Assessing the Usefulness of Different Feature Sets for Predicting the Comprehension Difficulty of Text

Brian Mac Namee[1], John D. Kelleher[2], and Noel Fitzpatrick[2]

[1] School of Computer Science, University College Dublin, Ireland
[2] Dublin Institute Of Technology, Ireland

Abstract. Within English second language acquisition there is an enthusiasm for using *authentic* text as learning materials in classroom and online settings. This enthusiasm, however, is tempered by the difficulty in finding authentic texts at suitable levels of comprehension difficulty for specific groups of learners. An automated way to rate the comprehension difficulty of a text would make finding suitable texts a much more manageable task. While readability metrics have been in use for over 50 years now they only capture a small amount of what constitutes comprehension difficulty. In this paper we examine other features of texts that are related to comprehension difficulty and assess their usefulness in building automated prediction models. We investigate readability metrics, vocabulary-based features, and syntax-based features, and show that the best prediction accuracies are possible with a combination of all three.

1 Introduction

Within English second language acquisition there is a fundamental difficulty in defining what is meant by *authentic* as opposed to *non-authentic* or artificial language usage. For example, is authentic usage only the remit of geographical countries where English is their first language or is an official language of communication? Within language teaching the opposition can be made between language usage that is fabricated for the teaching of English as a second language (ESL), and language usage which is not fabricated. This shift between forms of usages can be seen in text books which are used in the learning of ESL where fabricated sentences are often used to highlight specific forms of language or adapted material is incorporated into the reading and listening material.

The proponents of authentic usage tend to highlight the *authentic* as capturing what language is as socio-linguistic utterance in context [1]. For example, Cambridge University Press, one of the major text book publishers in English language teaching (ELT), has a discussion board that highlights the main advantages of using authentic materials in the classroom. The advantages listed include: helping students to learn how to communicate in the *real* world, learning language in context, and increased motivation for learners[3].

[3] `http://bit.ly/2xLHXWh`

There are, however, some disadvantages to using authentic material in English language teaching. Foremost amongst these is that the language is not primarily designed for learning but for communication between native speakers. This can mean that that level of language used in authentic material can be too difficult, in terms of the complexity of sentences, and, more importantly, the use of unfamiliar words or idiomatic expressions. Authentic, but difficult, texts can make the gap between the presumed level of the student or the class and the difficulty of the text too big leading students to quickly lose their motivation. Reliable methods to automatically determine the *comprehension difficulty* of a text could greatly mitigate these disadvantages by making it easy for teachers or learners to source authentic materials of an appropriate level. *Readability metrics* are one long-standing approach to doing this.

Readability is a term used to refer to the overall *understandability* or *comprehension* level of a text. There are a number of established, widely used readability metrics in the literature, such as Flesch and FOG[4]. Whilst these metrics go some way towards determining the comprehension difficulty of text, in general they all tend to focus on specific, narrow features of the language used—most readability metrics are defined as functions over counts of word syllables and/or sentence length. As W.H. DuBay points out: "*The variables used in the readability formulas show us the skeleton of a text. It is up to us to flesh out that skeleton with tone, content, organization, coherence, and design* [6, p. 56]. DuBay's analysis highlights that there are many more features of the language used in a text, beyond those modelled by traditional readability metrics, that impinge on the comprehension difficulty of that text. Investigating these features is the motivation behind the work described in this paper. We analyse how useful different sets of features of the language used in a text are in modelling comprehension difficulty of a text. In this work we consider readability metrics, syntax-based features, and vocabulary-based features.

The paper is structured as follows: Section 2 describes how we designed and created a dataset of texts annotated by comprehension difficulty; Section 3 describes the features we created and used in our models; Section 4 describes the different models we trained and presented the results of our evaluation experiments; in Section 5 we conclude the paper by discussing our results and highlighting some areas of future research.

2 Data

In order to build models to assess the usefulness of different features of the language used in text to predict comprehension difficulty we needed a dataset of texts annotated by comprehension difficulty. The first design decision in creating this dataset was to decide on the comprehension difficulty levels that we would use for annotation. One option would have been to use the Common European Framework of Reference for Languages (CEFR) [8] levels for annotation. Indeed, over the last number of years the development of the CEFR has led to the

[4] See Section 3.1 for more details on readability metrics.

