studies have addressed CWI as a binary classification task.

That said, there have been multiple studies analyzing the shortcomings of approaching CWI as a binary classification task. Some studies have studied the relationship between classification performance and dataset annotation in an attempt to estimate the theoretical upper boundary of binary CWI systems (Zampieri et al., 2017) while others have investigated alternative ways to model the task. One study posed that comparative judgments are more consistent than binary classification for CWI (Gooding et al., 2019).

CWI is of direct interest to those working in lexical simplification as it forms the first part of the lexical simplification pipeline (Devlin & Tait, 1998). Before a word can be simplified, a decision must be made as to whether or not that word requires simplification. Simplification systems (Biran et al., 2011; Bott et al., 2012), then generate potential candidates for simplification and use a similar process to CWI to select the most simple candidate (Paetzold et al., 2017).

Comparative complexity is a related but distinct task to Lexical Complexity Prediction. In this task, two words are taken and a judgment is given to determine which is the most complex. A recent study found that annotations for comparative complexity were more consistent than binary classification (Gooding et al., 2019). Nonetheless, we have not focussed on comparative complexity in this work, but rather on continuous complexity. We are most interested in the complexity of a word in it's original context, rather than in relation to another word.

The increased interest from the research community in CWI was the primary motivation for the organisation of the two editions of the aforementioned CWI shared task in 2016 and 2018. These shared tasks have made important benchmark datasets available to the community that are widely used beyond these competitions. In the next sub-sections we provide an overview of these two editions: CWI–2016 organized at SemEval 2016 (Paetzold & Specia, 2016a) and CWI–2018 organized at the BEA workshop in 2018 (Yimam et al., 2018). We describe the task setup, present the datasets, and briefly discuss the approaches submitted by participants in the two editions of the competition. We also present the approaches and the features used by each system. Finally, we analyze the results obtained by the participants and the main challenges of each edition of the CWI Shared Task.

2.1 CWI–2016

The first shared task on CWI was organized as Task 11 at the International Workshop on Semantic Evaluation (SemEval) in 2016.[1] CWI–2016 provided participants with a manually annotated dataset in which words in context were labeled as complex or non-complex, where complexity is interpreted as whether a word was understood or not by a pool of 400 non-native speakers of English. CWI–2016 was therefore modelled as a binary text classification task at the word level. Participants were required to build systems to predict lexical complexity in sentences of the unlabeled test set and assign label 0 to non-complex words and 1 to complex ones. Two examples from the CWI–2016 dataset are shown below:

[1] http://alt.qcri.org/semeval2016/task11/.

(1) A **frenulum** is a small fold of tissue that secures or **restricts** the **motion** of a mobile organ in the body.
(2) The name 'kangaroo mouse' refers to the species' **extraordinary** jumping ability, as well as its habit of **bipedal locomotion**.

The words in bold: *frenulum, restricts,* and *motion* in Example 1, and *extraordinary, bipedal,* and *locomotion* in Example 2 were annotated by at least one of the annotators as complex and thus they were labeled as such in the training set. Adjacent words like *bipedal locomotion* do not represent multi-word expressions (MWEs) as they were annotated in isolation because the task set-up of CWI–2016 only considered single word annotations. Whilst MWEs were not considered in CWI–2016, they were studied in CWI–2018 (see Sect. 2.2).

The dataset provided by the organizers of CWI–2016 contained a training set of 2237 target words in 200 sentences. The training set was annotated by 20 annotators and a word was considered complex in the training set if at least one of the 20 annotators assigned it as so. The test set included 88,221 target words in 9,000 sentences and each word was annotated by only one annotator. Therefore, the ground truth label for each word in the test was attributed based on a single complexity judgement. According to the organisers of CWI–2016, this setup was devised to imitate a realistic scenario where the goal was to predict the individual needs of a speaker based on the needs of the target group (Paetzold & Specia, 2016a). Finally, the data included in the CWI–2016 dataset comes from various sources such as the CW Corpus (Shardlow, 2013a), the LexMTurk Corpus (Horn et al., 2014), and Simple Wikipedia (Kauchak, 2013).

CWI–2016 attracted a large number of participants. A total of 21 teams submitted 42 systems to the competition. A wide range of features such as word embeddings, word and character n-grams, word frequency, Zipfian frequency-based features, word length, morphological, syntactic, semantic, and psycholinguistic features were used by participants. A number of different approaches to classification were tested, ranging from traditional machine learning classifiers such as support vector machines (SVM), decision trees, random forest, and maximum entropy classifiers to deep learning classifiers, such as recurrent neural networks. In Table 1, we list the approaches submitted to CWI–2016 by the 19 teams who wrote system description papers presented at SemEval.

In terms of performance the top-3 systems were team PLUJAGH (Wróbel, 2016), LTG (Malmasi et al., 2016), and MAZA (Malmasi & Zampieri, 2016) which obtained 0.353, 0.312, and 0.308 F1-score respectively. The three teams used rather simple probabilistic models trained on features such as n-grams, word frequency, word length, and the presence of words in vocabulary lists extracted from Simple Wikipedia, introduced by PLUJAGH. The relatively low performance obtained by all teams, including the top-3 systems, evidences how challenging the CWI–2016 shared task was. Both the data annotation protocol and the training/test split, where 40 times more testing data than training data is available, contributed to making CWI–2016 a difficult task.

A post-competition analysis was carried out using the output of all 42 systems submitted to CWI–2016 (Zampieri et al., 2017). Each system output to each test

Table 1 Systems submitted to the CWI–2016 in alphabetical order. We include team names and a brief description of each system including features and classifiers used. A reference to each system description paper is provided for more information

Team	Classifiers	Features	References
AI-KU	SVM	Word embeddings of the target and surrounding words	Kuru (2016)
Amrita-CEN	SVM	Word embeddings and various semantic and morphological features	Sanjay and Soman (2016)
BHASHA	SVM, Decision Tree	Lexical and morphological features	Choubey and Pateria (2016)
ClacEDLK	Random Forests	Semantic, morphological, and psycholinguistic features	Davoodi and Kosseim (2016)
CoastalCPH	Neural Network, Logistic Regression	Word frequencies and word embeddings	Bingel et al. (2016)
HMC	Decision Tree	Lexical, semantic, syntactic and psycholinguistic features	Quijada and Medero (2016)
IIIT	Nearest Centroid	Semantic and morphological features	Palakurthi and Mamidi (2016)
JUNLP	Random Forest, Naive Bayes	Semantic, lexicon-based, morphological and syntactic features	Mukherjee et al. (2016)
LTG	Decision Tree	n-grams and word length	Malmasi et al. (2016)
MACSAAR	Random Forest, SVM	Zipfian frequency distribution, word length	Zampieri et al. (2016)
MAZA	Meta-classifier	n-grams, word probability, word length	(Malmasi and Zampieri 2016)
Melbourne	Weighted Random Forests	Lexical and semantic features	Brooke et al. (2016)
PLUJAGH	Threshold-based methods	Features extracted from Simple Wikipedia	Wróbel (2016)
Pomona	Threshold-based methods	Word frequencies	Kauchak (2016)
Sensible	Ensemble Recurrent Neural Networks	Word embeddings	Gillin (2016)
SV000gg	System voting with threshold	Morphological, lexical, and semantic features	Paetzold and Specia (2016b)
TALN	Random Forest	Lexical, morphological, semantic, and syntactic features	Ronzano et al. (2016)
USAAR	Bayesian Ridge classifiers	Hand-crafted word sense entropy metric and language model perplexity	Martínez Martínez and Tan (2016)
UWB	Maximum Entropy	Word occurrence counts on Wikipedia documents	Konkol (2016)

instance was used as a vote to build two ensemble models. The ensemble models were built using plurality voting which assigns the highest number of votes as the label of a given instance, and exploits an oracle which assigns the correct label for an instance if at least one of the systems predicted the ground truth label for that instance. The plurality vote serves to better understand the performance of the systems using the same dataset while the oracle is used to quantify the theoretical upper limit performance on the dataset (Kuncheva et al., 2001). The study showed that the potential upper limit for the CWI–2016 dataset considering the output of the participating systems is 0.60 F1 score for the complex word class. The outcome confirms that the low performance of the systems is related to the way the data has been annotated. Finally, this study also confirmed the relationship between word length and lexical complexity annotation in this dataset, a feature used by many of the teams participating in CWI–2016 as well as in our present work.

2.2 CWI–2018

Following the success of CWI–2016, the second edition, CWI–2018, was organized at the Workshop on the Innovative Use of NLP for Building Educational Applications (BEA) in 2018.[2] Unlike CWI–2016 which focused only on English, CWI–2018 featured English, French, German, and Spanish datasets opening new perspectives in research in this area.

A total of four tracks were available at CWI–2018: English, German, and Spanish, in which training and testing data was available for each language, and French. The organizers released a French test set with no corresponding training set with the goal of deriving models for French CWI from the English, Spanish, and German datasets. CWI–2018 featured two sub-tasks: (i) a binary classification task similar to CWI–2016 where participants were asked to label the given target word in a particular context as complex or simple; (ii) a probabilistic classification task where participants were asked to give a probability of the given target word in a particular context being complex.

In terms of data, CWI–2018 used the *CWIG3G2* dataset (Yimam et al., 2017) in English, German, and Spanish. The English dataset contains texts from three domains, *News*, *WikiNews*, and *Wikipedia* articles and the evaluation was carried out per domain. To allow cross-lingual learning, a dataset for French was collected using the same methodology as the one used for the CWIG3G2 corpus. Another important difference between CWI–2016 and CWI–2018 is that the *CWIG3G2* featured annotation of both single words and MWEs while the dataset used in CWI–2016 only considered single words.

In terms of participation, CWI–2018 attracted 12 teams in different task/track combinations. In Table 2, we list the approaches submitted to the English binary classification single word track by the 10 teams who wrote system description papers presented at BEA. Most teams tried multiple approaches and here we describe the teams' best-performing ones according to their system description papers.

[2] https://www.sites.google.com/view/cwisharedtask2018/.

Table 2 Systems submitted to the CWI–2018 English binary classification single word track. We include team names and a brief description of each system including features and classifiers used. A reference to each system description paper is provided for more information

Team	Classifiers	Features	Paper
CAMB	Adaboost	N-grams, WordNet features, POS tags, dependency parsing relations, psycholinguistic features.	Gooding and Kochmar (2018)
CFILT_IITB	Voting ensemble	Word length, syllable counts, vowel counts, WordNet-based features.	Wani et al. (2018)
hu-berlin	Naive Bayes	Character n-grams	Popović (2018)
ITEC	LSTM	Word length, word and character embeddings, frequency count, psycholinguistics features.	De Hertog and Tack (2018)
LaSTUS/TALN	SVM, Random Forest	Word length, word embeddings, semantic and contextual features.	AbuRa'ed and Saggion (2018)
NILC	XGBoost	N-grams, word length, number of syllables, WordNet-based features.	Hartmann and dos Santos (2018)
NLP-CIC	Tree ensembles and CNNs	Word frequency, syntactic and lexical features, psycholinguistic features, and word embeddings.	Aroyehun et al.et al. (2018)
SB@GU	Extra trees	Word length, number of syllables, n-grams, frequency distribution.	Alfter and Pilán (2018)
TMU	Random Forest	Word length, word frequency, probability features derived from corpora.	Kajiwara and Komachi (2018)
UnibucKernel	Kernel-based learning with SVMs.	Character n-grams, semantic features, and word embeddings.	Butnaru and Ionescu (2018)

For the English binary classification single word track, the organizers reported the performance by all teams per domain. Team CAMB obtained the best performance for the three domains: 0.8736 F1-score on News, 0.8400 F1-score on WikiNews, and 0.8115 F1-score on Wikipedia. We observed that for all teams the performance on the News domain was generally substantially higher than the performance obtained in the two other domains. Several teams used the opportunity to compare multiple approaches for this task and many of them reported that traditional machine learning classifiers were more accurate than deep neural networks (Hartmann & dos Santos, 2018; Alfter &Pilán, 2018).

