Comparing Annotation Guidelines and Annotators: A Preliminary Investigation into the Lexical Complexities Assigned in Two Complex Word Identification Datasets

Kai North
Rochester Institute of Technology
Rochester, NY
kn1473@g.rit.edu

ABSTRACT

This preliminary investigation explores the effects of two types of annotation guidelines, together with two types of annotators, have on the quality of two complex word identification (CWI) datasets. The former refers to comparing an annotation guideline that exhibits a 5-point likert scale compared with the use of an annotation guideline that carries out pairwise or 6-point likert scale annotation. The latter refers to comparing US, UK and Australian annotators with international expert annotators.The two datasets are the CompLex dataset (Shardlow, Cooper and Zampieri, 2020) and the Word Complexity Lexicon dataset (Maddela and Xu, 2018). It was discovered that both datasets assigned word complexity differently. The Word Complexity Lexicon dataset was also found to have a low inter-annotator agreement. In addition, word length was discovered not to be a defining feature of complexity within the CompLex dataset. However, it was found to be an influential characteristic in the Word Complexity dataset. With these findings in mind, several recommendations have been provided so as to improve the quality of future CWI datasets.

Datasets:
The CompLex dataset and the Word Complexity Lexicon dataset have both been made publicly available at https://github.com/MMU-TDMLab/CompLex and https://github.com/mounicam/lexical_simplification.

1 INTRODUCTION

Complex Word Identification (CWI) is the task of automatically identifying which words within a given text may be considered complex (Zampieri, Tan and Genabith, 2016). It is a precursor for Lexical Simplification (LS) and in turn, Text Simplification (TS). LS aims to "replace complex words and expressions with simpler alternatives" (Paetzold and Specia, 2016). TS subsequently attempts to improve a text's "readability and understandability" (Shardlow, 2014) by reducing it's overall complexity.

1.1 Defining Word Complexity

Within CWI, complexity is used as a "synonym for difficulty" (Malmasi and Zampieri, 2016). A complex word is defined as a word that a variety of individuals, or a particular sample, may find difficult to recognise and/or understand.

1.2 Wider Impact of CWI

Prior studies have constructed CWI systems to meet differing text comprehension needs. Several systems have been designed to improve text readability for children (Kajiwara, et al., 2013; Maddela and Xu, 2018) or for those suffering from reading disabilities, such as dyslexia (Rello, et al., 2013). Other CWI systems specialize in "making texts more accessible to language learners" (Malmasi, Drasi and Zampieri 2016).

1.3 Lexical Complexity Prediction

In recent years, a new term has been coined: "Lexical Complexity Prediction" (Shardlow, Cooper and Zampieri, 2020). The introduction of this term has signified a shift in CWI literature. Previous research has investigated CWI as a binary classification task: classifying words as being either complex (1) or non-complex (0). New research has presented CWI as a multi-label classification task: classifying a word on a scale of lexical complexity, ranging from: 1). Very Easy to 5). Very Difficult (See Section 2.1.;Shardlow, Cooper and Zampieri, 2020) .

1.4 Guidelines and Annotators

Be it either binary or multi-label classification, all CWI tasks require a unique dataset. This dataset needs to be labelled so that its complex or non-complex words, along with their associated features, can be identified and then learnt by a classification algorithm. Shardlow, Cooper and Zampieri (2020) created such a dataset through the use of: 1). a manual annotation guideline that enforced likert scale annotation, and 2). English-speaking annotators that were obtained via crowd sourcing. A prior study by Maddela and Xu (2018) likewise created such a dataset for binary CWI. However, they alternatively used: 3). a manual annotation guideline that carried out pairwise annotation, and 4). non-native yet fluent English-speaking annotators from various international backgrounds.

1.5 Aim

The aim of this preliminary investigation is to see what effects these annotation guideline and annotator variables: 1). to 4)., may or may not have had on the types of words identified as being complex and, as a result, the overall quality of their corresponding CWI datasets. This paper hopes to provide a brief insight into what annotation guidelines and types of annotators work well in the creation of a CWI dataset.