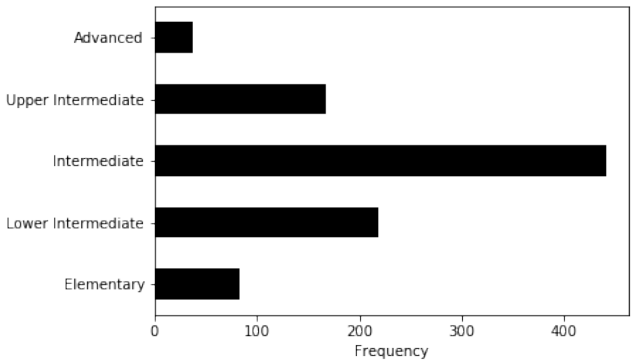

Fig. 1. The distribution of comprehension difficulty levels within the collected corpus

increased awareness of more nuanced understandings of language levels for learners. However, after a review of the CEFR it was decided that for the purposes of this project the CEFR did not give enough detail in terms of language difficulty for comprehension levels for them to be incorporated. Instead we based our comprehension difficulty annotations on the traditional English as Second Language levels which closely follow the Cambridge levels: *Beginner*, *Elementary*, *Lower Intermediate*, *Intermediate*, *Upper Intermediate* and *Advanced*.

Next we collected a corpus of texts whose original purpose was not ESL. The corpus contained 948 texts from a range of international English language online news sources that we expected to include texts at different comprehension difficulty levels. The average length of these texts in words is 457.5 (with a standard deviation of 379.7). We hired a number of ESL teachers to annotate these texts with difficulty levels through a bespoke annotation tool that presented texts to annotators in random order. Our review of the annotations revealed that there were no low level beginner texts in the corpus, this is to be expected as authentic texts at this level are rare. This left us with five levels of difficulty (ESL levels Elementary to Advanced). The distribution of difficulty levels within the corpus is shown in Figure 1.

3 Feature Design

There are a wide range of potential descriptive features that could be used in building a predictive model of text comprehension difficulty. The high-level *domain concepts* identified as important in this problem were: existing readability measures and related features, features based on the vocabulary in a text, and features based on a syntactic analysis of the text. In the following sections we describe the sets of features we developed and used from each of these domains.

3.1 Readability Metrics

There are a number of well-known readability metrics, for example: FOG [10], Flesch [9], and Coleman-Liau [4]. These metrics attempt to measure how easy it is to read a piece of text and are generally a function over the word length (either in terms of syllables or characters) and/or sentence length in a text. For example, Equation 1 defines the calculation of the Flesch readability metric [9]. In the case of the Flesch metric the readability scores range between 0 and 100, where 0 indicates that the text is unreadable and 100 indicates that the text is extremely easy to read. For several of these readability metrics mappings between the metric scores and school levels have been proposed.

$$\begin{aligned} Flesch = 206.835 - 1.015 \left(\frac{total\ words}{total\ sentences} \right) \\ - 84.6 \left(\frac{total\ syllables}{total\ words} \right) \end{aligned} \tag{1}$$

Figure 2 presents a scatter plot matrix (SPLOM) illustrating the linear relationships between a number of standard readability metrics: Flesch, Automated-Readability Index (ARI) [17], Fog, Lix [2], SMOG [14], and Coleman-Liau. The graphs along the main diagonal of the SPLOM present a density plot of the scores generated by the related readability metric when it is applied to documents in our corpus. The off-diagonal scatterplots reveal that many of these readability metrics have strong linear relationships. For example, the Lix and SMOG metrics have a very strong positive linear relationship. These strong linear relationships between many of these readability metrics indicate that many of these metrics are measuring close variants of the same thing. Some of the metrics, however, do appear to be capturing other aspects of readability. For example, examining the scatter-plots that include the ARI metric versus Flesch it appears the linear pattern evident in many of the other scatter-plots breaks-down.

