Incorporating Word-level Phonemic Decoding into Readability Assessment

Christine Pinney*, Casey Kennington*, Katherine Landau Wright*, Maria Soledad Pera†, Jerry Alan Fails*
*Boise State University, †TU Delft
{christinepinney, caseykennington, katherinewright, jerryfails}@boisestate.edu
M.S.Pera@TUDelft.nl

Abstract

Current approaches in automatic readability assessment have found success with the use of large language models and transformer architectures. These techniques lead to accuracy improvement, but they do not offer the interpretability that is uniquely required by the audience most often employing readability assessment tools: teachers and educators. Recent work that employs more traditional machine learning methods has highlighted the linguistic importance of considering semantic and syntactic characteristics of text in readability assessment by utilizing handcrafted feature sets. Research in Education suggests that, in addition to semantics and syntax, phonetic and orthographic instruction are necessary for children to progress through the stages of reading and spelling development; children must first learn to decode the letters and symbols on a page to recognize words and phonemes and their connection to speech sounds. Here, we incorporate this word-level phonemic decoding process into readability assessment by crafting a phonetically-based feature set for grade-level classification for English. Our resulting feature set shows comparable performance to much larger, semantically- and syntactically-based feature sets, supporting the linguistic value of orthographic and phonetic considerations in readability assessment.

Keywords: readability, education, phonetics

1. Introduction

Efforts towards automatically assessing readability levels that can help gauge the reading capabilities required to understand a text (e.g., school grade levels) range from using simple formulas to complex models. Allen et al. (2022) points out that simple formulas often use shallow, surface-level features such as the average number of words in a sentence or the number of sentences in a text, which do not capture important aspects of language complexity such as syntax or semantics. More recently, Lee et al. (2021) proposed a model that uses hundreds of linguistic features (for example, the average number of phrase structure parses per sentence in a text) and they incorporated those features in a large language model (LLM) fine-tuning regime to classify readability levels.

While these kinds of advancements in automatic readability scoring show improvement on multiple readability datasets (Meng et al., 2020; Deutsch et al., 2020), we see a common denominator among formulas and complex models (Benjamin, 2012; Lee et al., 2021) of leveraging linguistic patterns in text without explicitly addressing linguistic aspects that emerge during (and represent an integral part of) the reading and writing learning process in the classroom. In this paper we focus on an intended audience that makes use of automated readability scoring: the *teachers* who are directly involved in teaching children how to read and motivating them in their learning (Shuman, 2022). Teachers often use readability formulas to evaluate texts for use in their classrooms Gunning (2003), which could result in false positives and false negatives (i.e., identifying texts that would be inappropriately challenging or too easy for students, respectively), influencing the effectiveness of the materials chosen for the children. However, it is not yet known how existing formulas and models mirror curricular best practices for reading instruction.

For teachers, reading is understood to be a multifaceted process involving (but not limited to) word recognition, phonics, orthography, comprehension, phonemic awareness, and individual motivation (Stahl et al., 2019). Although the general goal of reading is to draw meaning from written text, children must engage with different aspects of language in their attempts to reach this goal as they develop their reading skills. In the beginning stages of reading development and word study (primarily between kindergarten and 5^{th} grade), understanding the relationship between speech sounds and letter symbols (i.e., phonics) is foundational to subsequent stages of development (Tompkins, 2001) through a process referred to as *decoding*. Recent work used phonetic patterns (previously identified for instructional purposes by educational researchers) to determine a set of rules for a state-of-the-art spellchecker designed for children who are learning how to spell (Downs et al., 2020), suggesting that phonetic features could be used to automatically identify readability levels.

While LLMs and transformer architectures have rightfully earned their place as the top perform-

ers in grade-level classification for readability assessment, we posit that they do not serve well the purposes of teachers. The tools teachers use to determine text readability should result in output that is interpretable; the blackbox nature of LLMs lends itself to producing unexplainable results that educators are unable to utilize effectively. Using simpler, more transparent machine learning techniques and classifiers, we echo prior work that showcased feature-based assessment of text readability (Zhang et al., 2013), specifically building on the handcrafted feature set put forth more recently by Lee et al. (2021) to include phonetically-based, orthographic features. By including calculations of features that surround the aforementioned decoding process that every child must progress through when learning to read, we can more closely emulate the assessment that is carried out by educators when determining the readability of a given text.

Our **primary research question** is: How do phonemic/phonetic features that characterize the decoding process contribute to current feature-based readability assessment methods, focusing on children who are in the *learning how to read* category of readers (Beier et al., 2022)?

Empirical explorations conducted using a number of standard readability assessment datasets, along with a dataset comprised of texts explicitly crafted to support reading development, reveal that phonetically-based features that map to orthographic development stages are valuable in readability assessment for early grade levels (i.e., kindergarten to 5^{th} grade). Moreover, outcomes from this work provide insight regarding the current reflection of instructional reading practices in datasets designed for readability and showcase new directions in future research for the consideration of specific linguistic patterns present in classroom reading instruction and the incorporation of these patterns in readability assessment methods and calculations. While we limit our explorations to American English datasets within the context of US/Western Education, results indicate the potential to apply similar approaches with other languages and within further educational contexts.