2 ANNOTATION GUIDELINES

An annotation guideline, or scheme, is often referred to as "the most critical component" (Ide and Putsejovsky, 2017) for many rule-based or machine learning tasks, including CWI. It is a document that defines the labels and sometimes their associated features that are required for later classification. An annotation guideline may either be designed for human annotators, being a manual annotation guideline, or for automatic machine annotators, being an automatic annotation guideline (Ide and Putsejovsky, 2017).

CWI commonly relies on human annotators and thus manual annotation guidelines. These annotation guidelines have "operational definitions so that [each] human [annotator] looking at the same piece of data [will be] more likely to assign it the same label" (Ide and Putsejovsky, 2017). As such, the definitions provided within an annotation guideline, together with how many definitions there are and how they are presented: 5-point likert scale, 6-point likert scale or pairwise, are extremely influential variables in determining the overall quality of a CWI dataset. The overall quality of a CWI dataset is determined by whether the assigned labels provide an accurate sample representation of complex and non-complex words.

2.1 Likert Scale Annotation

Shardlow, Cooper and Zampieri (2020) constructed the CompLex dataset. This dataset consists of 9,699 words (Shardlow, Cooper and Zampieri, 2020). These words were ranked by their annotators in accordance to their "5-point likert scale annotation guideline" (Shardlow, Cooper and Zampieri, 2020). This annotation guideline allowed for the creation of a dataset with 5 distinct word complexity labels that a later classifier could then distinguish. These labels were defined in their annotation guideline as follows:

- **Very Easy** - A word which you are very familiar with.
- **Easy** - A word which you are aware of its meaning.
- **Neutral** - A word which you find neither difficult nor easy.
- **Difficult** - A word whose meaning you do not understand but can infer through context.
- **Very Difficult** - A word you have never seen before or do not understand.

2.2 Pairwise Annotation

Shardlow, Cooper and Zampieri (2020) were not the first to use a 5-point likert scale within a CWI annotation guideline. Previously, Maddela and Xu (2018) used a 5-point likert scale for the annotation of their own 15,000 word CWI dataset: the Word Complexity Lexicon dataset. Through experimentation, they found that a "6-point [likert scale annotation guideline] worked better than a 5-point likert scale [annotation guideline]" (Maddela and Xu, 2016). They claimed that this was because a 6-point likert scale "had a more natural two-step approach" (Maddela and Xu, 2016), whereby each ranking had an opposite ranking on the other side of the complexity spectrum. This "two-step approach" (Maddela and Xu, 2016) is hereby referred to as pairwise annotation. Their paper, however, does not detail the exact definitions of the 6 labels used.These label definitions may, therefore, not have been provided within their annotation guideline. These labels were:

- **Simple**
  1). Very simple
  2). Moderately Simple
  3). Simple
- **Complex**
  4). Complex
  5). Moderately Complex
  6). Very Complex

Unlike, Shardlow, Cooper and Zampieri (2020), Maddela and Xu (2018) did not test their Word Complexity Lexicon dataset on multi-label classification for CWI. Instead, they "injected [their 6-point likert scale dataset] into two state-of-the-art-systems [for binary CWI]" (Maddela and Xu, 2018). These being SV000gg and the nearest centroid classifier (Paetzold and Specia, 2016). Both of these system saw a significant increase in their performance (Maddela and Xu, 2018). This improvement may have been a consequence of Maddela and Xu's (2018) use of pairwise annotation. This being as prior to Maddela and Xu's (2018) new dataset, the two systems relied on a binary annotated dataset that only consisted of the two labels complex (1) or non-complex (0) (Paetzold and Specia, 2016).