As noted in the introduction, readability metrics do not provide a measure of the comprehension difficulty of a text. For example, text that includes many idiomatic phrases, or novel turns of phrase may get a good readability score but this does not indicate that it will be easy to understand or comprehend such text. That said, readability metrics do provide an objective standard and do provide some information regarding comprehension difficulty. In developing our predictive models we considered all of the readability metrics shown in Figure 2 as input features.

3.2 Vocabulary-based Features

The words used in a text can have a direct impact on the comprehension difficulty of the text. The use of complex words is likely to make a text more difficult to read and to comprehend. This is why so many readability metrics use some measure of word length (syllable or character count) as a proxy for word complexity in the calculation of readability. A striking example of this is

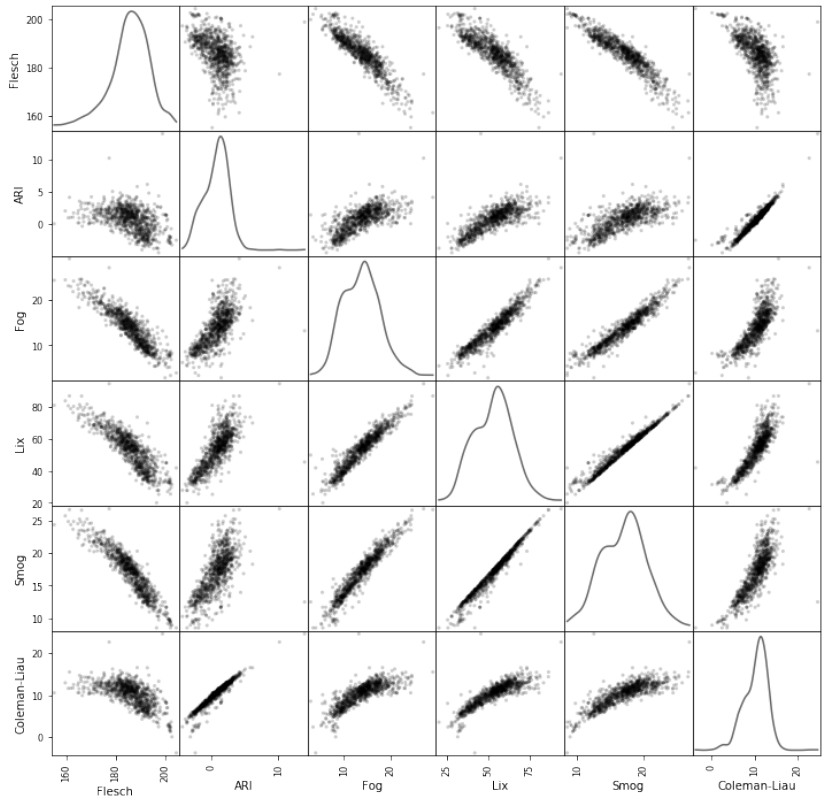

Fig. 2. SPLOM of readability metrics

the FOG , also known as Gunning-Fog, readability metric which explicitly takes the number of *complex words* into account in its calculation, see Equation 2. FOG defines complex words as those words with three or more syllables (where common suffixes are not counted as syllables; e.g., *-ed*, *-ing*, etc.) and which are not proper nouns, familiar words, or compound nouns.

$$FOG = 0.4 \times \left[\left(\frac{total\ words}{total\ sentences} \right) + 100 \left(\frac{complex\ words}{total\ words} \right) \right] \tag{2}$$

The FOG metric is an example of a readability metric that relies on word categories (e.g., familiar words, proper nouns, etc.) as operationalised by pre-specified lists. A challenge for these models is how best to define these word lists.

While the occurrence of specific words might work as a predictor of text difficulty, intuitively documents that include larger numbers of words that are generally rare are likely to be more difficult to understand than documents that primarily use more common words. To achieve this we used what we refer to as *rare-word* features which capture the predominance of rare words within a document in a generalisable way.

