Proceedings of the 2nd READI Workshop @ LREC2022, pages 9–16
Marseille, 24 June 2022

Agree to Disagree: Exploring Subjectivity in Lexical Complexity

Matthew Shardlow
Manchester Metropolitan University
M.Shardlow@mmu.ac.uk

Abstract

Subjective factors affect our familiarity with different words. Our education, mother tongue, dialect or social group all contribute to the words we know and understand. When asking people to mark words they understand some words are unanimously agreed to be complex, whereas other annotators universally disagree on the complexity of other words. In this work, we seek to expose this phenomenon and investigate the factors affecting whether a word is likely to be subjective, or not. We investigate two recent word complexity datasets from shared tasks. We demonstrate that subjectivity is present and describable in both datasets. Further we show results of modelling and predicting the subjectivity of the complexity annotations in the most recent dataset, attaining an F1-score of 0.714.

Keywords: Complex Word Identification, Lexical Complexity Prediction, Text Simplification

1. Introduction

Lexical Complexity Prediction (LCP) has applications in Text Simplification (Zampieri et al., 2017), as well as Readability Assessment (Ehara, 2020). It is the task of identifying how complex a word is likely to be for an end user. Similarly, Complex Word Identification (CWI) is the task of identifying whether a word is complex or not. In both these tasks, disagreements naturally arise between annotators seeking to faithfully give their subjective opinions on the difficulty of the words in question. Take, for example the following sentence, taken from the CWI2018 shared task data (Yimam et al., 2017):

> "A man and a woman questioned on suspicion of assisting an **offender** have been released."

The marked token (*offender*) may be considered complex by some and simple by others. In fact this example split the pool of annotators, being marked complex by 50% of the annotators and simple by the rest. This is not always the case though, and there are also words that are consistently annotated. For example, in the LCP2021 data (Shardlow et al., 2022), the following example is given:

> "Similarly, changes in **synaptic plasticity** due to Ca2+-permeable AMPARs [51,52,60], e.g., in piriform cortex, might alter odor memorization processes."

Clearly, here the entire context is very hard to understand, and the term in that context (*synaptic plasticity*) is inaccessible to a non-domain expert. As such, the term was annotated as the highest level of difficulty by all but one annotator.

Similarly, in the following context, also taken from LCP2021 all annotators chose the easiest level of difficulty for the token *hand*:

> "But he, beckoning to them with his **hand** to be silent, declared to them how the Lord had brought him out of the prison."

We can draw from these few examples that there are clear cases where annotators agree, and clear cases where annotators do not agree. These exist across multiple datasets and are not merely a factor of the token's complexity (i.e., we may naïvely assume that everyone agrees on simple words, but differs on complex words, or vice versa). For sake of ease, we will refer to *subjectivity* in the remainder of this paper in the context of the subjectivity of complexity.

These initial insights allow us to form the following research hypotheses and questions:

RQ1: Can we distinguish words with subjective or consistent complexity? Are they the same across different datasets?

- **RH1.1:** We can identify from existing datasets clear patterns of subjective and non-subjective complexity annotations.
- **RH1.2** The subjective and non-subjective complex words will be the same across datasets.

RQ2: What factors model subjectivity?

- **RH2.1:** Lexical ambiguity will correlate to subjectivity.
- **RH2.2:** Lexical frequency will correlate to subjectivity.
- **RH2.3:** Psycholinguistic norms will correlate to subjectivity.

RQ3: Can we reliably predict which words are likely to be consistently annotated as complex or simple, and which words are likely to be subjectively complex?

- **RH3.1:** Classical machine learning classifiers can predict subjectivity based on the lexical factors identified.

To answer these questions, the remainder of the paper is structured as follows: We define the notions of complexity and subjectivity in Section 2 and explore this in a concrete manner in Sections 3 and 4, which cover datasets from two shared tasks. We also discuss the internal mechanisms that were used during annotation and demonstrate the subjectivity that is present, which addresses RH1.1. Section 5 compares the two datasets in terms of the words that are found to be consistent or subjective and addresses RH1.2 accordingly. Section 6 identifies a number of pertinent features taken from the CWI/LCP literature and uses statistical methods to determine their relation to the subjectivity, addressing RH2.1–3. We build various classifiers to predict subjectivity in Section 7, which allows us to answer RH3.1. The paper concludes with a discussion of the work (Section 8) and a short discussion of the limited related works that exist (Section 9).