3 ANNOTATORS

3.1 US, UK and Australian Crowd-Sourcing

Shardlow, Cooper and Zampieri (2020) employed annotators from several English speaking countries: USA, UK and Australia. These annotators were gained through crowd-sourcing and were paid "around 3 cents per annotation" (Shardlow, Cooper and Zampieri (2020). In an attempt to maintain annotation quality, they undertook three measures: 1). they only selected annotators who "were known to... provide high quality work" (Shardlow, Cooper and Zampieri (2020), 2). they paid each annotator generously in order to encourage them to take their time over each annotation, and lastly 3). they "filtered out annotators who had not participated in the task properly" (Shardlow, Cooper and Zampieri, 2020).

There are advantages as well as disadvantages regarding Shardlow, Cooper and Zampieri's (2020) choice of annotator. By choosing annotators only from English speaking countries, it may have narrowed their sample. For instance, the complex words identified in their CompLex dataset may only be complex to those from English-speaking countries, whereas those of a differing nationality may find such words more or less complex. Thus, future researchers using this dataset may be able to generalise their results across English-speaking countries more so than other countries. Nevertheless, even though these annotators were sourced from the UK, USA and Australia, it does not guarantee that they were native-English speakers. Hence, the prior assumption may be somewhat flawed.

By paying each annotator "3 cents per annotation" (Shardlow, Copper and Zampieri, 2020), they claimed that this would motivate their annotators to perform more extensive and thorough annotation. It could equally be argued that this would have the opposite desired effect. For instance, by having a financial incentive, each annotator may have been motivated to complete as many annotations as possible as quickly as possible. As such, they may have put little thought into which words were complex or non-complex or may have been prone to the observer's paradox: rating a word as being complex simply because that is what they assumed they were tasked to do. This being said, Shardlow, Cooper and Zampieri (2020) do state that they removed those annotators who "had not participated in the task properly" (Shardlow, Cooper and Zampieri, 2020). This may have included such individuals.

3.2 International Experts

Maddela and Xu (2018) chose "11 non-native but fluent English speakers" (Maddela and Xu, 2018) to be their annotators. These annotators were known to the researchers and were likely to have been college students with a certain degree of expertise in data annotation or linguistics. As a result, it may be assumed that their data annotation would be of a higher quality than compared to that of Shardlow, Cooper and Zampieri's (2020) crowd-sourced annotators.

Maddela and Xu's (2018) annotators had differing first-languages and they originated from differing countries. Maddela and Xu (2018) state that this "may have contributed to the variance in their judgements" (Maddela and Xu, 2018). They found, for instance, that they had a relatively low inter-annotator agreement of 0.64. This means that for a number of instances, their annotators disagreed upon which label to assign to a particular word (See Section 5.3.). It may, therefore, be assumed that some words may have been assigned the wrong label. In fact, Yimam et al. (2017) discovered that "native speakers have higher inter-annotator agreement than non-native speakers, regardless of the text genre". As such, Shardlow, Cooper and Zampieri's (2020) US, UK and Australian annotators may have produced more accurate annotations than compared to those created by Maddela and Xu's (2018) non-native English speaking annotators; this being regardless of their potential lack of expertise due to being crowd-sourced.

3.3 Summary

Shardlow, Cooper and Zampieri (2020) as well as Maddela and Xu (2018) have likewise built extensive CWI datasets. These datasets respectively being the CompLex dataset and the Word Complexity Lexicon dataset. Both datasets were built through different means, with each having their own merits and limitations. CompLex was generated through the use of a 5-point likert scale. It was also created quickly by a large number of crowd-sourced annotators residing in the US, UK and Australia. These annotators may have lacked validity, reliability and expertise knowledge in data annotation which may have affected the quality of the CompLex dataset.

The Word Complexity Lexicon dataset was created through the use of pairwise annotation and a 6-point likert scale. It was labelled by a handful of annotators with potential experience in linguistics or CWI, yet due to their differing first languages and regional origins, displayed a relatively low inter-annotator agreement of 0.64. This may have reduced the quality of the Word Complexity Lexicon dataset. Nevertheless, unlike the CompLex datatset, this dataset was found to further enhance two state-of-the-art binary classifiers: SV000gg and the nearest centroid classifier (Paetzold and Specia, 2016).