We based our rare-word features on word frequencies from the British National Corpus (BNC). We chose to use the BNC as our background corpus because it is a balanced sampled corpus so it is reasonable to extrapolate from the frequencies found in BNC to general English [12]. We defined our rare-word features by binning the words in the BNC into 9 bins based on word-frequency. A challenge faced in the definition of any binning process, however, is to define appropriate threshold's between bins. In this case, the challenge was to define thresholds between common, rare and very rare words. Noting that the words frequencies in the BNC follow a Zipf distribution we defined our bins such that each subsequent bin contained the most common remaining words and the set of words in each bin would account for a predefined percentage of the tokens in the corpus. For example, Bin 1 contained the set of most frequent words in the BNC such that these words accounted for approximately 50% of the tokens in the BNC (this bin contained the 63 most common words in the corpus). Bin 2 contained the set of next most frequent words such that together these words accounted for 25% of the tokens in the BNC (this bin contained 822 words). The other bins were defined in a similar way: Bin 3 contained the remaining most common words that together accounted for 10% of the tokens in the corpus, the words in Bin 4 account for 5% of the tokens in the corpus, Bin 5 also accounted for 5% of the tokens, Bin 6 accounted for 2%, and Bins 7, 8 and 9 accounted for 1% each. Once we had defined our bins we represented the distribution of common and rare words in a document by calculating the percentage of words in a document that belong within each bin. For example, the Bin 1 percentage feature recorded the percentage of words in the document that belonged to Bin 1. Consequently, we developed 9 features based on our word frequency bins: Bin1%, Bin2%, ..., Bin9%. Together, these bin percentage features give an overall sense of the number of very common and very rare words, as well as everything in between, in a document.

We created two other vocabulary based features, one measured the *lexical diversity* of a document and the other the frequency of *named entities* in a document. Lexical diversity measures the range of different words used in a document, with a greater range indicating a higher diversity [15]. Lexical diversity is often used as a measure of text difficulty and to measure the language competency of writers. For example, lexical diversity has been used in studies to measure language competency skills of foreign and second language learners [7]. In our work we used a basic and intuitive measure of lexical diversity known as the *type-token ratio* (TTR) [19]. TTR is calculated as the number of unique words in a text (types) divided by number of words in the text (tokens), see

Equation 3. TTR values range from 0 to 1 with a higher number indicating greater lexical diversity.

$$TTR = \frac{count\ of\ unique\ words}{total\ words} \tag{3}$$

The final vocabulary feature we used was the percentage of words within a document that are part of named entity expressions. We identified the named entities in the text using the named entity recognition module of the Stanford CoreNLP software [13]. The motivation for including a feature based on named entities in our work was that named entities often pose difficulties to ESL student's, particularly those who come from different cultural backgrounds to that from which the authentic text was generated from.

3.3 Syntax-based Features

The occurrence of particular parts of speech and/or syntactic structures may affect the difficulty of a text from an ESL perspective. For example, prepositions and prepositional clauses, conjunctions, subjunctions, adverbs and adverbial clauses can all pose difficulties to ESL students. To capture these syntactic phenomena within our models we generated a set of features by first parsing the texts and then generating features from the parse tree annotations. We parsed the texts using the Stanford CoreNLP software [13]. The Stanford CoreNLP outputs parse trees annotated with the Penn Treebank tagset, for more details on the tagset see [18].

The first set of features we created from the parse trees were the percentages of each word-level part-of-speech (POS) tag in each text. These POS percentages were generated by simply dividing the count of occurrences of each POS tag in a text by the total number of POS tags in the text. The second set of syntactic features generated from the parse trees measured the distribution of syntactic tags in each text (e.g. tags such as *ADJP* adjective phrase, *SBAR* subordinate clause, etc.). These features were defined in a very similar manner to the POS percentage features: we simply counted the number of occurrences of each syntactic tag in parse trees generated from a text and divide these counts by the total number of syntactic tags in this parse tree set.