2. Definitions

We make an initial definition of the notion of subjectivity as follows. We build on this definition in the context of two datasets in Sections 3 and 4.

> The complexity of a word is considered *subjective* if the returned complexity labels for that word span a range of complexity values.

More formally, we can define a complexity annotation scheme as taking vocabulary items v_i from some vocabulary V and presenting them to a discrete set of n human annotators $h_1, ..., h_n$, drawn from a pool H of size at least n who each return some label l drawn from a discrete ordinal integer label set L. An annotation a_i can be defined as a point in the relation $A = H \times L$ and each v_i receives n annotations which can be represented as a vector $\overrightarrow{a}$ (with indices $a_1 ... a_n$). Given these conditions, we can define 2 properties, complexity and subjectivity as follows. The complexity of a vocabulary item v_i is the mean of the ordinal values of the labels in the annotations:

$$Complexity(v_i) = \frac{\sum_{j=1}^{N} a_j}{n} \tag{1}$$

Similarly, we can use these definitions to define a formal measure of subjectivity modelled on the average absolute deviation of $\overrightarrow{a}$:

$$Subjectivity(v_i) = \frac{\sum_{j=1}^{n} |Complexity(v_i) - a_j|}{n} \tag{2}$$

We may also define thresholds for complexity T_c and subjectivity T_s by which we define a vocabulary item as holding the property of complex or subjective:

$$Complex(v_i) \rightarrow Complexity(v_i) > T_c \tag{3}$$

$$Subjective(v_i) \rightarrow Subjectivity(v_i) > T_s \tag{4}$$

T_c may be sensibly set at 0.5 in complexity research, although this could be varied depending on the requirements of an application. T_s will be some function of the magnitude of L (i.e., the more categories to choose from, the wider deviation is acceptable before crossing the subjectivity threshold) and also of N (i.e., the more annotators that we have, the more potential for subjectivity). We propose the following definition for determining a subjectivity threshold as follows:

$$T_s = \alpha \times |L| \times n \tag{5}$$

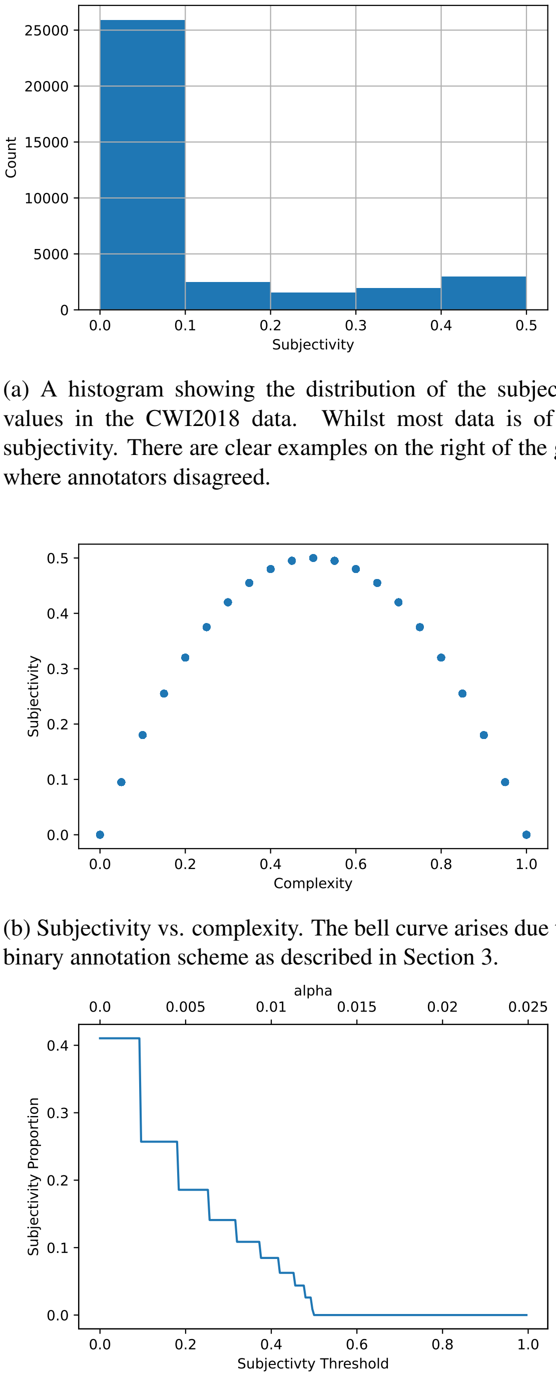
where α is a normalising constant set to some small value between 0 and 1. We report on empirical values of α in the next two sections.