3 Analysis of features of complex words

Upon analysing the datasets and system features used in CWI–2016 and CWI–2018, we noticed several intuitive explanations as to why a word may be judged as complex, or not:

- The word is archaic.
- The word is a borrowing from another language or refers to a concept that is atypical in the culture of the reader.
- The word is uncommon and many people are not generally exposed to it.
- The word refers to a very specialised concept.
- Although the word is common, it is being used with an uncommon meaning in the given context.

These possible characteristics motivated us to represent input words as sets of indicative linguistic features for the purpose of CWI. We used 378 features to represent words in our data set. These include psycholinguistic features derived from the MRC database (Wilson, 1988), word embeddings, and several other features with the potential to capture our intuitions about lexical complexity.

Values of the psycholinguistic features of words were obtained using the API to the MRC database. Many of the resources included in the database were built before 1998. These were derived through rigorous psycholinguistic testing, and as a result are of restricted size (offering relatively poor coverage of current English vocabulary). For this reason, in addition to specifying the values of these features directly from the database, we included binary features to indicate whether or not the word occurs in the MRC database.

We used information about whether or not the Wikipedia entry for the word includes an infobox element to indicate its degree of specialisation. Wiki-pedia[3] describes infoboxes as:

> […] a fixed-format table usually added to the top right-hand corner of articles to consistently present a summary of some unifying aspect that the articles share and sometimes to improve navigation to other interrelated articles. Many infoboxes also emit structured metadata which is sourced by DBpedia

[3] https://en.wikipedia.org/wiki/Help:Infobox. Last accessed 16th September 2021.

and other third party re-users. The generalized infobox feature grew out of the original taxoboxes (taxonomy infoboxes) that editors developed to visually express the scientific classification of organisms.

We observed that entries for specialised vocabulary (e.g. *Gharial*) frequently contain infobox elements of various types (e.g. *biota*). We extracted features encoding information about the occurrence and type of infobox element as an indicator of the level of specialisation of the word. We view this as a type of coarse-grained semantic information which is available for a relatively large proportion of words: more than 76% of those occurring in the CWI-2016 and CWI-2018 datasets.

The full feature set is displayed in Tables 3 and 4. Given that it encodes well-motivated psycholinguistic information and includes features which capture our intuitions about lexical complexity, we consider this feature set to be suitable for use in the derivation of models for CWI. We processed the human-annotated CWI–2016 and CWI–2018 datasets to represent words as feature vectors using the features in these tables.

Features P and S (Table 4) can be categorised as high coverage (holding for more than two thirds of the tokens in the annotated corpora); features G, E, J, H, Q, and I (Tables 3 and 4) as medium coverage (holding for more than one third but less than two thirds of the tokens in the corpora); and features F, R, N, O, K, B, D, M, L, and O (Tables 3 and 4) as low coverage (holding for less than one third of the tokens in the corpora).[4]

Considered individually, the great majority of features/feature sets listed in Table 3 have no linear relationship with the averaged human judgement of word complexity in the CWI 2016 and CWI 2018 datasets. The only exceptions are word length (feature group C) and the word's frequency count in the London-Lund corpus (feature group H). As the distributions of these two features are non-normal, we measured correlation with the averaged complexity ratings of words using Spearman's rho. We found that normalised word length has a low positive correlation ($\rho(28\,677) = 0.435, p < 0.001$) while the frequency of the word in the Brown corpus has a low negative correlation with word complexity ($\rho(28\,677) = -0.354, p < 0.001$). It is worth noting that MWEs in the CWI-2018 data are always complex and this may have influenced the results for word-length as MWEs are typically longer than single words.

There is no linear relationship between the values of features/feature sets listed in rows K–S of Table 4 and the averaged values of word complexity assigned by the annotators. In our experiments, we did not investigate the strength of correlations between individual word embedding features and average complexity ratings.

Given that the distributions of our features are non-normal, we used Levene's test (Levene, 1960) to assess the homogeneity of variance between word feature values and complexity scores. In all cases, the Levene test statistic exceeded critical values and obtained $p < 0.01$, indicating no equality of variance between complexity scores and feature values.

[4] These features are listed in decreasing order of coverage provided.

Table 3 Features (A–J) used to represent words

ID	Feature	Type	Definition
A	Frequent	Binary	One of the 10 000 most frequent words listed in Wiktionary
B	Archaic	Binary	Listed in an archaic word list.[†]
C	Length (normalised)	Numerical	Length of the word divided by 50.[‡]
D	Plurality	Binary	5 features indicating whether the word is plural, has no plural form, is a singular form, is both singular and plural form, or is plural but acts singular.
E	Familiarity	Numerical (100-700)	Familiarity score, derived by merging three sets of norms: Paivio (unpublished; these are an expansion of the norms of Paivio et al. (1968)), Toglia and Battig (1978), and Gilhooly and Logie (1980)). See Wilson (1988) for more details on these metrics
F	Concreteness	Numerical (100-700)	Concreteness score, listed in the MRC Database
G	Imageability	Numerical (100-700)	Imageability score of the word, listed in the MRC Database
H	Brown	Numerical	Frequency count of the word in the London-Lund Corpus of English Conversation (Svartvik and Quirk 1980)
I	KF_{FREQ}	Numerical	Frequency count of the word in the Kučera and Francis (1967) frequency list, derived from the Brown corpus.
J	TL_{FREQ}	Numerical	Frequency listed in Thorndike and Lorge (1944) L count, which combines the counts of morphological variants of the word in a reference corpus

[†]Available at https://archive.org/stream/dictionaryofarch028421mbp/dictionaryofarch~028421mbp_djvu.txt. Last accessed 26th February 2019

[‡]Longest word in English being 45 characters (pneumonoultramicroscopicsilicovolcanoconiosis)

Table 4 Features (K–T) used to represent words

ID	Feature	Type	Definition
K	MEANC	Numerical (100-700)	Meaningfulness rating of the word as provided by the Colorado norms of Toglia and Battig (1978)
L	MEANP	Numerical (100-700)	Meaningfulness rating of the word as provided by the norms of Paivio (unpublished)
M	AOA	Numerical (100-700)	Age of acquisition, as provided by the norms of Gilhooly and Logie (1980)
N	$TQ2_Q$	Binary	Morphological variant of another word in the dictionary
O	$TQ2_2$	Binary	Ends in the letter R and this R is not pronounced except when the next word begins with a vowel
P	WTYPE	Binary	9 features indicating the word type (adverb, conjunction, interjection, adjective, noun, past participle, pronoun, verb, or other) as listed in the Shorter Oxford English Dictionary or Webster's New International Dictionary
Q	STATUS	Binary	7 features indicating the word status (archaic, alien, obsolete, colloquial, rare, and standard) as listed in the Dolby database (Dolby et al., 1963)
R	STRESS	Binary	14 features indicating the stress pattern of the word when pronounced. Where 2 is a strongly stressed syllable, 1 is medium stressed, and 0 is an unstressed syllable, the 14 stress patterns are: 0, 01020, 010200, 02, 020, 0200, 10020, 102, 1020, 10200, 20, 200, 2000, and 22
S	INFOBOX	Binary	13 features indicating the type of infobox present in the English Wikipedia page for the word. Infobox types are: AMBIGUOUS, BIOGRAPHY_VCARD, BIOTA, BORDERED, COLLAPSIBLE_ AUTOCOLLAPSE, DEFAULT, GEOGRAPHY_VCARD, HPRODUCT, NONE, VCARD, VCARD_PLAINLIST, VEVENT, and VEVENT_HAUDIO
T	Word Embeddings	Numerical	300 features are the vector representation of the word derived using GloVe (Pennington et al., 2014)

Clearly, this is a surprising result. Research in psycholinguistics indicates, for example, that the frequency of a given word (feature groups A, H, I, and J) affects its perception (Segui et al., 1982; Dupoux & Mehler, 1990; Marslen-Wilson, 1990), that word familiarity (feature group E) and frequency affect visual and auditory word recognition (Connine et al., 1990), and that word imageability (feature group G) significantly impacts word reading accuracy and rate of word learning among first and second graders at risk for reading disabilities (Steacy & Compton, 2019). Further, the word "concreteness effect" (feature group F) is a well-established concept in psycholinguistics with the tendency of words with tangible physical referents being learned earlier, recognised faster, and recalled with less effort than words with abstract referents (Paivio, 1991; Schwanenflugel, 1991). Schwanenflugel et al. (1988) proposed that abstract words are more difficult to recognise because their interpretation is more reliant on context than is the case for concrete words. Word meaningfulness (feature groups K and L) has been observed to have a positive effect on word recognition (Leeds, 1976) and words with great meaningfulness have been found to be easier to recall than words with less meaningfulness (Kinoshita, 1989). Finally, the age of acquisition of words (feature group M) has been reported to be a predictor of the speed of reading words aloud and lexical decision tasks (in which participants are asked to judge whether particular sequences of characters are real words), with words acquired early in life being responded to more quickly than words acquired later in life (Morrison & Ellis, 2000). We would therefore expect to see more of our features correlating with complexity. This is likely to be a factor of the annotation protocols used in the datasets we analysed and motivates our wider argument in this work that there is a need for new CWI datasets. The two features that we did identify as showing correlation with word complexity (length and frequency) are both features that are used in almost all of the systems for the shared tasks at CWI–2016 and CWI–2018 as shown in Tables 1 and 2 respectively. This indicates that these features are useful for complexity both in our correlation analysis and in the empirical results of the systems that have submitted using these features. We include this here to show the lack of correlation between sensible features and those datasets. In our next section, we will discuss the deficiencies of these datasets, as well as proposed our specification for an improved CWI dataset.

4 Specification for CWI data protocol

In the previous Section we analysed differing features of complexity. In this section, we first highlight some of the design decisions that were taken in the creation of prior CWI datasets. We continue by proposing a specification, based on our prior analysis, for a new CWI dataset that improves on prior work. Our specification is designed to enable CWI research in areas that have not previously been explored. As well as providing a specification, we also provide a list of features for future datasets to implement in Table 6.

Table 5 CWI Datasets compared according to their features

Dataset	Binary	Probabilistic	Continuous	Context	Multi-Genre
CWI–2016	×			×	×
CWI–2018	×	×		×	×
Maddela–2018 (Maddela and Xu 2018)			×		

'Binary', 'Probabilistic' and 'Continuous' refer to the nature of the annotated labels. 'Context' refers to the presence of sentential context at annotation time and 'Multi-Genre' refers to the dataset drawing from sources across many genres

4.1 Building on previous datasets

The previous datasets for CWI have interesting characteristics that make them useful for the CWI task. A quick overview of these datasets is presented in Table 5, where they are compared according to some of their basic features.