Inspired somewhat by the relationship between lexical diversity and text difficulty we created two features to capture the diversity of POS and syntactic tags within a text: the first feature simply counted the number of different parts of speech tags that occurred at least once in the trees generated from a text; similarly, the second diversity feature counted the number of different syntactic tags that occurred at least once in the trees generated from a text. This was based on an intuition that a greater range of POS tags or syntactic tags within a single document could cause comprehension difficulties.

The last two features we generated from the parse trees were designed to capture the complexity of the sentences in a text. In 1979 Flesch motivated the inclusion of a parameter based on the length of a sentence within his readability scores as follows:

> *The longer the sentence, the more ideas your mind has to hold in suspense until its final decision on what all the words mean together. Longer sentences are more likely to be complex—more subordinate clauses, more prepositional phrases and so on. That means more mental work for the reader. So the longer a sentence, the harder it is to read.* [9, p. 22]

Certainly Flesch's argument is a good motivation for including sentence length in a measure of readability, and also in a measure of text difficulty. However, sentence length alone does not do justice to the potential differences in difficulty between sentences of the same length. For example, a sentence that includes multiple clause embeddings is likely to be more difficult to comprehend than a sentence of a similar length that is composed of a simple (if long) list. To capture these aspects of complexity we created three other features based on the set of parse trees generated from the text. These were:

1. *Max Embeddedness*: that maximal phrasal parse tree depth for sentences within a text (implemented by iterating through the parse trees for a text, linearising each parse tree, then reading each linearised parse tree from left to right and during this iterative process keeping track of the maximum number of open brackets encountered at any point)
2. *Average Embeddedness*: the average phrasal parse tree depth for sentences within a text (implemented in a similar way to Max Embeddedness.
3. *Average Phrasal Parse Tree Nodes*: simply the average number of phrasal parse tree nodes in sentences within a text.

4 Models

To evaluate the differing power of the different feature sets described in Section 3 to predict the comprehension difficulty of a document we built and evaluated multi-variate prediction models using each feature subset: readability metrics, vocabulary-based features, and syntax-based features. We also consider the performance of a model using the full combined set of features generated, and a model where only a subset of the what appear to be the most useful features are used.

To address the class imbalance in our dataset (we have many more documents at the intermediate level than at any other level, see Figure 1) we have converted from a categorical classification problem across the five different levels to a numeric prediction problem which each level is associated within a numeric score. The mappings are as follows *elementary*: 10, *lower-intermediate*: 30, *intermediate*: 50, *upper-intermediate*: 70, and *advanced*: 90. for prediction problems with ordinal targets this is a sensible approach to handling class imbalance.

To select a subset of the most useful features from the set available we use a simple *rank and prune* feature selection approach [11]. An importance score for each feature that captures the strength of its relationship with the numeric comprehension difficulty target is calculated and features are ordered from strongest to weakest according to these scores. In this case we calculate the F-score [3] for

Feature	Score
Number of Unique POS Tags Used	904.97
Number of Unique Syn. Tags Used	714.67
Word Count	692.37
Maximum Embededness	527.81
Lexical Diversity	463.02
Smog Readability Metric	257.50
Average Sentence Length	236.43
Flesch Readability Metric	236.43
Fog Readability Metric	230.59
Average Embededness	228.75
Average Phrasal Parse Tree Nodes	227.13
Lix Readability Metric	199.99
Coleman-Liau Readability Metric	160.04
% Cardinal Number (CD) POS Tags	88.74
ARI Readability Metric	82.18
Average Word Length	82.18
% Adjective (JJ) POS Tags	50.72
% Noun, Plural (NNS) POS Tags	49.33
% Preposition (IN) POS Tags	44.23
% Fragment (FRAG) Syn. Tags	37.23
% Prepositional Phrase (PP) Syn. Tags	34.58
% Verb, Gerund (VBG) POS Tags	34.47
% Symbol (SYM) POS Tags	33.66
% Bin-7 Vocabulary	33.32
% Unknown (X) Syn. Tags	32.40
% Proper Noun (NNP) POS Tags	31.41
% Subordinate Claus (SBAR) Syn. Tags	30.87
% Wh-determiner (WDT) POS Tags	30.07
% Bin-2 Vocabulary	24.20

Table 1. The top 30 features selected by the feature selection process and their feature importance scores.

each feature. The features with the top 30 scores (chosen to reduce to $frac13$ of the features) are selected for inclusion in the feature selection set. These features and their importance scores are shown in Table 1. It is interesting to note that the most useful features found are the syntax-based counts of the variety of POS and syntactic tags used in a text. It is also interesting to note that a mixture of simple measures (e.g. word count), readability metrics, and vocabulary-based and syntax-based features are included rather than simply a large set of one type.