2. Related Work

Previous work has established that the transformer architecture of LLMs in conjunction with handcrafted feature sets is a powerful methodological structure for readability assessment (Meng et al., 2020; Deutsch et al., 2020; Lee et al., 2021; Li et al., 2022; North et al., 2023; Imperial et al., 2022). Other work highlights neural network approaches, e.g., supervised and unsupervised learning (Martinc et al., 2019) as well as transfer learning for multilingual readability assessment (Azpiazu and Pera, 2019; Madrazo Azpiazu and Pera, 2020; Imperial et al., 2022). While state-of-the-art results were achieved on some readability benchmarks, the authors of Lee et al. (2021) assert that the more urgent contribution of their work was that which highlighted the importance of more traditional machine learning methodologies in areas of natural language processing (i.e., handcrafted features sets for readability assessment). Moreover, Kumar et al. (2023) notes that despite research that shows that LLMs can acquire syntactic (e.g., parts-of-speech) and semantic knowledge implicitly, LLMs tend to overrely on specific vocabulary words making them brittle for a readability-like task. As performance analysis of state-of-the-art methods is often reported on differing datasets and readability levels (Allen et al., 2022), consistency in performance is difficult to determine despite effectiveness.

Other related work looks into readability assessment for second language acquisition (Vajjala and Meurers, 2012) or adults with intellectual abilities (Feng et al., 2009), but here we focus on children who are learning their first language. More directly related to our work, Reyes (2019) explores the impact of phonemic frequencies in readability and word difficulty assessment. They develop a phoneme frequency scale using phonetic transcriptions of commonly-used English words, ranking the frequencies of phonemes from highest to lowest. The authors found that the phonemic frequency score method for readability assessment results in similar readability scores as the Spache score method, which is most suitable for lower grade levels. This indicates that phonemic/phonetic analysis is best at predicting readability assessment scores when applied to lower grade levels. We build on this work by using phonetic and orthographic features; we also systematically test feature importance of resulting models, emulating the feature evaluation and interpretation framework described by Imperial and Ong (2021).

3. Background

Research in Education provides teachers with many different approaches towards reading instruction and word study. A student's reading ability is frequently thought of using the Simple View of Reading (Hoover and Gough, 1980) which posits that an individual's reading ability is the product of their language comprehension and decoding skills (i.e., the ability to accurately assign phonemes to graphemes to "sound out" words). As overall reading is seen as the multiplicative product of these two skills, if a child lacks skills in one area, their reading ability is impacted. A readability formula would ideally also consider both avenues in determining reading difficulty: How challenging would this text be to decode, and how challenging would the text be to comprehend?

Educational researchers have identified the order in which learners typically master decoding, spelling, and comprehending different word patterns. We specifically applied the patterns identified in Bear et al. (2020)'s *Words Their Way*, described below, as this instructional program explicitly identifies the order in which word patterns are typically mastered, and the grade levels where learners can be expected to master each.

3.1. *Words Their Way* Elementary Spelling Stages

Words Their Way, a widely used resource among educators in North America, characterizes each spelling stage with specific orthographic and phonemic/phonetic patterns (Bear et al., 2020). Table 1 shows a synopsis of the orthographic and phonetic/phonemic patterns, which we discuss in-depth below, as they provide the theoretical basis for the features we leverage in our model. A more extensive table of stages with corresponding features and examples can be found in Appendix A.

Letter-Name Alphabetic In this stage, learners are just beginning to recognize that sounds can be visually connected to alphabetic letters in text. Learners typically have an easier time with letters that "say their name" (e.g., b is to *bee*, d is to *dee*, f is to *ef*) versus letters with more complicated sound-name relationships (e.g., w is to *doubleyoo*, y is to *wie*, h is to *aitch*). Within this stage, children will develop articulation skills by isolating letter sounds, including affricates (*ch, j, dr, tr*), voiced/unvoiced pairs (*b* vs. *p*, *v* vs. *f*), stop consonants (*b, d, g, k, p, t*), and continuants (*f, l, m, n, r, s, v, z*). Other features learned during this stage include short vowels (particularly in consonant-vowel-consonant, single-syllable words, such as the *a* sound in *cat* and *sat*, or the *e* sound in *let* and *bet*), consonant digraphs (*sh* in *fish* or *th* in *thin*) and blends (*cl* in *clap* or *sp* in *lisp*), preconsonantal nasals (*mp* in *jump* or *nk* in *pink*), and some consonant-influenced vowels (like the *a* sound in *car* or the *o* sound in *for*). Reading materials for learners in this stage should focus on contrasting these sounds in single-syllable words that are part of similar word families.

Stage	Typical Grade Range	Patterns
Letter-Name Alphabetic	K-1	blends digraphs CVC short vowels
Within Word Pattern	1-2	long vowels basic inflectionals complex consonants
Syllables & Affixes	2-5	compound words syllable junctures advanced inflectionals
Derivational Relations	5+	advanced suffixes Greek & Latin roots assimilated prefixes

Table 1: Stages and Patterns found by Educational research

Within-word Pattern In this stage, learners move from the alphabetic layer of language into the pattern level. In other words, learners (usually children) begin to recognize sound patterns within words, graduating from individual letters to combinations and clusters of language, or phonemes. Long vowels (particularly the effect of a final *e* on otherwise short vowels, as in *hat* vs. *hate*), diphthongs (*ai* in *rain*, *ea* in *team*) and more ambiguous vowels (*oi* in *coin*, *ou* in *count*) are studied, and more complex consonant clusters (*tch* in *catch*, *dge* in *edge*) are explored. Special attention is paid to homophones, or words that sound the same but have different meanings. Reading materials for learners in this stage of reading ability should contrast short and long vowel sounds, incorporating additional consonant-vowel patterns (e.g., CVCe, CVVC) and introducing more complex consonant clusters, basic plural nouns, and irregular past tense verbs. The syllable count should still be largely singular.