The first dataset we have considered is the CWI–2016 dataset, which provides binary annotations on words in context. 9,200 sentences were selected and the annotation was performed as described below in Paetzold and Specia (2016a):

> Volunteers were instructed to annotate all words that they could not understand…A subset of 200 sentences was split into 20 sub-sets of 10 sentences, and each subset was annotated by a total of 20 volunteers. The remaining 9,000 sentences were split into 300 subsets of 30 sentences, each of which was annotated by a single volunteer.

The annotators were asked to identify any words for which they did not know the meaning. Each annotator had a different proficiency level and therefore will find different words more or less complex - giving rise to a varied dataset with different portions of the data reflecting differing complexity levels. Further, each instance in the test set was annotated by 20 annotators, whereas each instance in the training set was annotated by a single annotator. For the test set, any word which was annotated as complex by at least one annotator was marked complex (even if the other 19 annotators disagreed). This is problematic as the training data is not representative of the testing data, making it hard for supervised systems to do well on this task. Binary annotation of complexity requires an annotator to impose a subjective threshold on the level at which they transition from considering a word complex as opposed to simple. An annotator's background, education, etc. may affect where this threshold between complex and simple terms should be set. Further, it is likely that one annotator may find words difficult that another finds simple and vice-versa. Factors such as the annotator's native language, educational background, dialect, etc. all affect the type of words they are familiar with. In the case of the training data where 20 annotators have all annotated the same instance and any instance with at least one annotation is considered complex, it may be taken that the annotations represent some form of maximum complexity - i.e., that any word is above the lowest possible threshold of complexity. However, in the case of the test set where each

word is annotated by a single annotator, the annotations are harder to interpret. Each instance is personal, reflecting only a single annotator's judgment.

Moving on from the CWI–2016 dataset, the CWI–2018 dataset also provides binary annotations, which were aggregated to give a 'probabilistic' measure of complexity. CWI–2018 invited participants to submit results on both the binary complexity annotation setting and the probabilistic annotation setting. To collect their data, the organisers of CWI–2018 followed a similar principle as in CWI–2016. Sentences were presented to annotators and the annotators were asked to select any words or phrases that they found to be complex. As in CWI–2016, the annotation task in CWI–2018 was subjective, with potentially low agreement between annotators. In the probabilistic setting, at least 20 annotations were collected from native and non-native speakers and each word was given a score indicating the proportion of annotators that found that word to be complex. (i.e., if 10 out of 20 annotators marked the word, then it would be given a score of 0.5). A useful property of this style of annotation is that words are seen on a probablistic scale of complexity. However, the aggregation of binary annotations to give continuous annotations does not necessarily tell us about the complexity of the word itself. Instead it tells us about the annotators, and how many of them will consider a word complex. So, for example a score of 0.5 does not indicate a median level of complexity (or some sort of neutrality between simple and complex), but instead should be interpreted as indicating that 50% of the annotator pool will consider this word complex.

The final dataset we have covered was published in 2018 by Maddela and Xu (2018). We refer to this as Maddela–2018 for brevity. In this dataset, 11 annotators who spoke English as a second language were employed to annotate a portion of 15,000 words on a 6-point Likert scale with 5–7 annotations being collected for each vocabulary item. Words were presented without context, with the annotators guessing or making assumptions about the sense of the word at annotation. Different annotators may have considered the word to have a different sense or to have been used in a different context. Almost all words are polysemous and the different senses of the words are likely to have different levels of complexity - particularly in a coarse grained sense setting (e.g., *mean* average vs. a *mean* person). The main effect here is that the varied complexities of the multiple senses and usages of a word are conflated into a single annotation. There is no information as to which word sense the annotators were giving the annotations for, and as such the annotations may be unreliable in cases where a word is used in an uncommon sense. In the Likert-scale type annotation, it is less of an issue that annotators' opinions will vary than in the binary setting used in CWI–2016 and CWI–2018, as each annotator's judgment is aggregated on a common continuous scale. This means that the final averaged annotation is reflective of the average complexity that a word might have in a general setting. This is making an assumption that the annotations are normally distributed and that a mean average is valid in this case. A normality test could be used to quantify whether instances are likely to have normal distributions, however with only 5–7 annotations per instance, this may not be reliable.

So far, we have mainly considered complex words. However, the complexity of multi-word expressions is a valuable addition to the CWI literature. MWEs can be considered as compositional or non-compositional. Compositional MWEs (e.g., christmas tree, notice board, golf cart, etc.) take their meaning from the constituent words in the MWE, whereas non-compositional MWEs do not (e.g., hot dog, red herring, reverse ferret,etc.). It is reasonable to assume that complexity will follow a similar pattern to semantics and that compositional MWEs will be dependent on the constituent words to give the complexity of the expression, whereas the complexity of non-compositional MWEs will be independent of the constituent words. In the previous datasets, only the CWI–2018 dataset asked annotators to highlight phrases as well as single words, giving a limit of 50 characters to prevent overreaching. Participants in the task were asked to also give complexity annotations for the highlighted phrases. The system with the highest overall score reported that they found it easier to always consider MWEs as complex in the binary setting (Gooding & Kochmar, 2018). The work of Maddela and Xu (2018) also considers MWEs. Although they do not annotate for these, instead using average pooling to combine the embeddings of each token in a phrase into a single embedding, which is then processed in the same way as for single words. As described previously, this assumes compositionality, which will not always be the case.

Little treatment has been given to the variations in complexity between different parts of speech. None of the previous datasets annotate specifically for part of speech except for the CWI–2016 shared task data, which explicitly asks annotators to only highlight content words in the target sentences. Again, this is an important consideration as the roles of nouns, verbs, adjectives and adverbs are different in a sentence and considering them as different entities during annotation will help to better structure corpora. Developers of the existing corpora that span POS tags all suggest the use of POS as a feature for classification—demonstrating its importance in CWI.

All of the corpora recognise the importance of a diversity of reader backgrounds in their corpus construction. Native speakers of English might not realise that certain words they know well (depending on their socio-cultural biases) are not commonly known or may falsely assume that they "find all words easy". All three of the corpora that we have studied include annotations by non-native speakers. The CWI–2016 dataset used crowdsourcing to get annotations from 400 non-native speakers, the CWI–2018 dataset used native and non-native speakers (collecting at least 10 annotations from each for every instance). The Maddela–2018 data used 11 non-native speakers. The use of non-native speakers for CWI annotation may lead to models trained using these datasets being useful for identifying words which are complex to non–native speakers, but may not be applicable to other groups.

All the datasets are heavily biased towards text which has not been professionally edited. The CWI–2016 dataset compiles a number of sources taken from Wikipedia and Simple Wikipedia, the CWI–2018 dataset takes Wikipedia, WikiNews and one formal set of news text sources. The Maddela-2018 dataset uses the Google Web1T (Brants and Franz 2006) (taken from a large web-crawl) to identify the most frequent 15,000 words in English and re-annotates each for complexity. Except for the news texts in the 2018 data, all of these sources are written for informal purposes

and will contain spelling mistakes, idioms, etc. There has been little prior work exploring cross-genre learning for CWI, however it is unlikely that models trained on such informal text will be appropriate for identifying complexity in formal texts.

4.2 Specification

In the remainder of this section we will describe some of the qualities of an ideal dataset for CWI. Our recommendations are summarised at the end of this Section in Table 6. This specification is intended to give general purpose recommendations for anyone seeking to develop a new CWI dataset.

The key issue with the shared task datasets was the subjectivity that arose during the annotation process due to their treatment of complexity as a binary notion. When multiple annotators are asked to "mark any complex word" they will each draw on their subjective definition of complexity, and each will choose a different subset of words to be annotated as complex. The annotations that result from this are probabilistic in nature and tell us more about the annotators than the words themselves. Future datasets should consider providing measures which attempt to give more objectivity and move towards consensus between annotators. Of course, any complexity annotation involving human participants will always rely on the participants subjective knowledge and hence will be dependent on the participants. More objective measures of continuous complexity could be given by asking annotators to mark words on a Likert scale as by Maddela-2018, or by looking at external measurements of the ability of people to read the words, such as lexical access time, eye tracking, etc.

There are two factors to be considered here when measuring word complexity. One is the perceived complexity of a word (how difficult an annotator estimates a word to be) and the other is the actual complexity of a word (how much difficulty that word presents to the reader) (Leroy et al., 2013). Clearly these are both important factors in estimating a word's complexity and although we may expect them to be correlated there is no guarantee they will be aligned. Whereas perceived complexity affects how a user may prejudge a text, actual complexity determines the degree with which a reader is likely to struggle.

Of course, any measure of complexity which is derived by asking humans to give a subjective judgment of how difficult they find a word is bound to give a measure of perceived rather than actual complexity. In fact, measuring actual complexity would only be possible if the human was taken out of the loop altogether (even a setting where the reader doesn't know they are being assessed would rely on a participant's innately subjective assessment of each word). Any annotation scheme which focusses on continuous complexity judgments is still inviting perceived complexity assessment. By giving more levels to the assessment of complexity (i.e., through a Likert Scale assessment) the annotators have more ability to better record their perception of the complexity of the words that are being assessed.

The only previous dataset to present continuous annotations (Maddela-2018) did so in the absence of context. Context is key to determining the usage and meaning of a word and the same word used in different contexts can vary greatly in both

Table 6 A list of recommended features for future CWI dataset development

ID	Feature	Description
1	Continuous annotations	Complexity labels should be on a continuous scale ranging from least to most difficult
2	Context	Tokens should be presented in their original contexts of usage
3	Multiple token instances	Each token should be included several times in a dataset
4	Multiple token annotations	Each token should receive many annotations from different annotators
5	Diverse annotators	The fluency and background of annotators should be as diverse as possible
6	Multiple genres	The text sources used to select contexts should cover diverse genres
7	Multi-word expressions	These should be considered alongside single word tokens as part of an annotation scheme

semantics and complexity. Indeed, a familiar word in an unfamiliar context may be just as jarring as a rare word for a reader, who is forced to quickly update their mental lexicon with the new sense of the word they have encountered (e.g. words like *base*, *boss*, and *fanning* in the domains of chemistry, architecture, and geology, respectively). Datasets should include context for any words that annotations are provided for. This will help systems to identify how contextual factors affect the complexity of a given instance. When presenting context, researchers may wish to either ask annotators to mark every word in a sentence according to some complexity judgment (dense annotation) or they may wish to pick a target word in a context and ask only for a judgment of the complexity of this word (sparse annotation). In the dense annotation setting, it is likely to be possible to get a much higher throughput of complexity annotations, as the reader will need to only read a sentence once to give multiple annotations, however they are likely to be deeply influenced by the meaning of the sentence, and may struggle to disassociate this from their annotation of complex words themselves. In the sparse annotation setting more contexts are required to give a comparative number of instances compared to the dense annotation setting, however the annotation given is more likely to be a direct result of the token itself, rather than the sentence. Any such sparse annotation task should be set up to ensure that an annotator gives judgments based on the word in its context (i.e., that they read and understand the context), rather than just giving a judgment based on the word, as if no context were presented.