3. Subjectivity in CWI2018 Annotations

The CWI2018 data covered Wiki text and Newswire data. Annotators were asked to identify any word or span that they found to be complex in a context. Each context was presented to 20 annotators, of which 10 were native speakers of English and 10 were not. This resulted in 20 binary annotations for each identified term which indicated whether an annotator considered that term complex. These binary annotations were represented by the ordinal labels 0 and 1 such that if every annotator agreed a word was complex it would have 20 positive annotations and get a score of 1. If no annotator considered a word complex it would have 20 zeroes and be given an overall score of 0.

Interestingly from the point of view of subjectivity, annotator disagreement is directly modelled in the complexity labels. As in the initial example given in the introduction, if 10 annotators found a word to be complex, whereas 10 found it to be simple, the word would be given a score of 0.5, according to the formulae given in Section 2.

We investigate the nature of subjectivity in the English portion of CWI2018 data through the 3 plots in Figure 1. We firstly show in Figure 1a the distribution of subjectivity values in the CWI2018 dataset. These values were calculated using the formula for subjectivity given above. It is clear that most items in the dataset fall in the lower end of the subjectivity — coming in the 0.0–0.1 bin. These represent both complex and simple words, although the majority are simple words due to the nature of the dataset. There are a number of words in the subsequent bands, with the highest bin (0.4–0.5) having just under 3000 examples. Figure 1b shows the relationship between subjectivity and complexity in the binary annotation setting of CWI2018. The bell curve that arises represents the fact that the lowest-subjective elements are those with high or low complexity (everyone agreed either way), whereas the most subjective elements are those with a mid-level complexity (half the annotators said simple, the other half said complex). Finally, Figure 1c shows the effect of varying alpha (And hence the threshold) on the proportion of words



(a) A histogram showing the distribution of the subjectivity values in the CWI2018 data. Whilst most data is of low-subjectivity. There are clear examples on the right of the graph where annotators disagreed.

(b) Subjectivity vs. complexity. The bell curve arises due to the binary annotation scheme as described in Section 3.

(c) The result of varying the subjectivity threshold according to α. Around 40% of the instances are considered subjective at low values of α.

Figure 1: Analysis of the subjectivity values in the CWI2018 dataset annotations.

that are considered subjective. A subjectivity threshold above 0.5 ($\alpha = 0.0125$) leads to no words being considered subjective. Figure 1c demonstrates that the subjectivity threshold can be empirically set to determine the words that are determined as subjective. A subjectivity threshold of 0.4 ($\alpha = 0.01$) would result in 5% of instances being considered subjective, whereas a lower threshold of 0.2 ($\alpha = 0.005$) would result in 15% of instances considered subjective. A few examples of words across subjectivity values are described in Table 1.

Subj	**Terms**
0.0	back, bomb, censorship, death, instilled
0.25	assets, cushion, launches, previously
0.5	approaching, credence, overspending, slash