Feature Set	Mean Squared Error
readability metrics	10.5943
vocabulary-based features	11.8468
syntax-based features	9.3417
all features	9.2349
selected features	9.1373

Table 2. The performance of models trained using each of the five different feature sets.

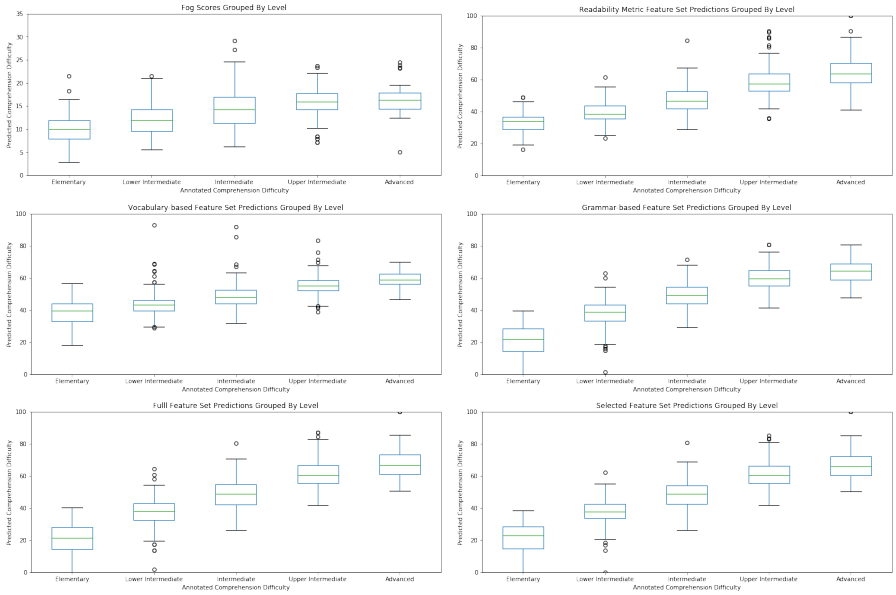

Fig. 3. Box plots illustrating the distribution of model predictions for each level for the different feature sets. From left-to-right and top-to-bottom these are the Fog readability metric alone, the readability metric features, the vocabulary-based features, the syntax-based features, the full set of features, and the subset of the 30 most important features.

In all cases the models used are *support vector regression* (SVR) models [5], as implemented in the Python scikit-learn package[5]. SVR models are chosen as they have been widely shown to be effective across a broad range of multivariate prediction problems well and deal well with features displaying strong co-linearity (for example the different readability metrics).

To evaluate the performance of each model we perform a 10-fold cross validation experiment measuring model performance using *mean absoloute prediction*

[5] `http://scikit-learn.org`

error. The performances of the models built using the five different features sets are shown in Table 2. We can see that the model built using the selected feature subset performs best out of the models test, although it is only very marginally better than the model trained using the full feature set.

We can illustrate the ability of these models to distinguish between the document of different comprehension ability through boxplots that illustrate the distribution of predictions for each difficult level. These are shown in Figure 3. We include the boxplot for the FOG readability metric as a baseline as well as the other feature sets. These boxplots clearly show a progression in the ability of models trained using different feature sets to separate texts into the different comprehension difficulty levels labelled in the text.