Given the aim of this preliminary investigation as well as the above prior research, the following research question was explored:

- **Research Question.** Does there exist any significant difference between the words assigned as being complex within the CompLex dataset compared to those assigned as being complex within the Word Complexity Lexicon dataset?

4 METHODOLOGY

Being a preliminary investigation, this study only sought to partially answer its research question. This was done so as to see whether further research is, or is not, warranted. As such, only two potential differences between each datasets' complex words were explored. These being: 1). complex word frequency, in the form of positive or negative keyness, and 2). complex word length.

4.1 Datasets

The two datasets: CompLex and the Word Complexity Lexicon dataset, were both made publicly available by their corresponding authors.They were, therefore, downloaded and converted to txt files.

4.1.1 Data Cleaning. A short python script was implemented in order to remove words within both datasets which were not assigned as being complex. These being words with a score less than 0.5 within the CompLex dataset and words with a score less than 3 within the Word Complexity Lexicon dataset. Both datasets were then compared. Any remaining complex words which did not appear at least once within the opposing dataset were then also removed. This produced two datasets which consisted of entirely complex words that were shared between both datasets. The two outputted datasets will now be referred to as the shared complex (SC) CompLex dataset and the shared complex (SC) Word Complexity Lexicon dataset.

4.2 AntConc

The corpus analysis software AntConc (Anthony, 2019) was used to examine and compare the two SC datasets. The following functions of AntConc were exploited: Word List, Keyword List, Concordance Tool, Clusters/N-Grams and lastly, Collocations. These functions allowed the most frequently assigned complex words to be presented. It also detailed their n-grams, hence context words, as well as their embedded sentence. These were useful for determining the meaning of the word and would likely be needed for further analysis within future research.

4.3 Calculating Keyness

Frequency values were automatically generated through AntConc's (Anthony, 2019) Keyword Function to determine a word's keyness. Keyness is a measure that signifies how "significantly more [or less] frequent [a word occurs] ... in the target dataset than compared to the reference dataset" (Evison, 2014). The target dataset being the SC CompLex dataset. The reference dataset being the SC Word Complexity Lexicon dataset.

Positive keyness describes a word that "appears significantly more frequent[ly]... in the target dataset than compared to a reference dataset" (Evison, 2014). Negative keyness, on the other hand, refers to when a word "appears significantly less frequent[ly]... in the target dataset than compared to a reference dataset" (Evison, 2014).

Keyness was calculated by Antconc through chi-squared:

$$\tilde{\chi}^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

The observed frequency values of each complex word were calculated: 'O'. Using these observed frequency values: 'O', their expected values: 'E' were then determined for each complex word. This was achieved by multiplying the observed frequency value: 'O' by the total number of times that each complex word appeared within each corpus: 'n'. This was then divided by the sum of the observed frequency value: 'O' over 'n'. The observed frequency values of each complex word: 'O' were then subtracted from their expected values: 'E' and then squared. Lastly, the returned values were divided by their expected values: 'E'. This outputted their chi-square value: 'x' (Stanford, 2019). The degree of freedom (df) associated with the complex word was then determined by the sum of the two datasets minus 1. This was then multiplied by the number of complex words in question minus 1. This always equated to 1x1 = df = 1. This df value was then used alongside the returned chi-square values: 'x' to determine whether any significant difference existed between the frequency of each complex word used within each SC dataset: their 'p-value' (Stanford, 2019). P-values over 0.05 were considered to indicate a significant difference (Kim, 2015).

It was assumed that per the differences in their annotation guidelines as well as between their annotators, the SC CompLex dataset would contain some complex words that demonstrated positive and/or negative keyness in comparison to those found within the SC Word Complexity Lexicon dataset. These complex words would, therefore, exhibit a significant difference between the two SC datasets. For this reason, the following hypothesis was drawn:

- **Hypothesis 1).** There shall exist several complex words that display positive keyness within the SC CompLex dataset when compared to their presence in the SC Word Complexity Lexicon dataset.