Syllables & Affixes Within this stage, disyllabic words are explored with special attention paid to types of syllable junctures and syllabic stress and accent patterns of each vowel. Learners will begin to recognize morphemes within words, focusing in on inflectional endings (e.g., *-ing* and *-ed* in past tense verbs), basic affixes (e.g., comparative suffixes *-er* and *-est*, prefixes *mis-* and *un-*), and compound words. Final unaccented syllables are examined (like the *al* sound in *normal* or the *et* sound in *comet*), as well as special, more complex consonants (e.g., silent initial *k* or *g* as in *know* or *gnat*, hard vs. soft *c* as in *call* vs. *cent*, hard vs. soft *g* as in *gossip* vs. *gentle*, the *f* sound of *ph* as in *graph*). Reading materials for this stage should introduce words with more syllables, as well as more complicated accent/stress patterns of different syllable juncture types. Inflectional endings and simple derivational affixes should be highlighted.

Derivational Relations In this final stage, more complicated affixes are introduced, including more advanced suffixes (e.g., the *shun* sound resulting from *-sion* or *-tion*, adding *able/ible* to root words like *audible* vs. base words like *breakable*) and assimilated or absorbed prefixes (e.g., *il-* in *illogical*, *op-* in *opposite*). Significant focus is given to Latin and Greek roots and prefixes, progressing from most concrete to more abstract in meaning. Sound alternations in vowels upon adding suffixes are also highlighted (e.g., *bomb* versus *bombard* or *sign* versus *signature*). Reading materials should include these more advanced orthographic features.

4. Method: Orthographic, Education-based Patterns for Readability Assessment

We describe the functions used to calculate features derived from the patterns presented in Sec-

tion 3, along with individual tests to show that our functions accurately identify words containing said features. We use the WTW word lists (i.e., lists of words wherein a given feature appears) to ensure the functions behave as expected. We also introduce the datasets that we use in Sections 5 and 6.

4.1. Education Patterns as Features

We map the progression of orthographic patterns throughout each stage into a set of computable features. To assess the relationship between alphabetic letters/letter combinations and resulting phonetic sounds, we use the traditional alphabetic text and corresponding IPA (International Phonetic Alphabet) translations. IPA offers a standardized, written representation of the speech sounds in oral language (IPA). For instance, to distinguish the letter *k* as silent in the word *knowledge*, one can compare the alphabetic text (*knowledge*) and its corresponding IPA translation /nɑɫəʒ/; the absence of *k* in the IPA translation indicates the silent characteristic of the letter within the word.

IPA The use of IPA helps account for sounds that result from different spellings. For example, the advanced suffix pronounced as “shun" or “zhun" (meaning “act, process, or the result of an act or process" (Bear et al., 2020)) can be spelled in many different ways. Sometimes it manifests as *-tion* (as in *reaction*), other times it is represented with a *-cian* (as in *musician*), and in other instances it appears as *-sion* or *-ssion* (as in *illusion* or *compression*). With IPA, these various spellings can be represented with only two phonetic notations: /ʃən/ and /ʒən/. This helps simplify the search for this advanced suffix feature. The algorithm below shows how we formalized this process with "shun" as the example in Algorithm 1.

Algorithm 1 Searching for advanced suffix "shun" in *musician*, *illusion*, and *compression*

```
shun_suffix ← [ʃən, ʒən]
ipas ← [mjuzɪʃən, ɪɫuʒən, kəmpɹɛʃən]
words_with_feature = 0
total_words = 0
feature_score = 0
for ipa in ipas do
    total_words + = 1
    for suffix in shun_suffix do
        if suffix in ipa then
            words_with_feature + = 1
        end if
    end for
end for
feature_score ← (words_with_feature / total_words)
return feature_score
```

Regular Expressions For certain features/patterns, we use regular expressions to find specific substrings in words as the phonetics of the alphabet reflect many of the patterns, but not all. To improve efficiency and limit these searches to those words that are relevant for a specific feature/pattern, we take into account the parts-of-speech (POS) for each word. For example, when searching for words that have the inflectional ending *-ed* (past tense verb), we want to avoid flagging words like *bed* or *red* because this orthographic feature only applies to verbs, not nouns or adjectives. By eliminating words with POS other than verb, we limit our regular expression searches to only those words that qualify as relevant.

To illustrate this algorithm, suppose we have the word “asking” and its corresponding POS tag "VBG" and we would like to check if it contains the *-ing* inflectional ending for verbs. First, to avoid incorrectly flagging a word like *king*, we would want to check that the POS tag is indeed in the tagger’s list of verb tags: *VBD*, *VBG*, *VBN*, *VBZ*. If the tag is included in the list (which, in this case, it is), we then use a regex search to look for the *-ing* suffix using the regex [ing$]. If this search returns true, then this feature exists within the given word.