5 Conclusions

The ability to automatically rate the comprehension difficulty of texts would greatly reduce the challenge of using authentic text in ESL classrooms and online services. While, the use of readability metrics has been demonstrated as a very useful determination of the general level of a text, this is not sufficient for rating comprehension difficulty. Comprehension difficulty is influenced by more features than just the simple measures of word and sentence complexity incorporated into readability metrics. In this paper we describe an analysis into different types of features of texts that are useful for predicting readability. We consider three different groups of features: readability metrics, syntax-based features, and vocabulary-based features. We base this analysis on a corpus of 948 texts collected from a range of international English language online news sources that were expertly annotated into five of the traditional English as Second Language levels: *Beginner*, *Elementary*, *Lower Intermediate*, *Intermediate*, *Upper Intermediate* and *Advanced*. We perform our analysis by building an evaluating predictive models using different feature subsets extracted from the document corpus.

The first thing this analysis illustrates is a confirmation that, although readability metrics can provide some indications of the of the difficulty of comprehension of a text for ESL, they are not sufficient to do the job of automatic rating accurately. This result highlights that it is necessary to make a distinction between the level of difficulty of comprehension of a text, in particular for English as Foreign language, and the standard readability scores. The second thing we show is that none of the different groups of features lead to models that are better than one trained with features from the different groups combined.

There are many extensions that could be made to the analysis described in this paper. For example, the level of comprehension difficulty of a text can be linked to the cultural context, the use of idiosyncrasies, or the use of idioms, and novel turns of phrase. While the identification of these aspects of text can be done through the use of computational models (see for example see [16] for the identification of idiomatic structures), they all remain open research challenges for computational models of languages. Nevertheless features based on these

aspects could be examined. Similarly, there the use of specific grammar points (e.g. particular tenses) are known to cause comprehension difficulties. While the use of syntax-based features based on POS and syntactic tags captures these to some extent, representing their use more directly in specific features would be beneficial.

References

1. Benveniste, E.: Problèmes de linguistique générale: I. Gallimard (1975)
2. Bjornsson, C.H.: Lasbarhet. Stockholm: Liber (1968)
3. Chen, Y.W., Lin, C.J.: Combining svms with various feature selection strategies. In: Feature extraction, pp. 315–324. Springer (2006)
4. Coleman, M., Liau, T.L.: A computer readability formula designed for machine scoring. Journal of Applied Psychology 60(2), 283 (1975)
5. Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A.J., , Vapnik, V.N.: Support vector regression machines. In: Advances in Neural Information Processing Systems 9, pp. 155–161. MIT Press (1997)
6. DuBay, W.H.: The principles of readability. Tech. Rep. ED490073, Impact Information (2004)
7. Durán, P., Malvern, D., Richards, B., Chipere, N.: Developmental trends in lexical diversity. Applied Linguistics 25(2), 220–242 (2004)
8. of Europe, C.: Common European Framework of Reference for Languages Learning, Teaching, Assessment. Cambridge University Press (2001)
9. Flesch, R.F.: How to write plain English: A book for lawyers and consumers. Harpercollins (1979)
10. Gunning, R.: The technique of clear writing. McGraw-Hill, New York (1952)
11. Kelleher, J.D., Mac Namee, B., D'Arcy, A.: Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT PRess (2015)
12. Leech, G., Rayson, P., et al.: Word frequencies in written and spoken English: Based on the British National Corpus. Routledge (2014)
13. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014), `http://www.aclweb.org/anthology/P/P14/P14-5010`
14. Mc Laughlin, G.H.: Smog grading-a new readability formula. Journal of reading 12(8), 639–646 (1969)
15. McCarthy, P.M., Jarvis, S.: Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods 42(2), 381–392 (2010), `http://dx.doi.org/10.3758/BRM.42.2.381`
16. Salton, G.D., Ross, R.J., Kelleher, J.D.: Idiom token classification using sentential distributed semantics. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 194–204. Association of Computational Linguistics (2016)
17. Senter, R., Smith, E.A.: Automated readability index. Tech. rep., CINCINNATI UNIV OH (1967)
18. Taylor, A., Marcus, M., Santorini, B.: The penn treebank: an overview. In: Treebanks, pp. 5–22. Springer (2003)
19. Templin, M.C.: Certain language skills in children; their development and interrelationships. University of Minesotat Press (1957)