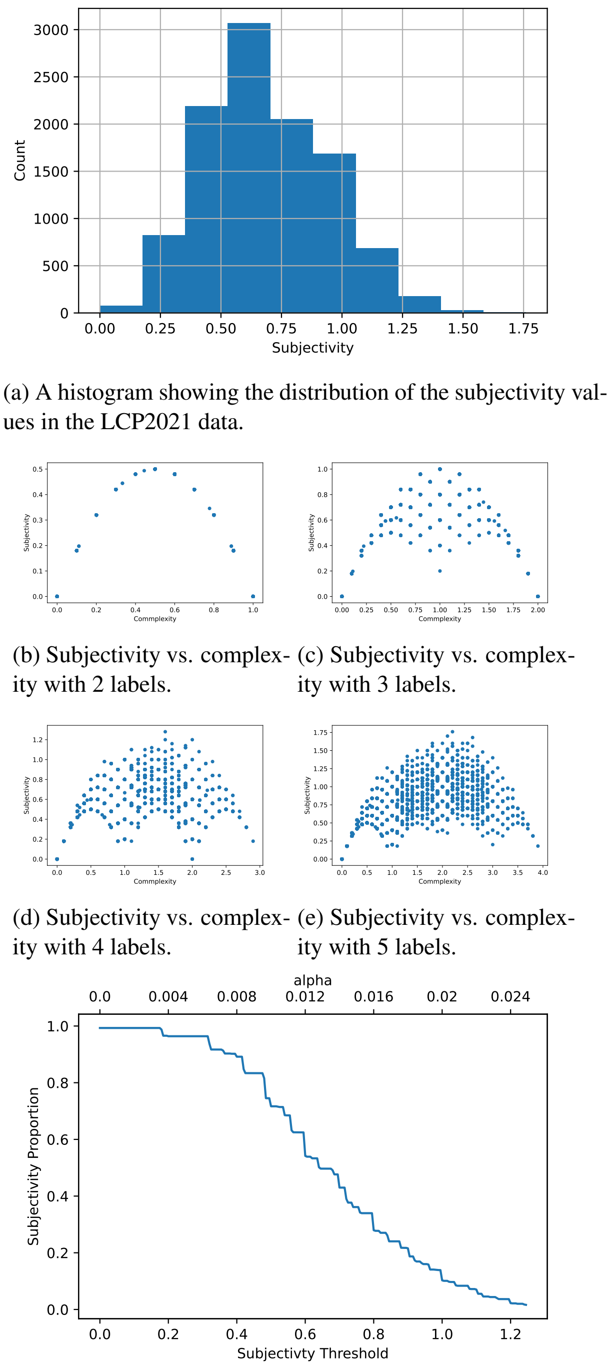
Table 1: Terms by subjectivity for CWI2018

We can see from Table 1 that both simple (back, town) and complex (censorship, instilled) terms were agreed upon by all annotators. The most controversial words are typically longer words that may require some subjective or domain knowledge to fully understand.

4. Subjectivity in LCP2021 Annotations

Whereas the CWI2018 data used binary annotation ($L = \{0, 1\}$) the LCP2021 task used a 5-point Likert scale ($L = \{0, 1, 2, 3, 4\}$). This allows annotators to agree on points in the Likert scale that do not represent the poles of the scale. For example, annotators may all agree that an instance is of medium complexity with a subjectivity of 0. Equally, annotators may nearly agree, centering around a given point, but disagreeing (within varying margins) from that point. Finally, it is possible that an instance might polarise the annotator pool. For example, if an instance is ambiguous one set of annotators may interpret in one way, whereas another take another interpretation. The first interpretation might lead to annotations of simplicity, whereas the latter leads to annotations of difficulty — creating a multi-modal distribution in the returned annotations. This has some negative ramifications for the definition of complexity used in this work, as the mean implicitly assumes a normal distribution. The complexity is still reflective of a central point in the annotations, but not a maximal point in this scenario. However, for our definition of subjectivity, the case of multi-modal distributions will still lead to high subjectivity values as the multiple modes will be separated from the centralised complexity value. In any case, this may become more of an issue with continuous annotations, as opposed to a 5-point Likert scale, where the few points in the scale force annotator decisions around common poles.

To investigate the phenomenon of subjectivity in the LCP2021 data, we applied the same transform following the equations from Section 2 to the original annotations to give a complexity and subjectivity value for each instance. The number of annotations for each instance in the LCP2021 data is 10. We demonstrate the subjectivity of these annotations by creating the same figures as for the CWI2018 data, as shown in Figure 2. Figure 2a demonstrates the distribution of subjectivity values in our dataset, with a mean around 0.6 and subjectivity ranging from 0 to 1.5. (N.b., subjectivity is



(a) A histogram showing the distribution of the subjectivity values in the LCP2021 data.

(b) Subjectivity vs. complexity with 2 labels.

(c) Subjectivity vs. complexity with 3 labels.

(d) Subjectivity vs. complexity with 4 labels.

(e) Subjectivity vs. complexity with 5 labels.

(f) The result of varying the subjectivity threshold according to α.

Figure 2: Analysis of the subjectivity values in the LCP2018 dataset annotations.

not capped at 1 as it is a function of the ordinal labels which range from 0-4). The distribution appears to be Gaussian, with a left skew. The range of the values can be interpreted in the context of the number of labels available. A subjectivity of 0.6 means that the annotations were within 0.6 points of a label of each other. The maximum possible subjectivity in a 5-label annotation setting would be where an equal number of annotators have selected polarised values. (i.e., 0 and 4). In this case, the complexity of the annotations for any N would be 2, as would the subjectivity. So a subjectivity of 1.0 in this setting is half of the theoretical maximum possible subjectivity. Almost all of the annotations fall below this mark.