Similarly, we use syllable counts for each word to further limit regular expression evaluations to qualifying text. For instance, when searching for words that follow the vowel-consonant-consonant-vowel doublet syllable juncture characteristic of the Within Word Pattern stage of reading (illustrated in words such as *pretty* and *blossom*), we limit the search to words with two syllables.

Variable-length Language Models To locate words within which a Greek or Latin root is present (e.g., matri*arch*y, *tang*ible, con*duct*or), we employ suffix tree language models (STLMs) (Kennington et al., 2012). An STLM acts as a variable-length *N*-gram language model, which can represent long sequences with better probabilities (i.e., lower perplexity) because the sequences are not limited to short n-grams. STLMs are suited to this task because they represent each character sequence present within a word while avoiding error-prone word segmentation characteristic of other approaches; they also require minimal storage space and faster computation time (Kennington et al., 2012). Because Greek and Latin roots are substrings of characters within a longer word, STLMs employed at the character-level work well in capturing these particular sub-sequences within words. We create and train individual STLMs for each root (30 Greek roots and 93 Latin roots) using the WTW word lists provided for each root. When applied to a word, the STLM returns the probability that the sequence of characters exists within its language corpus; a higher probability indicates the presence of this feature within the word. The probabilities for each word in a specific data entry are summed and averaged to represent the pattern as a feature.[1]

[1]We determined empirically that using traditional Latin alphabetic representations yields more reliable probabili-

It should be noted that the features are not perfect. For instance, though there are many strategies for identifying compound words within text, we are not aware of any algorithm that can do so without mistakes. Here, we have implemented a greedy algorithm that correctly identifies compound words with 98.7% accuracy (using a list of 314 words from WTW), but would mistake the word *asking* for a compound word (i.e., identified as *as* and *king*), consequently resulting in false positives.[2]

4.2. Data

In our exploration, we use the Reading AtoZ (RAZ),[3] WeeBit (Vajjala and Meurers, 2012), and Science (Nadeem and Ostendorf, 2018) datasets. In terms of grade level coverage, these datasets span from kindergarten to 12th grade, but we align the levels to consider kindergarten through 5^{th} grade. Phonetic instruction is characteristic of earlier stages of reading development, so we focus on this grade level range as it aligns with these stages.

Reading AtoZ (RAZ). This dataset was created and labeled by educators and did not require extensive cleaning. It has the strongest positive correlation between grade and text length (93.9% correlation), and includes coverage of kindergarten through 5^{th} grade with an additional label for any text recognized as beyond 5^{th} grade (5+).

WeeBit. This dataset required extensive cleaning to remove non-English characters and residual content from the original web scraping process.[4] This dataset provides coverage of 2^{nd} grade to tenth grade, so we restrict our samples to 2^{nd} through 5^{th} grade with 400 samples from each.

Science. This dataset provided the widest range of grade levels, offering coverage from kindergarten through twelfth grade, though kindergarten to 3^{rd} grade is compressed to one category (K-3). We limit the samples to kindergarten (K-3) to 5^{th} grade.

Comparison While the RAZ dataset has a very strong correlation between grade level and length of text (93.9%), that same correlation is not present to the same extent in the WeeBit (5.23%) and Science (-11.8%) datasets. It was discovered that some features/specific feature subsets of the Lee et al. (2021) feature set were calculating total counts without normalization. This resulted in 65 of the 255 features being over 90% correlated with length in the RAZ dataset. Granted, the length of text is important to consider in readability assessment, but it does not tell the whole story in terms of semantic, syntactic, and phonetic complexity. For our purposes, these features are removed from the calculations to avoid simply reinforcing this existing correlation. This is important because the RAZ data was designed by educators, thereby serving as a dataset more reflective of readability as determined by teachers and educators when compared with datasets like WeeBit or Science.

5. Experiment

To assess how our extracted orthographic features and patterns impact the accuracy of current readability assessment methods, we calculated word-level, phonetic feature scores and evaluate the performance of these features both in comparison to as well as combined with the Lee et al. (2021) feature set using several different models.

Procedure Similar to Lee et al. (2021), we preprocessed all data across datasets to remove special characters and stopwords, convert to lowercase, and tokenize. Feature vector extraction is performed at the document level. To assist with feature calculations and reduce time complexity, we apply these word-level processes before feature extraction:

- *IPA translation of text.* Some features require knowledge of both the textual representation and the phonetic representation of a word. The IPA translation allows for an alternate representation of a word that provides insight into the word's pronunciation. For this translation, we use a dataset with 135,006 words[5]; if a word does not have a corresponding translation, it is not included in the calculations. The words in this dataset offer 97.5%, 94.5%, and 95.8% coverage of the RAZ, WeeBit, and Science datasets, respectively.
- *Parts-of-speech*. Some features only apply to words with certain parts of speech (e.g., past tense only applies to verbs), so knowledge regarding the part of speech of a given word is required. We employ the POS tagger from NLTK (Bird et al., 2009).
- *Syllable counts.* Some features only apply to single-syllable words whereas others only apply to multisyllabic words, so awareness of the syllable count of a given word is necessary. We use a modified version of the syllable tokenizer from NLTK (Bird et al., 2009).[6] See Appendix B for the modified code.

ties than IPA representations.