4.4 Examining Word Length

Zipf (1935) stated that "the magnitude of words tend, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of [their] occurrences". Since making this statement, it has become a common belief that "the most frequent words in any language tend to be shorter [and thus, more simple]" (Malmasi and Zampieri, 2016). Complex words, on the other hand, are believed to be less common and longer (Zampieri, Tan and Genabith, 2016).

Maddela and Xu (2018) together with Shardlow (2014) argue that this assumption "is not always accurate and is often the major source of errors within the [CWI] pipeline" (Maddela and Xu, 2018). For instance, the word length effect is a well-known effect, whereby the shorter the word is the more likely it is to be remembered and recalled (Guitard et al., 2018). However, recent research is putting the word length effect into question. Studies, such as Guitard et al. (2018) and Jalbert, Neath and Surprenant (2011a; 2011b), have provided evidence that greater word length may not always inhibit a word's ability to be recalled, and thus, long words may not always be considered complex; given that CWI uses complexity as a "synonym for difficultly" (Malmasi and Zampieri, 2016).

Regardless of the above debate, there still "exists numerous studies showing that on immediate tests, short words are recalled better than long words" (Guitard et al., 2018). This may be especially true for second-language English speakers (Guitard et al., 2018). It could, therefore, be theorised that the complex words taken from the SC Word Complexity Lexicon dataset would be on average shorter than the complex words taken for the SC CompLex dataset. This being based on the low inter-annotator agreement found within the Word Complexity Lexicon dataset which was implied by Maddela and Xu (2018) to be a result of their annotators coming from different countries as well as having English as their second and not their first language; even though they were defined as being "fluent English speakers" (Maddela and Xu, 2018). As such, being non-native English speakers, they were likely more prone to the word length effect than Shardlow, Cooper and Zampieri's (2020) annotators from the US, UK and Australia. They may have perceived longer words to be harder, hence more complex, than shorter words. With this in mind, a second hypothesis was made:

- **Hypothesis 2).** On average the SC Word Complexity Lexicon dataset shall contain longer complex words than the SC CompLex dataset.

4.5 Summary

This preliminary investigation aimed to see what effects differing annotation guidelines as well as annotators may or may not have had on the types of words identified as being complex within both the CompLex dataset and the Word Complexity Lexicon dataset. In this endeavour, this paper sought to determine whether any significant difference exists between the complex words assigned within both datasets. Two hypotheses were drawn so as to discover whether such a significant difference could be found (See Sections 4.3. and 4.4.).

4.6 Limitation

After having compiled both the SC CompLex and SC Word Complexity Lexicon datasets, it became apparent that due to their small size as well as their reliance on differing text genres, a direct comparison between the two would be less rewarding and somewhat inconclusive than previously thought. This being as there were not found to be many shared words that had been assigned as being complex within both datasets. Future researchers who are interested in comparing the effects of annotation guidelines and annotators should subsequently examine their impact on one dataset rather than two separate and differing datasets. Nevertheless, some complex words did appear to be shared. These words are discussed throughout the following sections in relation to this paper's hypotheses.

5 FINDINGS

5.1 Complex Words and Positive Keyness

All of the shared complex words displayed positive keyness. Those with the highest positive keyness were found to be more frequently assigned as being complex by Shardlow, Cooper and Zampieri's (2020) US, UK and Australian annotators than compared to Maddela and Xu's (2018) international expert annotators (See Tables 1 and 2). These complex words being 'substrate', 'assay' and 'derivatives'. However, 'secretariat' and 'signaling' were not. In fact, these assigned complex words, with their correspondingly low positive keyness ratings of +98 and +38, were assigned as being complex moreso among the Word Complexity Lexicon dataset annotators than among the CompLex annotators; if the measure of keyness was changed from chi-squared to log-likelihood, these low positive keyness scores may have even changed so as to display negative keyness (See Tables 1 and 2).