To further investigate the effect that the number of annotators has on the distribution of the annotation with respect to complexity, we first recast the 5-point scale as a binary annotation. We further recast the problem as a 3 and 4 point annotation problem by relabelling in the manner described in Table 2, where the top row describes the original annotation point and the subsequent describe the transformed point. [1] Applying the transform allowed us to produce the graphs in Figures 2b–2e.

Original	0	1	2	3	4
2-label	0	0	1	1	1
3-label	0	0	1	2	2
4-label	0	1	2	2	3

Table 2: Label transforms used.

To describe the boundaries of the graphs in Figures 2b–2e, we can consider that the y-axis is determined by the complexity and the x-axis is determined by the subjecctivity. In the simplest case (Figure 2b) a parabola is formed as the subjectivity requires the complexity to be calculated, summing the number of instances once to calculate the mean and then again to calculate the subjectivity (hence x^2). It is logical to consider that when we have only labels 0 and 1, the subjectivity will be 0 when the complexity is 0 and the subjectivity will be 1 when the complexity is 1 (as these cases can both only arise when the vector is all zeroes, or all ones). Similarly, when the complexity is 0.5, the subjectivity is also 0.5 as this arises when the annotator pool is perfectly polarised (i.e., half have chosen zero and half have chosen one).

Let us then consider the more complex case of Figure 2c. In this graph there are 3 labels available to the annotators. We can see that the annotations fall in a space that can be described by three boundaries. The upper boundary is described as above, the case where the annotation vector contains only instances of 0 or 2 (2 being the largest possible annotation). It is the same curve as in Figure 2b, but is twice as high and twice as wide. There are also two clear lower bounds in the graph. The first, between 0 and 1 on the x-axis is described by the curve in Figure 2b as it is the case of annotations which contain only zeroes and ones (i.e., no twos). The second, falling between one and two on the x-axis is

[1] The relabelling is done in arbitrary manner to enable us to investigate the same effect in the 2, 3 and 4 point setting. It is not intended as a robust means of reducing the number of labels in an annotation setting.

described by a new curve, which is the same shape as the other two, but similarly described by the annotation vectors containing only 1's and 2's.

Given the description of Figures 2b and 2c above, it should be clear what is happening in the more complex Figures 2d and 2e. In these, the upper bound is similarly described by the polarised case between the first and last labels, whereas the lower bounds are described by the polarised cases between successive labels. This gives rise to the effect that subjectivity minima appear at each ordinal label (i.e., when all annotators selected that label) and that a single maxima appears at complexity = 0.5, when half the annotators selected the lowest possible annotation and the other half selected the highest.

Considering Figure 2e, which represents the original labels in the LCP2021 data, we see that the spread of annotations covers almost the entire possible space. We can observe that lower subjectivity occurs at the two ends of the scale (0.0 and 4.0), with similarly lower values for subjectivity appearing at 1.0 and 3.0. Interestingly, where there should be a minima at 2.0, this is missing, indicating that annotators were unlikely to agree on the 'Neutral' category in the annotation scheme. The observed maxima is around 1.75, indicating that the top portion of potential subjectivity values is missing as the maximum possible subjectivity would be 2.0.

We also analysed the threshold for subjectivity prediction and report our results in Figure 2f. This follows an inverse-S curve, in line with the normal distribution of subjectivity shown in Figure 2a. Again, we are not seeking to give a specific value for the subjectivity threshold here, but rather attempting to expose the behaviour of the thresholded values. We can see that a threshold of 0.25 ($\alpha = 0.005$) will result in around 95% of terms being considered subjective, whereas a threshold of 0.5 ($\alpha = 0.01$) will result in around 60% of the terms being considered subjective.