[2]Code to find the WTW patterns can be found at https://github.com/BSU-CAST/phonemic_feature_functions

[3]The RAZ data is proprietary from https://www.readinga-z.com

[4]Code to clean WeeBit: https://github.com/shlomihod/deep-text-eval/blob/master/data/weebit/prepare_weebit.py

[5]See https://github.com/open-dict-data/ipa-dict

[6]The NLTK syllable tokenizer often overlooked the 'e' rule for vowel pronunciation (e.g., *hat* vs. *hate*). We adjusted this tokenizer slightly to account for this rule.

To calculate the education-based features that represent the patterns in each of the WTW stages, we used regular expression evaluations informed by syllable counts and POS determinations, as well as the STLMs. Extensive word lists provided by WTW serve as positive and negative examples of the features and are used to test the functionality of the algorithms for finding the patterns (Bear et al., 2020). All instances are normalized for token (i.e., word or character) length. For features that employ a STLM (e.g., Greek and Latin roots), the calculation is measured by the probability that a given word belongs to the language corpus of a given language model. As these calculations are sensitive to words/text that are uncharacteristic of a given stage, Frye list words are excluded.[7]

We carried out an ablation study and feature importance analysis to analyze the individual contribution of each feature/feature set. These comparisons are made both in the case of a multi-class classification task involving all grade levels (i.e., kindergarten through 5^{th} grade), as well as a binary classification task wherein individual grade boundaries are examined (i.e., kindergarten to 1^{st} grade, 1^{st} grade to 2^{nd} grade, etc.). Highlighting grade boundaries in addition to classification with all grade levels provides insight into what individual grade levels or grade level transitions the feature set and/or model may be particularly good (or bad) at distinguishing between. Following the experimental framework used by Lee et al. (2021), the feature vectors and their corresponding grade-level labels are fed into three classifiers (Logistic Regression, Support Vector Machine, and Random Forest Classifier), where we include one neural network (Multilayer Perceptron) in place of a LLM (Pedregosa et al., 2012).

Metrics & Baseline We compared the accuracy scores of classification by grade level of the Lee et al. (2021) feature set alone, the Lee et al. (2021) feature set with WTW feature set appended, and the WTW feature set alone to measure the contribution of the orthographic features added, as well as the individual performance of these features when employed independent of syntactic and semantic features. The established Lee et al. (2021) feature set is used as our baseline against which to analyze our feature set's performance. We first evaluate performance with multi-class classification including all the grade levels within each dataset. To evaluate the feature sets' performance at grade boundaries (e.g., 1^{st} grade to 2^{nd} grade, 3^{rd} grade to 4^{th} grade, etc.), we also assess binary classification using each feature set at each grade boundary. For each classification task, the feature sets are fed into each of the four models and the average accuracy and F1 score from the best performing model is recorded.

To evaluate the importance of individual features in each feature set, we carried out feature importance analysis with *permutation feature importance*, a model inspection technique from sci-kit learn (Pedregosa et al., 2012). The best-performing model is given to the permutation feature function to determine how much the model's accuracy depends on each feature within the feature set, outputting those features upon which the model most depends. This analysis is performed on the Lee et al. (2021) feature set as well as the combined feature set with both the Lee et al. (2021) and WTW feature sets.

5.1. Results

Table 2 shows the results of this experiment. With the full-grade, multi-class classification task, the Lee et al. (2021) features outperformed the WTW features when employed independently, but feature importance analysis with the combined feature sets shows us that the WTW features contribute valuable information (Appendix C). Furthermore, the WTW features had a comparable performance to (sometimes better than) the Lee et al. (2021) feature set when tasked with binary classification at grade boundaries (Table 3). For both classification types, the two top-performing models were (1) Random Forest Classifier and (2) Logistic Regression.[8]

Multi-class Classification While the Lee et al. (2021) outperformed the WTW feature set with full grade level classification with each dataset, it is worth noting that classification with just the WTW feature set still achieved a comparable accuracy given the number of features in each set. With 190-255 features, the Lee et al. (2021) feature set resulted in 81% accuracy with RAZ, 61% accuracy with WeeBit, and 54% accuracy with Science; with just 15 features (a sizeable difference), the WTW feature set still achieved a result of 71% accuracy with RAZ, 57% accuracy with WeeBit, and 40% accuracy with Science. After removing the highly correlated features from the Lee et al. (2021) feature set in the RAZ dataset (those with a correlation with length greater than 90%) and appending the WTW feature set, accuracy returned to that achieved by the Lee et al. (2021) feature set alone before removing features. A statistical significance p-value of 0.19 when tested with the Lee et al. (2021) feature set and the Bear et al. (2020) feature set indicates no significance, but this reveals that a limited feature set of only 15 features that are designed to

[7] High frequency words, or sight words, are words that children learn (often uncharacteristically early in development) due to their regular and/or frequent use by surrounding language speakers and outlets (e.g., teachers, parents, digital media, books, etc.). Word lists by grade can be found at https://sightwords.com/sight-words/fry/.

[8] RFC hyperparameter: n_estimators=500. LR hyperparameter: max_iter=3000. Both models utilized a train/test split with test_size=0.08.

Dataset	Feature Set	Set Length	Accuracy	F1 Score
RAZ	Lee	190	81%	73%
	Lee & WTW	205	81%	71%
	WTW	15	71%	65%
WeeBit	Lee	190	61%	60%
	Lee & WTW	205	61%	60%
	WTW	15	57%	57%
Science	Lee	190	54%	52%
	Lee & WTW	205	50%	42%
	WTW	15	40%	37%

Table 2: Performance across feature subsets.

capture phonetic patterns is able to reach a comparable accuracy of a much larger feature set that incorporates measurements of semantics and syntax, implying that phonetically-based features provide value in readability assessment.