5.2 Word Length

The complex words within the CompLex dataset were found to have an average word length of 6.79 characters. This dataset was also discovered as having an average non-complex word length of 7.39 characters. The Word Complexity Lexicon dataset, on the other hand, was seen to have an average complex word length of 8.24 characters and an average non-complex word length of 6.35 characters.

5.3 Inter-Annotator Agreement

The complex or non-complex assignments made by Maddela and Xu's (2018) international expert annotators displayed in Table 2, would appear to exemplify their low rate of inter-annotator agreement of 0.64, or "variance in their judgements" (Maddela and Xu, 2018). For instance, in many cases the difference of one vote, or assignment, resulted in that particular classification. 'Substate', 'derivatives', and 'secretariat' have all been rated as being complex by 6 annotators. Another 5 annotators, however, have also rated these words as being non-complex. Furthermore, 'assay' and 'signaling' were classified as being complex and non-complex respectively, whereby a difference of 2 assignments for either of these words would have resulted in them being assigned the opposite label. This close judgement does not seem to have occurred among the CompLex annotators. For example, each word's assignment of complex or non-complex would appear to have a higher range, with a mean average range of 3. In contrast, those within the Word Complexity Lexicon dataset have a mean average range of 1.

Complex Word	Keyness
Substrate	+318
Assay	+288
Derivatives	+198
Secretariat	+98
Signaling	+38

Table 1: 5 Shared Complex Words and Keyness

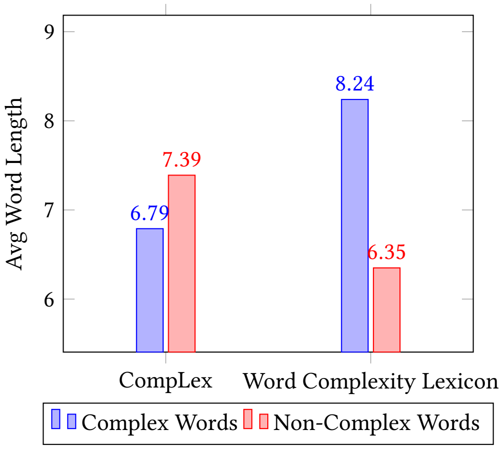

Figure 1: Average Word Lengths

	CompLex		Word Complexity Lexicon	
Word	Complex	Non-Complex	Complex	Non-Complex
Substrate	4	1	6	5
Assay	4	1	7	4
Derivatives	3	2	6	5
Secretariat	2	6	6	5
Signaling	1	4	4	7

Table 2: Assigned Labels

6 DISCUSSION

6.1 Datasets and Different Complex Words

Hypothesis 1). was found to be correct. Numerous shared complex words displayed positive keyness. Those with the highest positive keyness indicated that they were more commonly found to be complex within the CompLex dataset, whereas those with lower positive keyness were more frequently assigned as being complex with the Word Complexity Lexicon dataset.

Due to the limitation discussed in Section 4.6., this preliminary investigation was not able to theorise why such differences in complexity annotation occurred. Nevertheless, what is demonstrated in Tables 1 and 2 is that different annotation guidelines or annotators do, in fact, result in different complexity assignments. Therefore, this papers recommends that the accurate and consistent assignment of complexity should be taken as a top priority during the construction of a CWI dataset and classifier. This being as a dataset with inconsistent assignments and labels, is ultimately a dataset that may contain non-complex words being assigned as complex or vice-versa. This, in turn, defeats the purpose of a CWI classifier. For instance, by being trained on such a dataset, a CWI classifier may not be able to recognise unseen complex words given that it has been trained poorly and thus unable to distinguish between complex and non-complex word features.

6.2 Annotators and Word Length

The word length effect states that the shorter the word is the more likely it is to be remembered and recalled (Guitard et al., 2018).

As such, the word length effect implies that longer words may be thought as being more complex, whereas shorter words may be considered less complex. This was not found to be the case among the US, UK and Australian annotators of Shardlow, Cooper and Zampieri's (2020) CompLex dataset. For instance, it was found that the complex words assigned by these annotators were, on average, shorter than their assigned non-complex words by 0.6 characters. This being as their complex words were on average 6.79 characters long, whereas their non-complex words were on average 7.39 characters long. It may, therefore, appear that for CompLex's US, UK and Australian annotators word length was not an indicator of complexity.