5. CWI2018 vs. LCP2021

Using the data above we can draw several comparisons between the two prominent existing datasets for CWI/LCP annotation. First of all it is clear from Figures 1a and 2a that the underlying distribution of subjectivity in CWI2018 and LCP2021 is fundamentally different. This is due to the existence of many more agreed upon simple terms in CWI2018. By comparison, the LCP2021 data contains much more subjectivity than the CWI2018 data. Whereas the majority of instances in the latter dataset have a subjectivity close to 0, the subjectivity in the LCP data is centered around 0.4-0.6 (i.e., around half a point on the Likert scale). This is a factor of the way in which each dataset was annotated. In the CWI2018 data, annotators were presented with a context and asked to identify any complex terms. If a term was identified by at least one annotator, it was included in the dataset. This leads to the case where many terms were annotated by only a single annotator, having an annotation vector with a single 1 and the rest 0's. In our definition of complexity/subjectivity this is labelled as low-complexity, low-subjectivity. But it may be the case that the non-annotations of the term are really just the other annotators neglecting to annotate that term, rather than a confirmation of the term's simplicity. Contrastingly, the LCP2021 data presented annotators with specific terms and requested an annotation decision for every given term. This means that every annotation in the dataset is representative of a meaningful decision by the annotator. Clearly, this has led to more subjectivity in the range of annotations that are returned for LCP2021 than CWI2018.

The range of subjectivity with respect to complexity values is also larger in the LCP2021 data as a result of the labels on a 5-point Likert scale that were employed. This can be seen when comparing Figure 1b to Figure 2e. Whereas for the CWI2018 data, the subjectivity values are linked directly to the complexity, the LCP2021 data has a range of subjectivity values for each complexity value. This is because each possible complexity value could be made up of many different annotation vectors. E.g., a complexity value of 2 could be made up of 10 annotations of 2 or 5 annotation of 1 and 5 annotations of 3, as well as many other ways. Whereas the former would have a low subjectivity value, the latter would have a higher subjectivity as the annotators agreed less.

The subjectivity threshold behaves in a similar way between the two datasets. Both produce an inverse S-curve in Figures 1c and 2f. The α value was used to determine common thresholds and across our 2 datasets it allows for a similar threshold to be set given different values of α. Further work on datasets with different values of n and L is needed to determine the robustness of α to these values. Both curves follow a stepped curve, due to the different values that could be produced by the formula for subjectivity operating on a fixed size vector of integers. The LCP2021 data has more levels, producing a smoother curve as it has more labels in the annotation scheme — allowing for a wider range of final values.

We further compared the subjectivity values for common words between the CWI2018 and LCP2021 datasets. To do this, we took the subset of instances containing tokens that occurred in both datasets ($n = 26166$) and calculated Pearson's correlation between the subjectivity values in both datasets. The correlation was low at 0.189, indicating that the subjectivity for specific words in the two datasets is not well-aligned. This may seem surprising, as we would expect subjective words in one dataset to also be subjective in another dataset, however given the findings presented so far on the nature of subjectivity in each dataset and the description of the differing annotation protocols employed, it is conceivable that the discrepancy is in fact due to the differences in the datasets' construction and that future datasets following either protocol would have higher correlation.

6. Factors Affecting Subjectivity

To investigate our second research question, we adopt the LCP2021 data and perform a correlation analysis with a number of features which are used elsewhere in the literature to determine the complexity of a word. The feature categories and specific features, with identifiers are listed below:

Lexical Ambiguity:

- **Number of WordNet Senses (LA1):** The number of synsets that the wordform appears in within WordNet.
- **WordNet Tree Depth (LA2):** The depth at which this word appears in the WordNet Tree.
- **Number of WordNet Hyponyms (LA3):** The number of hyponyms (words with a more specific meaning) that this word has in WordNet.

Lexical Frequency:

- **Web1T Frequency (LF1):** The frequency of the term in the Google Web1T unigram dataset (Brants and Franz, 2006).
- **Subtlex Frequency (LF2):** The frequency of the term in the Subtlex dataset (Van Heuven et al., 2014).
- **log Web1T Frequency (LF3):** $log(\mathbf{LF1})$
- **log Subtlex Frequency (LF4):** $log(\mathbf{LF2})$

MRC Psycholinguistic Norms:

- **Familiarity (PN1):** How likely the word is to be known.
- **Concreteness (PN2):** The degree to which the word represents a grounded concept.
- **Imageability (PN3):** The degree to which the referent of a term can be visualised.

We use Pearson's correlation to determine the relationship between the subjectivity values for LCP2021 and the features we have determined above. These are presented in Table 3, where we also include the correlation with complexity for reference. The correlation between the complexity and subjectivity values was 0.641.