Feature Importance We conducted a feature importance analysis to identify valuable features with just the Lee et al. (2021) feature set alone, the Lee et al. (2021) feature set and the WTW feature set together, and finally the WTW feature set alone. Before excluding features or feature subsets from the Lee et al. (2021) set, the most important features in this feature set were often those highly correlated with length (greater than 90%). Removing these features did not significantly impact the accuracy of the feature set when given to a classifier (within 1-2%), but it resulted in less correlated features being more important. Before this removal, when appending the WTW feature set to the Lee et al. (2021) feature set, those highly correlated features from Lee et al. (2021) remained the most important features. After this removal, however, the features in the WTW feature set were more often recognized as important (Appendix C).

Binary Classification Between Grade Boundaries To assess how both feature sets performed with distinguishing between two grades (i.e., grade boundaries), the multi-class classification task was broken down into a binary classification task using only the RAZ dataset; using the dataset curated by educators seemed most appropriate for assessing grade-boundary classification with a phonetic feature set designed to map to educational development stages. Results are summarized in Table 3. When combined (after removal of Lee et al. (2021) features with over 90% correlation), the feature set does well with the kindergarten to 1^{st} (99%), 1^{st} to 2^{nd} (87%), and 2^{nd} to 3^{rd} (86%) grade boundaries. The WTW feature set alone outperforms both the Lee et al. (2021) feature set and the combined feature set with the 3^{rd} to 4^{th} grade boundary (81%) and outperforms the Lee et al. (2021) feature set alone with the 4^{th} to 5^{th} grade boundary (74%), indicating potential for phonetically-based features to better designate grade boundaries in readability assessment. A p-value of 0.043 with classification at the 3^{rd} to 4^{th} grade boundary (where the Bear et al. (2020) feature set outperforms the Lee et al. (2021) feature set and combined feature set) indicates statistical significance at this boundary in particular.

Boundary	Feature Set	Accuracy	F1 Score
kin | 1^{st}	Lee	98%	98%
	Lee & WTW	**99%**	**99%**
	WTW	92%	91%
1^{st} | 2^{nd}	Lee	87%	82%
	Lee & WTW	87%	82%
	WTW	81%	72%
2^{nd} | 3^{rd}	Lee	85%	84%
	Lee & WTW	**86%**	**85%**
	WTW	76%	73%
3^{rd} | 4^{th}	Lee	76%	73%
	Lee & WTW	78%	76%
	WTW	**81%***	**79%**
4^{th} | 5^{th}	Lee	71%	72%
	Lee & WTW	74%	72%
	WTW	74%	72%

Table 3: Performance across grade boundaries. Asterisks indicate statistical significance.

6. Analysis

Our results bring attention to some important aspects of current methods employed and datasets used in readability assessment. First, it would appear from the WTW feature set's comparable performance independently (with both multi-class and binary classification) and individual feature importance when appended to a larger semantic/syntactic feature set that phonetic-based features do indeed provide linguistic value that could be useful in readability assessment. Second, our findings provide insight regarding the current state of datasets used for readability assessment in terms of reflecting the instructional practices (e.g., phonics) of reading education. Third, the potential correlation between text length and grade level in data when carrying out calculations for readability assessment is worth properly accounting for.

The Importance of Phonetic Considerations The resulting feature importance of the orthographic features we have created when appended to the existing semantically- and syntactically-based feature set from Lee et al. (2021) implies that recognition of phonetic linguistic aspects in the text had not yet been addressed in calculations and proves valuable for readability assessment. Their linguistic value is further depicted by the comparable performance of these features when employed independently from other feature sets, with both multi-class and binary classification. With binary classification specifically (i.e., distinguishing grade boundaries), phonetic feature calculations can be even more informative than their semantic and syntactic partners, as is demonstrated at the 3^{rd} to 4^{th} and 4^{th} to 5^{th} grade boundaries. The use of phonetic features that map to familiar linguistic patterns in stages of reading development offers improved interpretability to teachers.

The Characteristics of Readability Assessment Data The RAZ dataset was created and classified by educators; i.e., domain experts. In contrast, both the WeeBit and the Science datasets were curated from existing online content and labeled by individuals outside of the Education domain. This would suggest that the data within the RAZ dataset has stronger roots in education and should therefore better align with the linguistic features that map to the orthographic stages of development that education research presents; our analysis suggests otherwise. As hypothesized, our phonetically-based features do have the most positive impact when applied to the education-based RAZ dataset in comparison to the other two datasets, but when further analyzing the presence of these orthographic features within the data, we encountered unexpected results. When looking at features characteristic of the Derivational Relations stage, for instance, we would not expect those features (e.g., Greek/Latin roots, advanced suffixes, etc.) to be present in data labeled as readable for kindergarten or 1^{st} grade students, but such appears to be the case. While we controlled for the uncharacteristic phonetic nature of high-frequency words, there were still considerable instances of features within a grade level of text that were expected to only be present in more advanced grade levels.