Given that we are recommending that the data is presented in context, there is a strong argument for presenting multiple instances of each word. If only one instance of a word were presented in context, then it may be the case that this word had a specific usage that was not representative of its general usage. Words are polysemous (Fellbaum, 2010) and this is true both at the coarse grained (tennis *bat* vs. fruit *bat*) and narrow grained levels (I *love* you vs. I *love* London). The coarse grained level represents different meanings or etymologies, whereas the fine-grained level may

represent a similar meaning but a different intensity (as in our example). The provision of multiple instances of a word allows both of these factors to be taken into account. This consideration should be held in balance with the need to have a diversity of tokens. If a dataset has N instances, constituting P occurrences of R words, then we suggest that $R \gg P$. I.e., the number of total words should be much larger than the number of instances of each word. There is more to be gained in a dataset by having a diversity of tokens than by having many annotations on each token. An interesting separate task would be to annotate many instances of one word form for complexity and analyse how the context affects this. However, this is a secondary task to the one we are presenting here of assessing a word's complexity.

Each instance in a new CWI dataset should be viewed and annotated by multiple people, ideally from a spectrum of ability levels. Multiple annotations have been a common theme of the previous CWI datasets we have discussed, with datasets using as many as 20 annotators per instance. All subsets (train, dev, test) of a dataset should be annotated by the same number of annotators, or at the very least by annotators drawn from the same distribution. This ensures that all subsets of the data are comparable. More annotations allows us to capture a wider array of viewpoints from annotators of varying ability levels. If the annotators are carefully selected to ensure they represent a mixture of ability levels then this will lead to annotations that are representative. Consider the case where all annotators are of low ability, or of high ability. The resulting annotations may lead to all words being assigned to the most or least complex categories respectively. This may be desirable in user- or genre-specific settings, but is not desirable for general-purpose LCP. There are two potential approaches to selecting a pool of annotators and distributing annotations between them. Firstly, a researcher may choose to use a fixed number of annotators, such that each annotator views every data instance once. In this setting, each data instance receives N annotations, where N is the number of annotators chosen. Secondly, the annotations may be distributed across a wider pool of annotators, where given N annotators each sees a randomised subset of the data. In this setting, a researcher may choose to control how many instances each annotator sees, ensuring an even distribution of annotators across the data instances. The second approach is more appropriate in a crowd-sourcing setting, where a researcher has diminished ability to control who takes on which job.

Previous CWI datasets for English have given a strong focus on non-native speakers as discussed above. Non-native speakers have learnt English as a foreign language and the assumption in using them for CWI research is that they will have only learnt a simple subset of English that allows them to get by in daily tasks. However, a non-native speaker may range from a new immigrant who has recently arrived in an English speaking country to someone who has lived there for decades. Further, both native and non-native speakers may simultaneously be specialists in some domains and novices in other domains. Non-natives may be specialists in domains where natives are not, and vice versa, influencing their complexity judgments. We would suggest, that whilst non-native speakers should not be excluded from the CWI annotation process, they should not be relied upon either. Instead the pool of annotators should be selected for their general ability in English, not for their mother tongue. Indeed, when selecting non-native speakers it may be worth considering

selecting a variety of mother-tongues, as it is the case that different languages, or language families will have cognates and near-cognates with English, making it easier for non-native speakers of certain backgrounds to understand words in English with roots in their mother tongue.

Allowing for multiple genres gives more diversity in the type of text studied and allows systems that are trained on it to generalise better to unseen texts. This prevents overfitting to one text-type, leading to results being more reliable and hence more interpretable, and ultimately leads to the creation of useful models that can be applied across genres. CWI resources should name the source genres that their texts are taken from and comply with licences placed on those genres. Whilst informal, or amateur text is in abundance (e.g., Twitter or Wikipedia), formal texts should also be considered for CWI such as professionally written news, scientific articles, parliament proceedings, legal texts or any other such texts that are written for a professional audience. These texts provide well structured language, which is typically targeted at a specific audience and is of a difficult quality for those outside that audience. These texts contain a higher density of complex words and as such are useful examples of the types of text that might need interventions to improve their readability for a lay reader.

As discussed previously, MWEs are an important element in complexity as previous studies have shown that MWEs are generally considered more complex by a user than individual words (Gooding & Kochmar, 2018). Any new CWI dataset should consider incorporating MWEs as they will certainly be useful for future CWI research. When we consider that MWEs can range from simple collocations (White House), to verbal phrases (pick up) and may span 2 or more words, across parts of speech—including phrasal MWEs (it's raining cats and dogs)—it is clear that the number of potential MWEs to consider is much wider than the number of single tokens. How do we select appropriate MWEs to cover? There is no particular advantage to CWI in selecting one category of MWE over another, but we suggest that any dataset covering MWEs explicitly names the types of MWE that it has covered. By incorporating MWEs, a dataset may be used to investigate both the nature of complexity in those MWEs and in the constituent tokens. Strategies for identifying MWEs, as well as the different types of MWEs are beyond the scope of this work and we would direct the reader to the MWE literature (Sag et al., 2002; Schneider et al., 2014) for a more comprehensive treatment of this problem.

5 CompLex 2.0

In this Section we describe a new CWI dataset that we have collected. Our new dataset, dubbed 'CompLex 2.0' builds on prior work (CompLex 1.0 Shardlow et al., 2020), in which we collected and annotated tokens in context for complexity. We have described the data collection process for CompLex 1.0 as below and then the annotation process that we undertook to extend this data to CompLex 2.0. CompLex 2.0 covers more instances than CompLex 1.0 and crucially, has more annotations per instance than CompLex 1.0, making it more reliable. We present statistics on

our new dataset and describe how it fits the recommendations we have made in our specification for new CWI datasets above. CompLex 2.0 was used as the dataset for the SemEval Shared Task on Lexical Complexity Prediction in 2021.

5.1 Data collection

The first challenge in dataset creation is the collection of appropriate source texts. We have followed our specification above and selected three sources that give a sufficient level of complexity. We aimed to select sources that were sufficiently different from one another to prevent trained models generalising to any one source text. The sources that we used are described below.

- Bible: We selected the World English Bible translation (Christodouloupoulos & Steedman, 2015). This is a modern translation, so does not contain archaic words (thee, thou, etc.), but still contains religious language that may be complex. The inclusion of this text gives language that combines narrative and poetic text that uses language typically familiar for a reader, yet interspersed with unfamiliar named entities and terms with specific religious meanings (*propitiation*, *atonement*, etc.).
- Europarl: We used the English portion of the European Pariliament proceedings selected from europarl (Koehn 2005). This is a very varied corpus concerning a wide range of issues related to European policy. As this is speech transcription, it is often dialogical in nature in contrast to our other two corpora. Again, the style of text is generally familiar as it is transcriptions of debates. However technical terminology relating to the topics of discussion is present, raising the difficulty level of this text for a reader.
- Biomedical: We selected articles from the CRAFT corpus (Bada et al., 2012), which are all in the biomedical domain. These present a very specialised type of language that will be unfamiliar to non-domain experts. Academic articles present a classic challenge in understanding for a reader and are typically written for a very narrow audience. We expect these texts to be particularly dense with complex words.

In addition to single words, we also selected targets containing two tokens. We used syntactic patterns to identify these MWEs, selecting for adjective-noun or noun-noun patterns. We discounted any syntactic pattern that was followed by a further noun to avoid splitting complex noun phrases (e.g., noun-noun-noun, or adjective-noun-noun). We used the StanfordCoreNLP tagger (Manning et al., 2014) to get part-of-speech tags for each sentence and then applied our syntactic patterns to identify candidate MWEs.

Clearly this approach does not capture the full variation of MWEs. It limits the length of each to 2 tokens and only identifies compound or described nouns. Some examples of the types of MWE that we identify with this scheme are given in Table 7. Whilst this inhibits the scope of MWEs that are present in our corpus, this does allow us to make a focused investigation on these types of MWEs. Notably, the

Table 7 The varied types of MWEs that can be captured by our syntactic pattern matching. NN indicates a Noun-Noun pattern, whereas JN indicates an Adjective-Noun pattern

Pattern	MWE	Type
NN	storage box	Compound Noun
JN	ready meal	Described Noun
JN	electric vehicle	Compositional
NN	hot dog	Non-compositional
JN	European Union	Named Entity

types of MWE that we have identified are those that are the most common (compound nouns, described nouns, compositional, non-compositional and named entities). The investigation of other types of MWEs may be addressed by other, more targeted studies following our recommendations for CWI annotation.

For each corpus we selected words using frequency bands, ensuring that words in our corpus were distributed across the range of low to high frequency. We selected the following eight frequency bands according to the SUBTLEX frequencies in order of least to most frequent (i.e., most to least complex): 2–4, 5–10, 11–50, 51–250, 251–500, 501–1400, 1401–3100, 3101–10000. We excluded the rarest words (those with a frequency of only 1) as well as the most frequent (those above 10,000) in order to ensure that our instances were well-attested content words. As frequency is correlated to complexity (Brysbaert et al., 2011), this ensures that our final corpus will have a range of high and low complexity targets. We chose to select 3000 single words and 600 MWEs from each corpus to give a total of 10,800 instances in our corpus. We selected a representative number of instances from each frequency band to give the desired total number of instances in each corpus. We automatically annotated each sentence with POS tags and only selected nouns as our targets, in-keeping with our MWE selection strategy. We allowed a maximum of 5 instances of a token to be selected in each genre (ensuring that contexts were different). This maximises the total number of examples of each instance, whilst still allowing some variation in the selection of tokens. There is a theoretical minimum of 600 instances of single words and 120 MWEs that could occur in our corpus (each with 5 occurrences in each of the three genres. Table 11 shows that the number of repeated instances is much lower. This is a factor of the stochastic selection procedure that we have employed. We have included examples of the contexts and target words in Table 8.

5.2 Data labelling

As has been previously mentioned, prior datasets have focused on either (a) binary complexity or (b) probabilistic complexity. Neither of which give a true representation of the complexity of a word. In our annotation we chose to annotate each word on a 5-point Likert scale, where each point was given the following descriptor:

1. Very Easy: Words which were very familiar to an annotator.
2. Easy: Words for which an annotator was aware of the meaning.

3. Neutral:	A word which was neither difficult nor easy.
4. Difficult:	Words for which an annotator was unclear of the meaning, but may have been able to infer the meaning from the sentence.
5. Very Difficult:	Words that an annotator had never seen before, or were very unclear.

We used the following key to transform the numerical labels to a 0-1 range when aggregating the annotations: $1 \rightarrow 0$, $2 \rightarrow 0.25$, $3 \rightarrow 0.5$, $4 \rightarrow 0.75$, $5 \rightarrow 1$. This allowed us to ensure that our complexity labels were normalised in the range 0–1.

We initially employed crowd workers through the Figure Eight platform (formerly CrowdFlower), requesting 20 annotations per data instance and paying \$0.03 per annotation. We selected annotators from English speaking countries (UK, USA and Australia). In addition, we used the annotation platform's in-built quality control metrics to filter out annotators who failed pre-set test questions, or who answered a set of questions too quickly.

After we had collected these results, we further analysed the data to detect instances where annotators had not fully participated in the task. We specifically analysed instances where an annotator had given the exact same annotation for all instances (usually these were all 'Neutral') and discarded these from our data. We retained any data instance that had at least 4 valid annotations in our final dataset.

This led to the version of the dataset we described as CompLex 1.0. Whilst this dataset evidenced the trends we expected to see, the conclusions we were able to draw from it were weaker than we hoped (Shardlow et al., 2020). The median number of annotators was 7 per instance, and we identified this as an area for improvement. The involvement of more annotators would allow more opinions to be expressed, leading to better average judgments.