The international expert annotators of the Word Complexity Lexicon dataset may have been influenced by the word length effect. Unlike CompLex's annotators, their assigned complex words were on average longer than their assigned non-complex words by 1.89 characters. This being as their complex words were on average 8.24 characters long. Their non-complex words, however, were on average 6.35 characters long. This may suggest that for these annotators, longer words were harder to recognise or recall and hence were perceived as being more complex.

Hypothesis 2)., was subsequently found to be correct. There are two possible explanations to why this may have been the case. In other words, explanations to why CompLex's annotators may not have been influenced by word-length whilst the Word Complexity Lexicon annotators likely were. On further investigation of the CompLex dataset, it was found that a large majority of its short complex words, these being under 3 characters long, were acronyms. It, thus, comes as no surprise that either due to annotation time constraints or lack of context, that CompLex's annotators marked these short words as being difficult to understand, hence complex. The international annotators of the Word Complexity Dataset Lexicon, on the other hand, may have found longer words harder to understand. It has been well documented that second-language users are more commonly influenced by the word length effect (Guitard et al., 2018). Therefore, as English was not their first language, they may have found shorter words easier to understand and thus found them to be less complex. Nevertheless, Maddela and Xu (2018) claimed that these annotators were fluent in English. However, without an exact measure of their English language proficiency, their fluency may be questionable.

6.3 Low Inter-Annotator Agreement

Table 2 illustrates the low inter-annotator agreement among Maddela and Xu's (2018) international expert annotators (See Section 5.3.). According to Maddela and Xu (2018), a probable cause of this disagreement is that they had "different native languages [and were from different countries]" (Maddela and Xu, 2018). For isntacne, it is well known that second-language English speakers with differing first-languages may find differing English vocabulary more or less difficult. Crowther, et al. (2015), for example, found that Chinese Mandarin and Hindi speakers found differing words more or less hard to comprehend depending on differing variables related to their first language. For this reason, the native language of an annotator should be continued to be taken into consideration when constructing a CWI dataset.

7 CONCLUSION

This preliminary investigation has sought to provide insight into what effects annotation guidelines and annotators have on the quality of a CWI dataset. It was found that word complexities were assigned differently within the CompLex dataset compared to the Word Complexity Lexicon dataset. It was also discovered that Shardlow, Cooper and Zampieri's (2020) US, UK and Australian annotators assigned shorter words as being complex more often than Maddela and Xu's (2018) international expert annotators. This may have been due to the CompLex dataset containing a high number of acronyms or Maddela and Xu's (2018) annotators having English as their second-language. Furthermore, the Word Complexity Lexicon dataset's low inter-annotator agreement was also confirmed. This was likely a result of the differing linguistic backgrounds of each annotator. Based on these findings, several future recommendations have been provided so as to improve the future quality of CWI datasets.

8 FUTURE RECOMMENDATIONS

Due to Maddela and Xu's (2018) use of pairwise or 6 rather than 5-point likert scale annotation for the annotation of a CWI dataset, together with its corresponding improvement of two state-of-the-arts systems: SV000gg and the nearest centroid classifier (Paetzold and Specia, 2016), this papers recommends that future CWI research should further investigate the use of pairwise annotation, hence the use of a 6-point likert scale for data annotation. In addition, future CWI researchers who seek to investigate guideline and annotator effects on dataset quality, should do so by comparing the effects of these variables on one dataset rather than two. Lastly, the use of second-language English speakers for CWI should entail detailed English language proficiency scores. These scores would, in turn, dismiss any critique surrounding wrong complexity assignment due to low English proficiency; unless, the CWI classifier is purely designed to identify words that are complex for English language learners.