The process of learning phonics (and, subsequently, how to read) often involves contrastive word-study with focus on those linguistic features characteristic of orthographic development stages that we have outlined here. For text to truly be readable for a given student, an educator must present reading material wherein a large majority of the words must only contain features that map to that student's current stage of development, perhaps offering a handful of individual words outside of that stage to provide a challenge or serve as an introduction to features from a subsequent stage; when creating reading materials and labeling content as readable for a certain grade level, this word-level attention to detail could prove unmanageable. This may help explain why features characteristic of later stages of spelling and reading development emerged in the data (even within the dataset that was designed by educators we can assume are knowledgeable in terms of phonics instruction), but this also suggests that annotating data for readability assessment is an intricate task.

The Effect of Text Length Before removing the highly correlated features from the Lee et al. (2021) feature set, the classifiers were more often within a grade level when they made the incorrect choice. In other words, if a given piece of text was labeled as 4^{th} grade and a classifier mislabeled the text, it was often labeled as 3^{rd} or 5^{th} grade and not often 2^{nd} or 1^{st}. After removing those highly correlated features, however, the classifiers were less likely to guess incorrectly within this range, indicating that those highly correlated features were leading the classifiers to depend on the length differences to distinguish at grade boundaries.

This dependence is not entirely misguided, as it is evident that text for kindergartners and text for 6^{th} graders does often have drastic differences in terms of length, but dependency on length becomes counter-productive when looking at texts more similar in length. For instance, the sentence "What is movement?" may be very short in length, but being able to comprehend the meaning of "movement" (Body movement? A political movement? Movements in the financial market?) is likely out of scope for younger students. Thus, over-dependence on text length would lead to mis-classification of text, so proper measures should be carried out (i.e., normalization) for this kind of error to be avoided.

7. Conclusion

Like many tasks in NLP, readability assessment is complex. Despite the ability of large language models to excel at this task when compared to more traditional machine learning methods, it is important to keep in mind the audience of readability assessment tools when we employ uninterpretable, blackbox techniques. Educators require a tool that assists them not only in measuring readability, but also in interpreting the semantic, syntactic, and phonetic complexities within a piece of text; more traditional techniques offer this improved interpretability. We extend the established semantically- and syntactically-based feature set created by Lee et al. (2021) to include phonetic feature calculations, demonstrating the linguistic value of measuring orthographic features that map to developmental reading/spelling stages grounded in Education. We also reflect on the nature of readability assessment data and its subsequent effect on readability assessment itself, especially as it pertains to word-level feature calculations; contrastive word-study practices in the classroom environment aren't demonstrated in readability data, and this has implications worth considering when trying to measure word-level feature scores.

Future work towards a fast, interpretable readability algorithm that provides informative output beyond a single grade level score is worth exploring. While in this iteration of our work we did not incorporate all the features in the WTW stages, they could have the potential to provide further linguistic value to an orthographic feature set. Additionally, phonetics can differ among languages, and exploring how phonetically-based features contribute to readability assessment in languages other than English presents another path for future investigation.

Acknowledgements This work is supported by NSF Award #1763649. Thanks to members of the CAST team including, particularly Benjamin Bettencourt for their support with this paper. Finally, we thank the anonymous reviewers for their valuable feedback.

Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. 2022. Supercalifragilisticexpialidocious: Why using the "right" readability formula in children's web search matters. In *Advances in Information Retrieval*, pages 3–18. Springer International Publishing.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Trans. Assoc. Comput. Linguist.*, 7:421–436.

Donald R. Bear, Marcia Invernizzi, Shane Templeton, and Francine Johnston. 2020. *Words Their Way: Word Study for Phonics, Vocabulary, and Spelling Instruction*. "Pearson Education, Inc.".

Sofie Beier, Sam Berlow, Esat Boucaud, Zoya Bylinskii, Tianyuan Cai, Jenae Cohn, Kathy Crowley, Stephanie L Day, Tilman Dingler, Jonathan Dobres, Jennifer Healey, Rajiv Jain, Marjorie Jordan, Bernard Kerr, Qisheng Li, Dave B Miller, Susanne Nobles, Alexandra Papoutsaki, Jing Qian, Tina Rezvanian, Shelley Rodrigo, Ben D Sawyer, Shannon M Sheppard, Bram Stein, Rick Treitman, Jen Vanek, Shaun Wallace, and Benjamin Wolfe. 2022. Readability research: An interdisciplinary approach. *Foundations and Trends® in Human–Computer Interaction*, 16(4):214–324.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ. Psychol. Rev.*, 24(1):63–88.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA $\rightarrow$ Online. Association for Computational Linguistics.

Brody Downs, Oghenemaro Anuyah, Aprajita Shukla, Jerry Alan Fails, Sole Pera, Katherine Wright, and Casey Kennington. 2020. KidSpell: A Child-Oriented, Rule-Based, phonetic spellchecker. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6937–6946, Marseille, France. European Language Resources Association.

Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics on - EACL '09*, Morristown, NJ, USA. Association for Computational Linguistics.

Thomas G Gunning. 2003. The role of readability in today's classrooms. *Topics in Language Disorders*, 23(3):175.

Wesley A Hoover and Philip B Gough. 1980. The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*.

Joseph Marvin Imperial and Ethel Ong. 2021. Under the microscope: Interpreting readability assessment models for Filipino. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 1–10, Shanghai, China. Association for Computational Lingustics.

Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antonio Ibanez, Ranz Sapinit, and Mohammed Hussien. 2022. A baseline readability model for Cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 27–32, Seattle, Washington. Association for Computational Linguistics.