For the second round of annotations we used the Amazon Mechanical Turk platform. We used exactly the same data as in the original annotation of CompLex 1.0 and requested new annotations for each instance. We gave the same instructions to annotators regarding the Likert-scale points. As there is no in-built quality control in Mechanical Turk, we opted to release the data in batches (1200 instances at a time). We asked for a further 10 annotations per instance and paid at a rate of \$0.03 per annotation. We reviewed the annotators work in between batches, rejecting accounts which submitted annotations too quickly, or without correlation to the other annotator's judgments. We also measured the correlation with lexical frequency to ensure that the annotations we were receiving were in the range we expected.

This allowed us to gather a further 108,000 annotations on the CompLex data. These new judgments were aggregated with those from CompLex 1.0 to give a new dataset—CompLex 2.0. We used this data to run a shared task on Lexical Complexity Prediction at SemEval 2021 (Shardlow et al., 2021).

5.3 Corpus statistics

The first round of annotations led to an initial version of the Corpus (CompLex 1.0), for which we have shown the statistics originally reported in Table 9. Due to

Table 8 Examples from our corpus, the target word is highlighted in bold text

Corpus	Context	Complexity
Bible	This was the **length** of Sarah's life	Low
Biomed	[…] cell **growth** rates were reported to be 50% lower […]	Low
Europarl	Could you tell me under which rule they were enabled to extend this item to have four rather than three **debates**?	Low
Europarl	These agencies have gradually become very important in the **financial world**, for a variety of reasons	Medium
Biomed	[…] leads to the **hallmark loss** of striatal neurons […]	Medium
Bible	The **idols** of Egypt will tremble at his presence […]	Medium
Bible	This is the law of the **trespass offering**	High
Europarl	They do hold elections, but candidates have to be endorsed by the conservative clergy, so **dissenters** are by definition excluded	High
Biomed	[..] due to a reduction in **adipose** tissue	High

The field *Complexity* refers to perceived complexity

the quality control that we employed for this round of annotation, we discarded a large portion of our original judgments and only kept instances with four or more annotations. This is evident in the fact that only 9476 instances out of our original 10,800 are present in this iteration of the corpus. Additionally, the median number of annotators was 7 across our corpus (with the range being from 4 to 20). Retaining only the annotations in which we could be certain of the quality was a difficult choice, as it reduced the amount of data available. However, the mean complexities of the sub-corpora were in line with our expectations. With Biomedical text being on average more complex than the other two genres.

This led us to undertake our second round of annotation in order to develop CompLex 2.0 ready for the SemEval shared task. We have included statistics on the annotations aggregated from both rounds in Table 10. 513 separate annotators viewed our data, with each annotator seeing on average 542 instances across all rounds of annotation (around 5% of our corpus). We gathered a total of 278,093 annotations, paying $0.03 per annotation. The average time spent per annotation was 21.61 seconds, which means that we paid our workers at an average rate of 5 US Dollars per hour. The task received reviews indicating that annotators found it to be well paid in comparison to other tasks on the platform. We gathered an average of 25.75 annotations per instance, this is an increase over CompLex 1.0, which only had on average 7 annotations per instance. We expect that by having more annotations per instance, we will have more reliable average estimates of the complexity of each word.

We report detailed statistics on our new dataset, CompLex 2.0, in Table 11. We can see that in total 5,617 unique tokens covering single words and multi-word expressions are distributed across 10,800 contexts. Whilst the contexts are split evenly between each genre (3,600 each) the number of repeated words is higher in the Biomed and Bible corpora, with more distinct words occurring in the Europarl corpus. The complexity annotations are low at 0.321 for the entire corpus, indicating

Table 9 The statistics for CompLex 1.0. We report on the entire corpus and also present a breakdown of statistics by **Genre**

Genre	Contexts	Unique words	Complexity
All	9476	5166	0.394
Europarl	3496	2194	0.390
Biomed	2960	1670	0.407
Bible	3020	1705	0.385

We include statistics on the number of **Contexts**, the number of **Unique Words** and the mean **Complexity** in each partition

that the average complexity of words is somewhere between points 2 (0.25—a word which that the annotator was aware of the meaning) and 3 (0.5—A word which was neither difficult nor easy) on our Likert scale. This indicates that annotators generally understood the words in our dataset. The annotations did use the full range of our Likert scale and the dataset contains words of all complexities. We can see from the data that the Biomedical genre was on average more difficult to understand (0.353) than the other genres (0.303 for Europarl and 0.307 for Bible respectively). Multiword expressions are markedly more complex (0.419) than single words (0.302), with the same genre distinctions as in the full data.

5.4 Inter-annotator agreement

Achieving strict adherence to annotation guidelines is difficult in the crowd-sourcing setting as there is little time to train, test or survey annotators. As a result, inter-annotator agreement tends to be lower in this context. We provided some controls as outlined above to ensure that annotators were fully participating in the task and that their annotations aligned with those of other annotators. In our setting, we do not necessarily expect annotators to agree in every case as one may legitimately consider a word to be complex, whilst another considers it to be simple. A reasonable expectation is that annotators will provide similar annotations to each other, and that the annotations will mostly fall into one category. We expect the distribution of annotations for one instance to be normally distributed. We have already made this assumption, as we take the mean to give the average complexity.

To test this, we used a Shapiro-Wilk test (Shapiro & Wilk, 1965), which gives a number in the range of 0-1 indicating how likely a given distribution is to follow the normal distribution. For each of our instances, we perform the test on the annotations for that instance. A higher number indicates that the instance has annotations which are more likely to be normally distributed, whereas a low number on this test indicates a non-Gaussian distribution, such as a multi-modal distribution. A histogram of this data is displayed in Fig. 1. This shows that the majority of our data obtains a score between 0.7 and 0.9 according to the Shapiro-Wilk test, with a peak around 0.85. This indicates that our data is generally normally distributed, and hence that annotators generally gave annotations that centered around a mean value.

In Table 12 we have shown a number of examples from our corpus that do not follow the distribution that we may have expected. These were infrequent in

Table 10 Statistics on the round of evaluation undertaken with Mechanical Turk

Number of annotators	513
Number of instances	10,800
Number of annotations	278,093
Annotations per Instance	25.75
Instances per annotator	542.09
Time per annotation	21.61 (s)

Table 11 The statistics for CompLex 2.0

Subset	Genre	Contexts	Unique words	Complexity
All	**Total**	**10,800**	**5617**	**0.321**
	Europarl	3600	2227	0.303
	Biomed	3600	1904	0.353
	Bible	3600	1934	0.307
Single	**Total**	**9000**	**4129**	**0.302**
	Europarl	3000	1725	0.286
	Biomed	3000	1388	0.325
	Bible	3000	1462	0.293
MWE	**Total**	**1800**	**1488**	**0.419**
	Europarl	600	502	0.388
	Biomed	600	516	0.491
	Bible	600	472	0.377

We report on the entire corpus and also present a breakdown of statistics by **Genre** and by **Subset**. We include statistics on the number of **Contexts**, the number of **Unique Words** and the mean **Complexity** in each partition

our corpus, but are displayed here to help the reader understand where annotators may have disagreed. In example 1 the simple word 'heaven' was given to annotators, most of whom assigned it to the *Very Easy* category. However, 3 annotators disagreed with this, assigning it to the *Neutral* category. Possibly, the annotators found the word easy, but the metaphorical usage harder to grasp. Example 2 shows a similar disagreement, albeit around a more difficult word. 'Election' is a word that most people living in a democracy will have encountered, yet 5 people felt it was neither easy, nor difficult—placing it in the *Neutral* category. Our third example, taken from the Biomedical genre, demonstrates a word (Granules) which is considered *Easy* by 14 annotators, yet is considered *Difficult* by 4 annotators. Whilst 'Granules' is not a particularly rare word, it may be considered complex by some in this instance due to its contextual usage in the biomedical literature. Example 4 shows a word which is specific to biblical language ('Cubit'). Although the annotations gave a reasonably Gaussian set of annotations (0.848 according to the Shapiro-Wilk statistic), they were split over all 5 potential categories. This is an example of annotators' previous familiarity with the text. Those who know a cubit is an ancient measure of length will score it on the easier side

of the Likert scale, whereas those who have not seen the word before will score it as more difficult. The remaining three examples (5–7) all score similarly highly on the Shapiro-Wilk test, however they have a wide spread of annotations. Again, this is likely due to the familiarity of the annotators with each word.

5.5 CompLex 2.0 features

We have presented a corpus that was developed according to the recommendations that we have set out earlier in this work (see Sect. 4.1). Whilst we have made every effort to follow these, practical concerns have led us to pragmatic design decisions that made the development of our corpus feasible. In the following list, we itemise the design decisions that were made during the construction of our corpus and show how these link to the recommendations from Table 6.

1. Continuous annotations: We have implemented this using a Likert Scale as described above. Unlike Maddela-2018 who used a 4-point Likert scale, we chose a 5-point Likert scale to allow annotators to give a neutral judgment. To give final complexity values we took the mean average of these annotations, transforming the complexity labels in the range 0–1.
2. Context: We presented annotations in context to the annotators and explicitly asked annotators to judge a word based on its contextual usage (but not on the context itself). Following (Peirce 1906), we distinguish between word types (the distinct words used in a text, which comprise its vocabulary) and word tokens (the different occurrences or instances of those words throughout the text). There are clear variations in the complexity of different tokens sharing the same word type. For example, the word 'table' receives a higher complexity rating in the less common sense of 'table a motion' than in the more frequent sense of something being 'on the table'.
3. Multiple tokens: We presented a maximum of 5 tokens per word type, per genre. This led to 5,617 word types across 10,800 tokens and contexts giving an average density of 1.92 contexts per word type. Although some word types do appear in multiple contexts 3,423 words appear with only a single context. 671 word types feature 5 or more tokens (and contexts). This is a compromise between our desire to include a wide variety of word types in the dataset and to include multiple tokens of each type. A dataset featuring a more rigorous treatment of contexts may reveal the role of context in complexity estimation in a way that our data is not able to.
4. Multiple token annotations: We have described our process of gathering an average of 25.75 annotations per token. We could have chosen to do fewer annotations in favour of annotating more tokens, however we prioritised having a large number of judgments per token to give a more consistent and representative averaged annotation.
5. Diverse annotators: We did not place many restrictions, or record demographic information regarding our annotators. Doing so may have helped to better understand the makeup of our annotations and identify potential biases. We did not

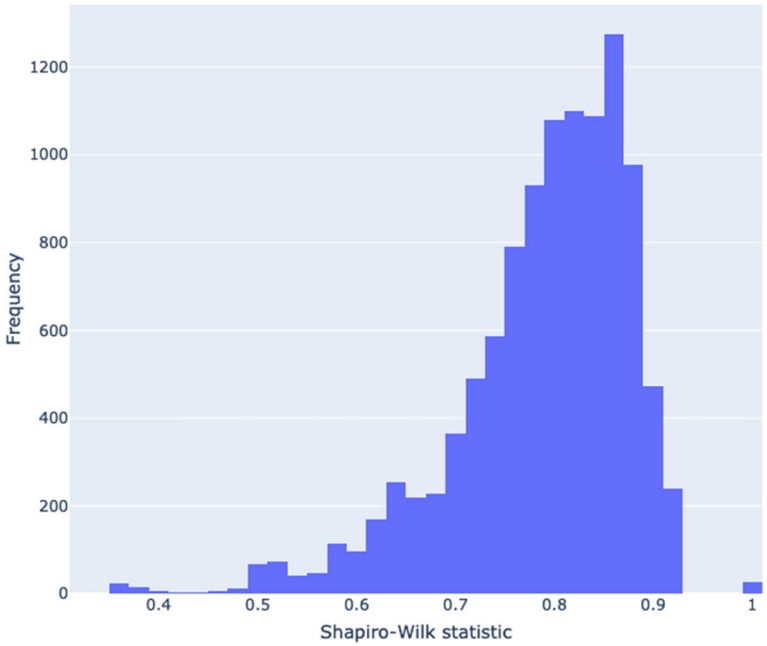

Fig. 1 A histogram of Shapiro-Wilk's test statistics, demonstrating the likelihood for each instance that the annotations are normally distributed

record this information due to the crowd-sourcing setting that we used. This is something for future LCP annotation efforts to consider.