REFERENCES

Crowther, D., Trofimovich, P., Saito, K. and Isaacs, T. (2015). Second Language Comprehensibility Revisited: Investigating the Effects of Learner Background. TESOL Quarterly, 49 (4), pp. 814-837.

Evison (2014) 'What are the basics of analysing a corpus?' In O'Keeffe, A. McCarthy, M. (eds.) The Routledge Handbook of Corpus Linguistics. New York: Routledge.

Guitard, D., Saint-Aubin, J., Gabel, A.J., Suprenant, A.M. and Neath, I. (2018). Word Length, Set Size, and Lexical Factors: Re-Examining What Causes the Word Length Effect. Journal of Experimental Psychology: Learning, Memory, and Cognition, 44 (11), pp. 1824-1844.

Ide, N. and Putsejovsky, J. (2017). Handbook of Linguistic Annotation. New York: Springer.

Jalbert, A., Neath, I. And Surprenant, A.M. (2011b). Does length or neighbourhood size cause the word length effect?. Memory and Cognition, 39 (1), pp. 1198-1210.

Jalbert, A., Neath, I., Bireta, T.J., and Surprenant, A.M. (2011a). When does length cause the word length effect?. Journal of Experimental Psychology: Learning, Memory and Cognition, 37 (1), pp. 338-353.

Kajiwara, T., Matsumoto, H. and Yamamoto, K. (2013). "Selecting Proper Lexical Paraphrase for Children." In Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING). Kaohsiung, Taiwan, The Association for Computational Linguistics and Chinese Language Processing. pp. 59-73.

Maddela, M. and Xu, W. (2018). "A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplifcaition." In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 31 October – 4 November. Association for Computational Linguistics. pp. 3749-3760.

Malmasi, S. and Zampieri, M. (2016). "MAZA at SemEval-2016 Task 11: Detecting Lexical Complexity Using a Decision Stump Meta-Classifier." In Proceedings of SemEval-2016. San Diego, California, 16-17 June. Association for Computational Linguistics. pp. 991-995.

Malmasi, S., Dras, M. and Zampieri, M. (2016). "LTG at SemEval-2016 Task 11: Complex Word Identification with Classifier Ensembles." In Proceedings of SemEval-2016. San Diego, California, 16-17 June. Association for Computational Linguistics. pp. 996-1000.

Paetzold, G. H. and Specia, L. (2016). "SemEval 2016 Task 11: Complex Word Identification." In Proceedings of SemEval-2016. San Diego, California, 16-17 June. Association for Computational Linguistics. pp. 560-569.

Rello, L., Baeza-Yates, R., Dempere-Marco, L. and Saggion, H. (2013) "Frequent words improve readability and short words improve understandability for people with dyslexia." In Proceedings of INTER-ACT. Paphos, Cyprus, 2-6 September. IFIP. pp. 203-219.

Shardlow, M., Cooper, M. and Zampieri, M. (2020). "CompLex – A New Corpus for Lexical Complexity Prediction from Likert Scale Date." In Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding Difficulties (READI). Masrseille, France, 11 May 2020. READI. pp. 1-6.

Shardlow, M. (2014). A Survey of Automated Text Simplification. International Journal of Advanced Computer Science and Applications, 4 (1), pp. 58-70.

Stanford (2019) 'The Chi-Square Test'. Available at https://web.stanford.edu/class (Accessed: 5 th of October 2020).

Yiman, S. M., Stajner, S., Riedl, M. and Biemann, C. (2017). "CWIG3G2 – Complex Word Identification Task across Three Text Genres and Two User Groups." In Proceedings of The 8th International Joint Conference on Natural Language Processing. Taipei, Taiwan, 27 November – 1 December. AFNLP. pp. 401-407.

Zampiechildrenri, M., Tan. L. and Genabith J.V. (2016). "MacSaar at SemEval-2016 Task 11: Zipfian and Character Features for Complex Word Identification." In Proceedings of SemEval-2016. San Diego, California, 16-17 June. Association for Computational Linguistics. pp. 1001-1005.