Table 3 shows that the features we tried have a weak negative correlation with subjectivity. The correlation with the lowest magnitude (LA2, WordNet Tree Depth) is -0.094 and the highest (LF4, Log SUBTLEX Frequency) is -0.412. The correlation values for subjectivity are typically in line with, although slightly lower than those for complexity, except in the case of LA2 and LF3, which both show a larger discrepancy, although the reason for this is unclear.

7. Predicting Subjectivity

Finally, we train several models to predict subjectivity in the LCP2021 dataset. This could enable future applications to not only determine which words are complex or simple, but also determine whether a word is likely to split the opinions of users. This may be useful for determining simplification and personalisation strategies, or for better understanding the nature of a complexity value that is returned by a system. For example, if a system returns a neutral complexity, it is helpful to know if that value is likely to be agreed upon, or if some users will find the word difficult, whereas others will find it easy (giving an average of neutral).

We first select a subjectivity threshold of 0.68, which splits the data into 50% subjective and 50% non-subjective. Duplicate tokens in the dataset were removed, leaving 5,617 instances. We did not take contexts into account, as our features are context-free. We then created a training (70%) and testing (30%) set for our experiments.

We selected a Support Vector Machine (SVM), Random Forest (RF) and AdaBoost (AB) classifier from SciKitLearn and trained each one on our dataset. We did not tune the hyperparameters. We used all features described previously in Section 6. All results are reported on a single final run on the test set. We report the Precision, Recall and F1 score for both the Subjective and Non-Subjective classes in Table 4.

Our results are intended to demonstrate that subjectivity can be predicted using the features we have identified, as well as to give some simple baseline results for performance on this task. The scores indicate a reasonable predictive power, with AdaBoost giving an F1 score of 0.713 on the subjective class and 0.645 on the non-subjective class.

8. Discussion

8.1. Answers to Research Hypotheses

The answers to our initial research hypotheses stated earlier are given below:

RH1.1: We demonstrated that we could identify subjective and non-subjective annotations through the use of an equation for determining a subjectivity value and setting a threshold. We investigated the nature of subjectivity in the CWI2018 and LCP2021 datasets and demonstrated that both datasets contain a range of subjective and non-subjective annotations.

RH1.2: We found a low correlation between the subjectivity values for common terms in the two datasets we studied. Our analysis showed that the nature of subjectivity in these datasets is different, leading to the discrepancy.

RH2.1–2.3: We demonstrated that all of our feature categories had a low, but meaningful correlation with subjectivity. The features that we selected are also correlative with complexity and, as subjectivity and complexity are correlative with each other, we were able to use these features for subjectivity too.

	LA1	**LA2**	**LA3**	**LF1**	**LF2**	**LF3**	**LF4**	**PN1**	**PN2**	**PN3**
Complexity	-0.387	-0.229	-0.197	-0.330	-0.246	-0.443	-0.573	-0.351	-0.331	-0.314
Subjectivity	-0.265	-0.094	-0.167	-0.283	-0.222	-0.271	-0.412	-0.274	-0.258	-0.245

Table 3: Correlation analysis between common lexical features and complexity/subjectivity in the LCP2021 dataset

Method	Subjective			Non-Subjective		
	P	R	F1	P	R	F1
RF	0.658	0.681	0.669	0.656	**0.632**	0.644
SVM	0.623	**0.836**	**0.714**	**0.736**	0.476	0.579
AB	**0.660**	0.775	0.713	0.715	0.586	**0.645**

Table 4: Results of predicting which instances in the dataset will be subjective

RH3.1: We were able to predict the subjective label of the words in the LCP2021 dataset with an F1 score of 0.71. This demonstrates that subjectivity is a predictable phenomenon and we hope that in light of this finding future researchers will consider complexity in light of subjectivity.

8.2. Threats to Validity

One deficiency in our work is that we have not taken context into account. In the LCP2021 and CWI2018 annotations words were presented in context and the labels were given for the word in context, not for the word itself. This meant that repeated instances of a word had different annotation vectors and hence complexity labels (as different word senses, etc. affected the complexity). In our work, we have selected a single instance of each token, reducing the dataset size and ignoring the context. We expect to be able to address this in future work by investigating the context sensitivity of lexical subjectivity, in relation to complexity as well as other tasks.