6. Multiple genres: We have selected three diverse genres with a potential for complex language. We deliberately avoided the use of Wiki text as this has been studied widely already in CWI.
7. Multi-word expressions: We have included these in a limited form as part of our corpus. The MWEs make up 16.66% of our corpus. We have included these as an interesting area to study and we hope that their inclusion will shed light on the complexity of MWEs. Further studies could focus on specific types of MWE, extending our research.

The CompLex 2.0 corpus is designed according to the recommendations we have set out. In particular, we do not record demographic information on our participants and as such cannot make reasonable claims as to the diversity of our annotators. Our corpus is intended as a starting point for future LCP researchers to build on. Using the methodology described in this section, further datasets encoding information about complex words can be annotated, focusing on the remaining open research questions in lexical complexity prediction.

Table 12 Examples of annotations with interesting distributions indicating disagreement among annotators respectively

ID	Corpus	Context	Annotations					
			VE	E	N	D	VD	S-W
1	Bible	You will have treasure in **heaven**	24	1	3			0.423
2	Europarl	**Election** of Vice-Presidents (first, second and third ballots)	19	1	5			0.544
3	Biomed	Annexin A7 was isolated as the agent that mediated aggregation of chromaffin **granules** and fusion of…		14	2	4		0.612
4	Bible	Ehud made him a sword which had two edges, a **cubit** in length; and he wore it under his clothing on his right thigh	2	3	4	12	8	0.848
5	Europarl	I therefore wanted to tell you that I inadvertently voted "yes' in the vote on the Cornelissen report on the first part of **recital** 0, when I intended to vote "no'	1	11	3	9	2	0.848
6	Europarl	The Rospuda valley is the last **peat** bog system of its kind in Europe	2	11	3	4	4	0.848
7	Biomed	Amyloid burden worsens significantly with age, and by 9 mo, the **hippocampus** and cortex of untreated mice are largely filled with aggregated peptide	1	8	7	9	2	0.901

The target word in the context is highlighted in bold. S-W stands for Shapiro-Wilk. Levels of Annotations are Very Easy (VE), Easy (E), Neutral (N), Difficult (D), and Very Difficult (D)

6 Predicting categorical complexity

We represented words and multiword units in the CWI–2016 (Paetzold & Specia, 2016a), CWI–2018 (Yimam et al., 2018), and the new CompLex 2.0 single word and multiword datasets using features which, on the basis of previous work in lexical simplification (Paetzold, 2016), text readability (Yaneva et al., 2017; Deutsch et al., 2020), psycholinguistics/neuroscience (Yonelinas et al., 2005), and our inspection of the annotated data, we consider likely to be predictive of their complexity (Sect. 3).

We used the `trees.RandomForest` method distributed with Weka (Hall et al., 2009) to build baseline lexical complexity prediction models exploiting the features presented in Sect. 3. In the experiments described in the current Section, we framed the prediction as a classification task with continuous complexity scores mapped to a 5-point scale. The points on these scales denote the proportions of annotators who consider the word complex (c): few ($0 \leq c < 0.2$), some ($0.2 \leq c < 0.4$), half ($0.4 \leq c < 0.6$), most ($0.6 \leq c < 0.8$), and all ($0.8 \leq c \leq 1$).

Table 13 displays weighted average F-scores and mean absolute error (MAE) scores obtained by the baseline models in the ten-fold cross validation setting. This table includes statistics on the number of instances to be classified in each dataset.

Table 14 displays the results of an ablation study performed in order to assess the contribution of various groups of features to the word complexity prediction task applied in the four datasets: CWI–2016, CWI–2018, CompLex (single words), and CompLex (multi-words). The feature sets refer to those studied previously in this work in Sect. 3. In the table, negative values of ΔMAE indicate that the features are helpful, reducing the mean absolute error of the classifier. The reverse is true of positive values.

Our results indicate that for prediction of lexical complexity in the CWI–2016 dataset, five of the ablated feature groups are useful. Features encoding information about word length and the regularity of the singular/plural forms of nouns, the typical age of acquisition of the words, and the broad syntactic categories of the words improve the accuracy of the classifier, as do word embeddings.

For words in the CWI–2018 dataset, no feature group was found to be particularly useful for prediction of lexical complexity, though a simple model based only on word length information outperformed the default baseline exploiting all features. Again, this may be due to the typically longer MWEs present in the CWI–2018 dataset, which are exclusively labelled as complex.

When predicting the lexical complexity of individual words in the CompLex 2.0 data, features encoding information about whether or not the word was archaic, about the regularity of the singular/plural forms of nouns, and about the stress patterns of the words were all found to be useful. When considering multiword units (bigrams), a far larger proportion of the feature groups was observed to be useful for lexical complexity prediction. In our ablation study of bigrams, we assigned the bigram the average value of each feature (all of the features were represented numerically, including binary and one hot representations, and none of the features were symbolic). We found that features encoding information about word frequency, whether or not the words were archaic, word length, regularity of singular/plural

Table 13 Evaluation results of the baseline `trees.RandomForest` classifier

Dataset	F-score (weighted average)	MAE	Instances
CWI 2016	0.915	0.04	2237
CWI 2018	0.843	0.0681	11, 949
CompLex 2.0 (single)	0.607	0.1782	7233
CompLex 2.0 (MWE)	0.568	0.2137	1465

forms, standard age of acquisition, broad syntactic category, the word's status as either archaic, alien, obsolete, colloquial, rare, or standard, the stress pattern of the word, and the occurrence of an INFOBOX element in the Wikipedia entry for the word were all useful predictors of lexical complexity. Averaged word embeddings also improved the accuracy of predictions made by the `trees.RandomForest` classifier in the CompLex (multi) dataset.

In the CWI–2016 and CWI–2018 datasets, we applied Weka's attribute (feature) ranking method with the unsupervised `Principal Components` Attribute Transformer evaluator to the 378 numerical features described previously in Tables 3 and 4 (Sect. 3). Table 15 displays the ten top-ranked groups of features for the four datasets. The main observations to be drawn from the feature selection study is the usefulness of information related to word familiarity, concreteness, and imageability in all datasets and information from the vector representations of words derived using GloVe (Pennington et al., 2014). These features occur in the systems that participated in the CWI Shared Tasks as shown in Tables 1 and 2. This corroborates our findings in line with previous work.

Interestingly, whereas the results presented previously using a correlation analysis did not find psycholinguistic features (Groups E,F,G,K) to be correlative with complexity, the principal component analysis indicates that these features are in fact likely to be useful for prediction in these datasets.

These results demonstrate that by using our new data from CompLex 2.0, the features that we expect to correlate well with complexity judgments are more likely to be effective features for classification than when annotations are done in a binary setting as in the CWI–2016 and CWI–2018 datasets.

7 Predicting continuous complexity

In our final section, we use the data we have collected to discuss the nature of complex words from a different perspective than in Sect. 6. Whereas in the previous Section we converted all labels into a categorical format to allow comparison, in this Section we use the labels assigned to CompLex 2.0 to discuss factors affecting the nature of lexical complexity, and its prediction. We first look at the effects of genre on CWI. We then continue in our exploration to study the distribution of annotations, to determine how and when annotators agree on the complexity of a word.

Table 14 Results of feature ablation

Ablated feature group	CWI–2016 ΔMAE	CWI–2018	CompLex (single)	CompLex (multi)
A	+1E-04	0	+0.0002	−0.0002
B	+0.0002	+0.0002	−1E-04	−0.0006
C	−0.0001	+0.0004	+1E-04	−0.0004
D	−0.0001	+0.0001	−1E-04	−0.0002
E, F, G	0	+0.0001	+0.0003	+0.0006
H, I, J	+0.0002	+0.0004	+0.0007	+0.0012
M	−0.0001	+0.0001	0	−0.001
N	0	0	+1E-04	+0.0005
P	−0.0002	+0.0001	+0.0002	−0.0007
Q	0	+0.0001	+1E-04	−0.0002
R	+1E−04	0	−1E−04	−0.0009
S	+1E−04	+0.0001	0	−1E-04
Linguistic features (A-S)	0	+0.0009	**+0.0018**	**+0.0027**
T	**− 0.0029**	**+0.002**	**+0.001**	**− 0.0065**
All but C	**+0.0093**	**− 0.0681**	**+0.0469**	**+0.0396**

Positive numbers represent a higher MAE after the named feature group was removed (hence the feature was helpful), whereas negative numbers represent the opposite. Most deltas are small, indicating minimal effect from many features. Values above 0.001 or below -0.001 are highlighted in bold

7.1 Prediction of complexity across genres

To better understand the effect of text genre on the LCP task we designed the experiments described in this Section. For these, we employed a simple linear regression with the features described previously in Sect. 3. We use the single words in the corpus and split the data into training and test portions, with 90% of the data in the training portion and 10% of the data in the test portion. We first created our linear regression using all the available training data and evaluated this using Pearson's Correlation. We used the labels given to the data during the annotation round we undertook to create CompLex 2.0. The prediction model based on linear regression achieved a score of 0.771, indicating a reasonably high level of correlation between its predictions and the labels of the test set.

This result is recorded in Table 16, where we also show the results for each genre. In each case, we have selected only data from a given genre and followed the same procedure as above, splitting into train and test and evaluating using Pearson's correlation. The linear regression model is less closely correlated when making lexical complexity predictions in the Europarl (0.724) and in the Bible data (0.735). This is expected, given the reduction in size of the training data. It is surprising to see that the linear regression model worked better for the Biomedical data than for any other subset (0.784). This may indicate that simple and complex words are more distinct in this corpus and that this distinction can be learnt from a more focused training set.

Table 15 Results of feature selection (`PrincipalComponents`)

Rank	CWI–2016	CWI–2018	CompLex (single)	CompLex (multi)
1	E, F, G, K	E, F, G	E, F, G	T (subset)
2	T (subset)	E, F, G	T (subset)	T (subset)
3	H, I, J , T (subset)		T (subset)	E, F, G
4	T (subset)	T (subset)	T (subset)	T (subset)
5	T (subset)	T (subset)	D, N, A	T (subset)
6	T (subset)	D, T (subset)	T (subset)	T (subset)
7	T (subset)	T (subset)	T (subset)	T (subset)
8	T (subset)	T (subset)	T (subset)	T (subset)
9	T (subset)	T (subset)	T (subset)	T (subset)
10	T (subset)	P, T (subset)	T (subset)	T (subset)

To further determine the effects of genre on lexical complexity prediction, we constructed a new linear regression model that was trained and tested using specific genres selected from our corpus. We trained on single genres and tested on each of the other 2 genres, as well as training on a combined subset of 2 genres and testing on the remaining genre. The results for this experiment are shown in Table 17. We were able to build a reliable predictive model for cross-genre complexity prediction in each case.