IPA. International phonetic association.

Casey Redd Kennington, Martin Kay, and Annemarie Friedrich. 2012. Suffix trees as language models. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC). Istanbul, Turkey*, pages 446–453, Istanbul, Turkey. European Language Resources Association (ELRA).

Yaman Kumar, Swapnil Parekh, Somesh Singh, Junyi Jessy Li, Rajiv Ratn Shah, and Changyou Chen. 2023. Automatic essay scoring systems are both overstable AndOversensitive: Explaining why and proposing defenses. *Dialogue and Discourse*, 14(1):1–33.

Bruce W Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana,

Dominican Republic. Association for Computational Linguistics.

Wenbiao Li, Ziyang Wang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and Length-Balanced loss.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2020. An analysis of transfer learning methods for multilingual readability assessment. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20 Adjunct, pages 95–100, New York, NY, USA. Association for Computing Machinery.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2019. Supervised and unsupervised neural approaches to text readability.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. ReadNet: A hierarchical transformer framework for web article readability analysis. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 33–49. Springer International Publishing, Cham.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9):1–42.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2012. Scikit-learn: Machine learning in python. pages 2825–2830.

Vera Paola E Reyes. 2019. Exploring the use of the phoneme frequency scale method in determining word difficulty levels and readability scores. In *Proceedings of the 2019 7th International Conference on Information and Education Technology*, ICIET 2019, pages 284–288, New York, NY, USA. Association for Computing Machinery.

Clarice K Shuman. 2022. *Exploring the Effects of Teachers' Motivation to Read on Students' Motivation to Read*. Ph.D. thesis, Kennesaw State University.

Katherine A Dougherty Stahl, Kevin Flanigan, and Michael C McKenna. 2019. *Assessment for reading instruction*. Guilford Publications.

Gail Tompkins. 2001. *Literacy for the 21st Century: A Balanced Approach 7th Edition*. Pearson.

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

Lixiao Zhang, Zaiying Liu, and Jun Ni. 2013. Feature-based assessment of text readability. In *2013 Seventh International Conference on Internet Computing for Engineering and Science*. IEEE.

A. WTW Feature Set

WTW development stages with associated features and examples from WTW word lists (Bear et al., 2020).

Stage	Feature	Examples
Derivational Relations	Greek Roots	**hydra**te, epi**derm**is
	Latin Roots	**circum**ference, re**spir**ation
	Advanced Suffixes	aud**ible**, hesit**ancy**, opti**cian**
	Assimilated Prefixes	**il**logical, **ag**gregate, **ob**scure
Syllables & Affixes	VV Syllable Juncture	cr**e-a**te, l**i-a**r, p**i-a**no
	VCCV Syllable Juncture	ch**ap-te**r, p**ub-li**c, s**an-da**l
	VCCV Doublet Syllable Juncture	bl**iz-za**rd, p**at-te**rn, m**am-ma**l
	VCCCV Syllable Juncture	p**ump-ki**n, d**ol-phi**n, b**ot-tl**e
	VVCV SyllableJuncture	s**ea-so**n, **eas-y**, fl**oat-e**d
	Compound Words	**bed**room, **head**light, **snow**flake
	Inflectional Endings for Adjectives	help**less**, bod**ily**, care**ful**
	Advanced Inflectional Endings	hopp**ed**, hopp**ing**, hop**ing**
Within Word Pattern	Basic Inflectional Endings	plant**s**, beach**es**, pick**ed**
	Complex Consonants	lu**nch**, fu**dge**, **k**nock
Letter-Name Alphabetic	CVC Short Vowels	**cam**p, **tes**t, s**tom**p, sh**run**k

B. Modified NLTK Syllable Tokenizer

Algorithm 2 Incorporating the "e" rule for vowel pronunciation in syllable tokenization

```
SSP ← SyllableTokenizer()
result ← SSP.tokenize(word)
if re.search("e$", result[len(result) - 1] then
    modified ← " ".join([result[i] for i in [len(result)
- 2, len(result) - 1]])
    result[len(result) - 2] ← modified
    del result[len(result) - 1]
end if
return result
```

C. Feature Importance

Analysis of most important features in the combined Lee et al. (2021) and WTW feature set with the RAZ dataset **before** removing features from the Lee et al. (2021) feature set with high correlations to text length. Feature names are provided by Lee et al. (2021) and the corresponding correlations are listed for each feature.

to_VeTag_C	0.982
ra_SuAvP_C	0.493
to_VePhr_C	0.976
to_ContW_C	0.994
to_SbL1C_C	0.995
to_SbCDC_C	0.981
ColeLia_S	0.997
as_Token_C	0.928
CorrAjV_S	0.895
to_FTree_C	0.997

Analysis of most important features in the combined Lee et al. (2021) and WTW feature set with the RAZ dataset **after** removing features from the Lee et al. (2021) feature set with high correlations to text length. WTW features are bolded.

compound_words
BClar20_S
at_SbSBW_C
at_SbFrL_C
BClar15_S
BClar05_S
ONois05_S
ra_SuNoT_C
adv_inflectional_endings_adj
ORich05_S
ra_AvVeP_C
at_FuncW_C
OClar05_S
advanced_suffixes
ORich10_S
ra_NoVeP_C
ra_SuAvP_C
vvcv_juncture