The definition of complexity was formalised for this paper. Whilst this is reflective of the processes undertaken in previous papers to the best of the authors knowledge and given the reporting in previous work, it is possible that some unreported factors of the process are missing from our definitions. The measure of subjectivity was also determined within the scope of this work and is not adopted widely by the community. We hope that this work will introduce the notion of subjectivity and allow researchers working on lexical complexity to consider their annotations in the context of subjectivity.

Finally, a threat to the validity is that the work is done on secondary datasets. In the scope of this work, we have no control over the quality of the annotations that have been undertaken. Each dataset is reported on extensively in its own paper which detail the quality control mechanisms used to ensure that the annotators were doing the task expected of them.

9. Related Work

Complex Word Identification was first proposed as an initial step in the lexical simplification pipeline (Devlin, 1998). Efforts to automatically predict complex words (Shardlow, 2013) using machine learning techniques showed this to be possible. The task was popularised by shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018), where winning systems typically used feature based approaches (Gooding and Kochmar, 2018). Recently, the LCP2021 shared task (Shardlow et al., 2021) introduced continuous complexity prediction, as opposed to the binary or probabilistic prediction seen prior. High-ranking systems used either transformer based models (Yaseen et al., 2021) or feature engineering approaches (Mosquera, 2021).

Further work in CWI/LCP has sought to adapt the problem to a personalising task (Lee and Yeung, 2018) in which the specific needs of a user are modelled and reflected in individualised complexity predictions. Recent work demonstrated that lexical complexity differs due to annotator background, such as native speakers vs. non-native speakers (Gooding et al., 2021).

The distribution of lexical complexity shown in this work is backed up by previous works from the literature analysing the CWI-2018 dataset (Quijada and Medero, 2016). This data, and by association the concept of lexical complexity, has been considered subjective previously by other authors (Finnimore et al., 2019).

In the field of sentiment analysis the term subjectivity is used to refer to the degree to which a user is drawing on their own personal opinion vs. stating objective fact (Maks and Vossen, 2012; Hill and Korhonen, 2014). That is a subtly different notion of subjectivity to the one used here. In the context of Lexical Complexity Prediction, we are assuming that a user's annotations are inherently drawn from personal experience, and instead our measure is whether those personal experiences converge or diverge.

10. Conclusion

We have investigated the nature of Lexical subjectivity within the scope of lexical complexity. We show that this exists across two prominent datasets and outline how it differs between them. We have also shown that subjectivity is not a stochastic phenomenon, but is correlated to several well-known features for lexical complexity and that we are able to predict subjectivity with simple machine learning classifiers in an unseen setting. We expect that transformer based methodologies will also provide strong scores on this task, and leave these experiments to future work. We release the datasets with subjectivity values, and the code used to create them via GitHub[2].

11. Bibliographical References

Devlin, S. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Ehara, Y. (2020). Interpreting neural CWI classifiers' weights as vocabulary size. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 171–176, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

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

Gooding, S. and Kochmar, E. (2018). Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Gooding, S., Kochmar, E., Yimam, S. M., and Biemann, C. (2021). Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449.

Hill, F. and Korhonen, A. (2014). Concreteness and subjectivity as dimensions of lexical meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–731, Baltimore, Maryland, June. Association for Computational Linguistics.

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

Maks, I. and Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688.

Mosquera, A. (2021). Alejandro mosquera at semeval-2021 task 1: Exploring sentence and word features for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Quijada, M. and Medero, J. (2016). HMC at SemEval-2016 task 11: Identifying complex words using depth-limited decision trees. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1034–1037, San Diego, California, June. Association for Computational Linguistics.

Shardlow, M., Evans, R., Paetzold, G., and Zampieri, M. (2021). Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16.

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*, pages 103–109.

Yaseen, T. B., Ismail, Q., Al-Omari, S., Al-Sobh, E., and Abdullah, M. (2021). Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78.

Zampieri, M., Malmasi, S., Paetzold, G., and Specia, L. (2017). Complex word identification: Challenges in data annotation and system performance. *NLPTEA 2017*, page 59.

12. Language Resource References

Brants, T. and Franz, A. (2006). Web 1t 5-gram version 1 (2006). *Linguistic Data Consortium, Philadelphia*.

Shardlow, M., Evans, R., and Zampieri, M. (2022). Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, pages 1–42.

Van Heuven, W. J., Mandera, P., Keuleers, E., and Brysbaert, M. (2014). Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

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

[2] https://github.com/MMU-TDMLab/LCP_Subjectivity