Our results show that there is a drop in performance when training on out-of-domain data, compared to training on in-domain data. This is true across all genres, where a reduction of between 0.119 and 0.297 can be observed in Pearson's correlation. In each genre, the scores improve when training on the other two genres, rather than just on one. This may be due to the effect of multiple genres helping the linear regression to generalise to global complexity effects, rather than overfitting to specific complexity features in one genre. If we were to test our results on an additional genre/domain, we may hope to see that training on three genres (as are present in our corpus) would yield even more generalised results.

7.2 Subjectivity

We previously used a Shapiro-Wilk test to demonstrate that our annotations are generally normally distributed. We obtained the mean of each annotation distribution to give a complexity score for each instance in our dataset. An interesting question to ask is how representative these means are of the true complexity of a word. One word may be considered easy by one annotator, yet difficult by another. Factors such as age, education and background may well affect which words a reader is familiar with. We can use the normally distributed annotations to understand this phenomenon by investigating the standard deviations of the annotations for each instance.

We have provided examples from our corpus in Table 18 with both the mean complexity and the standard deviation (σ) of the annotations. The top three rows show

Table 16 Results of training a linear regression on all the data, and on each genre

Subset	Correlation
All	0.771
Europarl	0.724
Biomed	0.784
Bible	0.735

Table 17 Results of training a linear regression on one genre, or pair of genres and testing on a different genre

Train	Test	Correlation
Biomed	Europarl	0.542
Bible	Europarl	0.484
Biomed + Bible	Europarl	0.651
Bible	Biomed	0.487
Europarl	Biomed	0.630
Bible + Europarl	Biomed	0.723
Biomed	Bible	0.605
Europarl	Bible	0.616
Biomed + Europarl	Bible	0.692

examples of high standard deviation, whereas the bottom three rows show examples of low standard deviations. It is clear from the table that annotators generally agree more about words which are less complex, with disagreements tending to happen around the more difficult words. An analysis of the mean complexity and standard deviation of the complexity yields a Pearson's correlation of 0.621, indicating that these are moderately correlated (disagreement is linearly related to complexity).

8 Discussion

Our work has sought to introduce a new definition of lexical complexity to the research community. Whereas previous treatments of lexical complexity have considered it a binary phenomenon in the Complex Word Identification (CWI) setting, we have extended this definition to lexical complexity prediction (LCP), considering complexity as a continuous value associated with a word. This new definition asks the question of 'how complex is a word' rather than 'is this word complex or not?'. This question allows us to give each token a complexity rating on a continuous scale, rather than a binary judgment. If binary judgments were required, it would be easy to create them using our dataset by imposing a threshold at some point in the data. By imposing thresholds at different points, binary labels can be obtained to suit different subjective definitions of complexity. Further, by implementing multiple thresholds, multiple categorical labels can be recovered from the data.

In Sect. 3 we showed that the types of features we would typically expect to correlate with word complexity did not show any correlation with the CWI–2016 and

CWI–2018 datasets. This motivated our analysis of the protocol underlying the annotation of these datasets and our development of a new protocol for CWI annotation. In Sect. 6, we were able to show through the use of feature ablation experiments that more of the feature sets that we used were relevant to the classification of CompLex 2.0, than were relevant to the annotation of CWI–2016 or CWI–2018. This implies that the annotations in our new dataset are more reflective of traditional measures of complexity.

We discussed the existing CWI datasets at length (Sect. 3), culminating in our new specification for LCP datasets in Sect. 4. Whilst we have gone on to develop our own dataset (CompLex 2.0), we also hope to see future work developing new CWI datasets following the principles that we have laid out. Future datasets could focus on multilinguality, multi-word expressions, further genres, or simply extending our analysis to further tokens and contexts. Certainly, we do not see the production of CompLex 2.0 as an end point in LCP research, but rather a starting point for other researchers to build from. This is why we have included our protocol in detail—in order to ensure the replicability of our work in future research.

In moving from binary annotations to Likert-scale annotations, we have provided a new dataset, which gives continuous annotations based on a more objective measure of complexity. The binary setting could also be improved if more objective guidelines were provided to the annotators (e.g., instructions such as "identify words that are appropriate for an adult", or "identify words that are specific to a domain", as opposed to "identify words that **you** find difficult). In our comparison, we are comparing a subjective binary dataset to a (more) objective continuous dataset (of course, our dataset still relies on some degree of annotator interpretation of the Likert scale labels). We do not have the ability to compare an objective binary dataset to our data, as it does not exist to the best of the author's knowledge, however doing so would likely yield further interesting insights into the differences between continuous and binary lexical complexity.

We implemented our specification for a new LCP dataset, following the recommendations established in Sect. 3. This led to the creation of CompLex 2.0. In Sect. 5.5 we have explicitly compared our dataset to the recommendations we made in Table 6, and we would encourage the creators of future LCP datasets to do the same. This will ensure that datasets can be easily evaluated and compared at a feature level. The CompLex 2.0 dataset is available via GitHub[5]. We have made this data available under a CC-BY licence, facilitating its reuse and reproducibility outside of our work.

Our new LCP dataset is the first to provide continuous complexity annotations for words in context. The role of context in lexical complexity has not been widely studied and we hope that this dataset will go some way towards allowing researchers to work on this topic. Indeed, the evidence from our annotations shows that for a single token in multiple contexts, the complexity annotation of that token does vary. Further work is needed to prove that the variation is an effect of the contextual

[5] https://github.com/MMU-TDMLab/CompLex.

Table 18 Examples of instances with subjective (wide standard deviation) and certain (narrow standard deviation) annotations

Corpus	Context	Complexity	σ
Biomed	The first step requires generating a floxed allele in ES cells that will serve as the **substrate** for subsequent exchanges (RMCE-ready ES cell, Fig. 1)	0.556	0.433
Bible	The second came, saying, 'Your mina, Lord, has made five **minas**	0.433	0.423
Europarl	'Budget support' refers to the transfer of financial resources from a funding agency outside the partner country's treasury, under the **proviso** that the country abide by the agreed conditions governing payments	0.567	0.382
Biomed	Similarly, changes in **synaptic plasticity** due to Ca2+-permeable AMPARs [51,52,60], e.g., in piriform cortex, might alter odor memorization processes	0.975	0.077
Bible	Or were you baptized into the **name** of Paul?	0.000	0.000
Europarl	Therefore, I would like to ask, in accordance with the Rules of Procedure, for the **matter** to be referred to the competent body	0.175	0.118

occurrence, or difference in sense and not due to the stochastic nature of annotations resulting from crowdsourcing.

Although we gave annotators in our task a 5-point scale ranging from Very Easy to Very Difficult, we chose to aggregate the annotations to give a mean-average for each instance. This makes a fundamental assumption that the distance in continuous complexity space between each point on the Likert scale is constant. Obviously, there is no guarantee that such an assumption is true. The danger of this is that annotations may be falsely biased towards one end of the scale. For instance, if the distance between Very Easy and Easy is shorter than the distance between Easy and Neutral, then considering these as the same distance will falsely inflate complexity ratings. Another strategy could have been to take the median or mode of the complexity annotations to give a final value. The disadvantage of that approach would be that every instance would have an ordinal categorical label instead of a continuous label as we have advocated for. This would be a different problem to the one we have explored, and is left to future research.

We used categorical complexity to provide a feature analysis of our dataset and the prior CWI datasets in Sect. 6. We observed that a number of features were identified as useful for the prediction task, indicating that complexity is a matter of many factors and no single factor can be used to determine a word's complexity. Interestingly, in Table 13, we showed that in the categorical setting the CWI-2018 and CWI-2016 datasets both outperformed the CompLex 2.0 dataset. We are not trying to use the dataset here to demonstrate some superior performance, but rather demonstrate a comparative analysis of features that are useful for complexity prediction. This may indicate that systems wishing to return a categorical label (as those used in Sect. 6), could use probabilistic or categorical data for training and get better results than when using our data. Our continuous labels allow us to perform further interesting analyses into the nature of complex words as presented in Sect. 7.

We were able to use our data to show that complexity can be predicted across genres. This is encouraging as our dataset contains three diverse genres, and we can expect that the complexity annotations we have identified will generalise well to other genres. A model trained on all three genres will learn features of complexity that are common to all genres, rather than to any one specific genre. We also demonstrated that our instances vary in subjectivity of complexity, with those rated as more complex typically being more subjective. Identifying the factors that make a word subjectively complex would be an interesting line of study, but is left for future research.

The ambiguity of a word is likely to play a role in its complexity. Words which are often mistaken for others are more likely to be confused and hence are likely to be rated as more difficult to understand by a reader. Conversely however, there is a well documented direct correlation between polysemy and frequency (i.e., infrequent words are typically monosemes, whereas frequent words have many senses. See the WordNet entries for 'run', 'bat', 'cat', etc.). It may be hypothesised that ambiguity and frequency need to be jointly taken into account when investigating lexical complexity, with a likely ordering from least to most complex being: (high-frequency, monosemous), (high-frequency, polysemous), (low-frequency, monosemous), (low-frequency, polysemous). Prior efforts have been undertaken to create sense annotated complexity datasets (Strohmaier et al., 2020), and building upon our research with sense annotations, using the specification given in Sect. 4.2 will lead to fruitful research outcomes.

9 Conclusion

We have demonstrated that previous datasets are insufficient for the task of Complex Word Identification. In fact, the very definition of the task—identifying complex words in a subjective binary setting rather than on an objective continuous scale is at fault. We have advocated for a generalisation of this task to Lexical Complexity Prediction and we have provided recommendations for datasets approaching this task. Further to this we have provided a new dataset, CompLex 2.0, which is the first publicly available dataset to provide continuous complexity annotations for words in context. We release the data in full to allow future researchers to join us in this exciting task of Lexical Complexity Prediction.

References

AbuRa'ed, A., Saggion, H. (2018). LaSTUS/TALN at Complex Word Identification (CWI) 2018 Shared Task. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Alfter, D., Pilán, I. (2018). SB@GU at the Complex Word Identification 2018 Shared Task. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Aroyehun, S. T., Angel, J., Alvarez, D. A. P., Gelbukh, A. (2018). Complex word identification: convolutional neural network vs. feature engineering . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W. A., Cohen, K. B., Verspoor, K., Blake, J. A., et al. (2012). Concept annotation in the craft corpus. *BMC Bioinformatics, 13*(1), 161.

Bingel, J., Schluter, N., Martínez Alonso, H. (2016). CoastalCPH at SemEval-2016 Task 11: The importance of designing your Neural Networks right. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1028–1033. Association for Computational Linguistics, San Diego, California.

Biran, O., Brody, S., Elhadad, N. (2011). Putting it simply: A context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers* (ACL-2011), pp. 496–501. Portland, Oregon.

Bott, S., Rello, L., Drndarevic, B., Saggion, H. (2012). Can Spanish be simpler? LexSiS: Lexical Simplification for Spanish. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science* (pp. 8–15). Springer, Samos, Greece.

Brants, T., Franz, A. (2006). The google web 1t 5-gram corpus version 1.1. LDC2006T13.

Brooke, J., Uitdenbogerd, A., Baldwin, T. (2016). Melbourne at SemEval 2016 Task 11: Classifying Type-level Word Complexity using Random Forests with Corpus and Word List Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 975–981. Association for Computational Linguistics, San Diego, California.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A.M., Bölte, J., Böhl, A. (2011). The word frequency effect. Experimental Psychology.

Butnaru, A., Ionescu, R. T. (2018). UnibucKernel: A kernel-based learning method for complex word identification . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Choubey, P., Pateria, S. (2016). Garuda & Bhasha at SemEval-2016 Task 11: Complex Word Identification Using Aggregated Learning Models. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 1006–1010. Association for Computational Linguistics, San Diego, California.

Christodouloupoulos, C., & Steedman, M. (2015). A massively parallel corpus: The bible in 100 languages. *Language Resources and Evaluation, 49*(2), 375–395. https://doi.org/10.1007/s10579-014-9287-y

Connine, C., Mullennix, J., Shernoff, E., & Yelen, J. (1990). Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(6), 1084–1096.

Davoodi, E., Kosseim, L. (2016). CLaC at SemEval-2016 Task 11: Exploring linguistic and psycho-linguistic Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 982–985. Association for Computational Linguistics, San Diego, California.

De Hertog, D., Tack, A. (2018). Deep Learning Architecture for Complex Word Identification . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Deutsch, T., Jasbi, M., Shieber, S. (2020). Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 1–17. Association for Computational Linguistics, Seattle, WA, USA Online. https://doi.org/10.18653/v1/2020.bea-1.1

Devlin, S., Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic Databases* pp. 161–173.

Dolby, J. L., Resnikoff, H. L., MacMurray, F.L. (1963). A tape dictionary for linguistic experiments. In *Proceedings of the American Federation of information processing societies: Fall Joint Computer Conference*, pp. 419 – 423. Spartan Books, Baltimore, MD.

Dupoux, E., & Mehler, J. (1990). Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code. *Journal of Memory & Language, 29*, 316–335.

Fellbaum, C. (2010). Wordnet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and applications of ontology: Computer applications* (pp. 231–243). Amsterdam: Springer.

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

Gilhooly, K., & Logie, R. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods, 12*(4), 396–427.

Gillin, N. (2016). Sensible at SemEval-2016 Task 11: Neural Nonsense mangled in ensemble mess. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 963–968. Association for Computational Linguistics, San Diego, California.

Gooding, S., & Kochmar, E. (2018). 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*. Association for Computational Linguistics, New Orleans, United States.

Gooding, S., Kochmar, E. (2019). Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1148–1153. Association for Computational Linguistics, Florence, Italy.

Gooding, S., Kochmar, E., Sarkar, A., Blackwell, A. (2019). Comparative judgments are more consistent than binary classification for labelling word complexity. In *Proceedings of the 13th Linguistic Annotation Workshop*, pp. 208–214.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explorations Newsletter, 11*, 10–18. https://doi.org/10.1145/1656274.1656278

Hartmann, N., & dos Santos, L. B. (2018). NILC at CWI 2018: Exploring Feature Engineering and Feature Learning . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

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

Kajiwara, T., & Komachi, M. (2018). Complex word identification based on frequency in a learner corpus . In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 1537–1546). Association for Computational Linguistics, Sofia, Bulgaria.

Kauchak, D. (2016). Pomona at SemEval-2016 Task 11: Predicting Word Complexity Based on Corpus Frequency. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 1047–1051. Association for Computational Linguistics, San Diego, California.

Kinoshita, S. (1989). Generation enhances semantic processing? The role of distinctiveness in the generation effect. *Memory & Cognition, 17*, 563–571.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit*.

Konkol, M. (2016). UWB at SemEval-2016 Task 11: Exploring Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 1038–1041. Association for Computational Linguistics, San Diego, California.

Kuncheva, L. I., Bezdek, J. C., & Duin, R. P. (2001). Decision templates for multiple classifier fusion: An experimental comparison. *Pattern Recognition, 34*(2), 299–314.

Kuru, O. (2016). AI-KU at SemEval-2016 Task 11: Word Embeddings and Substring Features for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic*

Evaluation (SemEval-2016, pp. 1042–1046). Association for Computational Linguistics, San Diego, California.

Kučera, H., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Leeds, B.G. (1976). Kindergarten children and the influence of letter shapes and meaningfulness of vocabulary as factors influencing word recognition. Technical Report ED136250, Department of Health, Education & Welfare, National Institute of Education.

Leroy, G., Kauchak, D., & Mouradi, O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *International Journal of Medical Informatics, 82*(8), 717–730.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in Honor of Harold Hotelling*. Palo Alto, CA: Stanford University Press.

Maddela, M., & Xu, W. (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*, pp. 3749–3760. Association for Computational Linguistics, Brussels, Belgium. https://doi.org/10.18653/v1/D18-1410

Malmasi, S., Dras, M., & Zampieri, M. (2016). LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 996–1000. Association for Computational Linguistics, San Diego, California.

Malmasi, S., & Zampieri, M. (2016). MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 991–995. Association for Computational Linguistics, San Diego, California.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60.

Marslen-Wilson, W. (1990). *Activation, competition, and frequency in lexical acess*. Cambridge, MA: MIT Press.

Martínez Martínez, J. M., & Tan, L. (2016). USAAR at SemEval-2016 Task 11: Complex Word Identification with Sense Entropy and Sentence Perplexity. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 958–962. Association for Computational Linguistics, San Diego, California.

Morrison, C., & Ellis, A. W. (2000). Real age of acquisition effects in word naming and lexical decision. *British Journal of Psychology, 91 Pt. 2*(2), 167-180 180.

Mukherjee, N., Patra, B. G., Das, D., & Bandyopadhyay, S. (2016). JU_NLP at SemEval-2016 Task 11: Identifying Complex Words in a Sentence. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 986–990. Association for Computational Linguistics, San Diego, California.

Paetzold, G., & Specia, L. (2016). SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 560–569. Association for Computational Linguistics, San Diego, California.

Paetzold, G., & Specia, L. (2016). 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), pp. 969–974. Association for Computational Linguistics, San Diego, California.

Paetzold, G. H. (2016). Lexical simplification for non-native english speakers. Phd thesis, University of Sheffield.

Paetzold, G. H., Alva-Manchego, F., Specia, L. (2017). MASSAlign: Alignment and Annotation of Comparable Documents. In *The Companion Volume of the IJCNLP 2017 Proceedings: System Demonstrations*, pp. 1–4.

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

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/ Revue canadienne de psychologie, 45*(3), 255–287.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76*(1, Pt.2), 1–25.

Palakurthi, A., & Mamidi, R. (2016). IIIT at SemEval-2016 Task 11: Complex Word Identification using Nearest Centroid Classification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 1017–1021. Association for Computational Linguistics, San Diego, California.

Peirce, C. S. S. (1906). Prolegomena to an apology for pragmaticism. *The Monist, 6*, 492–546.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP*.

Popović, M. (2018). Complex Word Identification using Character n-grams. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Quijada, M., & Medero, J. (2016). HMC at SemEval-2016 Task 11: Identifying Complex Words Using Depth-limited Decision Trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 1034–1037. Association for Computational Linguistics, San Diego, California.

Ronzano, F., Abura'ed, A., Espinosa Anke, L., & Saggion, H. (2016). TALN at SemEval-2016 Task 11: Modelling Complex Words by Contextual, Lexical and Semantic Features. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016, pp. 1011–1016). Association for Computational Linguistics, San Diego, California.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '02, p. 1-15. Springer, Berlin, Heidelberg.

Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, M. T., Conrad, H., & Smith, N. A. (2014). Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), pp. 455–461. European Language Resources Association (ELRA), Reykjavik, Iceland.

Schwanenflugel, P. J. (1991). Why are abstract concepts hard to understand? In P. J. Schwanenflugel (Ed.), *The psychology of word meaning* (pp. 223–250). Mahwah, NJ: Erlbaum.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27*(5), 499–520.

Segui, J., Mehler, J., Frauenfelder, U., & Morton, J. (1982). The word frequency effect and lexical access. *Neuropsychologia, 20*(6), 615–627.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples)†. *Biometrika, 52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

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

Shardlow, M. (2013). The CW corpus: A new resource for evaluating the identification of complex words. In *The Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (PITR 2013). Association for Computational Linguistics, Sofia, Bulgaria.

Shardlow, M., Cooper, M., & Zampieri, M. (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), pp. 57–62. European Language Resources Association, Marseille, France.

Shardlow, M., Evans, R., Paetzold, G., & Zampieri, M. (2021). Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation* (SemEval-2021).

Sanjay, S. P., & Soman, K. P. (2016). AmritaCEN at SemEval-2016 Task 11: Complex Word Identification using Word Embedding. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 1022–1027. Association for Computational Linguistics, San Diego, California.

Steacy, L. M., & Compton, D. L. (2019). Examining the role of imageability and regularity in word reading accuracy and learning efficiency among first and second graders at risk for reading disabilities. *Journal of Experimental Child Psychology, 178*, 226–250.

Strohmaier, D., Gooding, S., Taslimipoor, S., & Kochmar, E. (2020). SeCoDa: Sense complexity dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 5962–5967. European Language Resources Association, Marseille, France. https://aclanthology.org/2020.lrec-1.730

Svartvik, J., & Quirk, R. (Eds.). (1980). *Handbook of semantic word norms*. Lund: Liver/Gleerups.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Teacher's College: Columbia University, New York, NY, USA.

Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ, USA: Erlbaum.

Wani, N., Mathias, S., Gajjam, J. A., & Bhattacharyya, P. (2018). The Whole is Greater than the Sum of its Parts: Towards the Effectiveness of Voting Ensemble Classifiers for Complex Word Identification. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, United States.

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*, 6–10.

Wróbel, K. (2016). PLUJAGH at SemEval-2016 Task 11: Simple System for Complex Word Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016), pp. 953–957. Association for Computational Linguistics, San Diego, California.

Yaneva, V., Orăsan, C., Evans, R., & Rohanian, O. (2017). Combining multiple corpora for readability assessment for people with cognitive disabilities. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 121–132. Association for Computational Linguistics, Copenhagen, Denmark. https://doi.org/10.18653/v1/W17-5013

Yimam, S.M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., & Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of BEA*.

Yimam, S.M., Štajner, S., Riedl, M., & Biemann, C. (2017). Multilingual and Cross-Lingual Complex Word Identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, RANLP 2017, pp. 813–822. INCOMA Ltd., Varna, Bulgaria.

Yonelinas, A. P., Otten, L. J., Shaw, K. N., & Rugg, M. D. (2005). Separating the brain regions involved in recollection and familiarity in recognition memory. *Journal of Neuroscience, 25*(11), 3002–3008.

Zampieri, M., Malmasi, S., Paetzold, G., & Specia, L. (2017). Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications* (NLPTEA 2017, pp. 59–63). Asian Federation of Natural Language Processing, Taipei, Taiwan.

Zampieri, M., Tan, L., & van Genabith, J. (2016). MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for ComplexWord Identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation* (SemEval-2016, pp. 1001–1005). Association for Computational Linguistics, San Diego, California.