<Investigating the use of Phonemes as Readability Signals>

<An empirical exploration>

<Mark Musa Mitrani Sabah[1]>

Supervisor(s): <Maria Soledad Pera[1]>

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: <Mark Musa Mitrani Sabah>
Final project course: CSE3000 Research Project
Thesis committee: <Maria Soledad Pera>, <Avishek Anand>

Abstract

In this paper, we aim to uncover previously unknown relationships between readability and phoneme-related features of text. Phonemes are the smallest unit of sound that can distinguish one word from another. Phonemes encode pronunciation, therefore they encode phonetic/auditory details about words that letters do not capture. However, the use of phonemes to aid readability estimation has thus far been limited in literature. This paper aims to bridge this knowledge gap by investigating the relationship between readability and individual phonemes, groupings of phonemes, and phoneme-derived features. The experiments are carried out on the WeeBit corpus. Our findings indicate that phoneme-related properties on their own do not serve as accurate indicators of text complexity.

1 Introduction

The ability to read enables us to access information and become educated in important matters. However, not all texts are created equal in terms of their ease of reading and comprehension. Some texts effortlessly guide readers through their content, while others present challenges that demand greater cognitive effort. The readability of a text is defined as its "quality of being easy and enjoyable to read" [1].

Understanding and assessing the readability of texts is crucial for facilitating effective communication and education. In educational settings, readability plays a crucial role in assigning appropriate reading materials to learners [9]. Matching texts to learners' language and reading proficiency levels is essential for optimizing learning outcomes. Educators use readability assessments to ensure that students receive texts that align with their abilities, providing appropriate challenges while avoiding frustration [7].

Research into readability assessment aims to address this need by studying "the features of texts and readers that govern reading ease" [6]. Though readability was historically assessed through readability formulas, advances in natural language processing and machine learning have allowed the development of various models that are able to measure readability significantly more accurately [3]. The development and refinement of such models hinges partly on explorations into the effectiveness of various text features in assessing readability [11].

Phonemes are "the smallest class of sounds that leads, in a specific language, to differences in meaning" [8]. These distinct sound units differentiate words from one another and encode important phonetic and auditory information. A proper understanding of phonemes has been shown to play a crucial role in word recognition, language acquisition, and the development of proper reading skills [5; **?**].

Existing literature has hinted at the potential relationship between phoneme use and readability. For instance, it has been suggested that restricting phonemes to a certain subset could enhance the readability of a text for beginning readers [2]. This idea is supported by findings in the field of speech-language pathology: According to a study regarding English children's mean age of consonant acquisition, children tend to acquire certain groups of phonemes earlier than others [4]. Due to their increased familiarity with the phoneme groups that were acquired earlier, children may have a preference towards them, therefore perceiving texts consisting of those groups as more readable. It is plausible, then, to expect that certain phonemes or groups of phonemes would be more prevalent in lower readability levels. Moreover, researchers have used phoneme-related measures to measure the decoding difficulty of a text, and decoding difficulty is "important for accurately matching young readers to appropriate text and scaffolding reading development" [13]. Decoding difficulty can thus factor into the readability of a text, therefore phoneme-related measures could also prove useful to assess readability. Lastly, a previous study has shown that estimating readability by analyzing the frequency of the phonemic composition of words produces scores similar to the Spache score, an established metric for assessing readability [12]. However, an initial literature review reveals that the relationship between phonemes and text complexity has not been explored further in current research.

Given the aforementioned findings of related literature, we posit that there is a need to explore this gap further. In this work, we set out to advance knowledge about how phonemes influence text complexity. To that end, we conduct an empirical exploration of the relationship between readability and phonemes. In order to drive our exploration, we formulate our research question **RQ** as follows:

Can phonemes serve as indicators of the level of complexity of English texts?

In order to address our RQ, we conduct data experiments on the WeeBit corpus, a collection of text samples categorized under 5 readability levels covering the ages 7-16 [15]. Our approach is based on the exploration of potential relationships between phonemes and text complexity through three distinct perspectives, hereafter referred to as 'lenses'. Each lens essentially represents a different aspect or layer of phonemic data that we aim to examine:

- The first lens looks at individual phonemes and their relationship to readability, providing us with a view of the correlations of each phoneme's frequency to the readability level of a text.
- The second lens considers groups of phonemes categorized by age of acquisition and manner of pronunciation. Through this perspective we aim to investigate the frequency of each group's phonemes across readability levels.
- Lastly, with the the third lens, we focus on numeric features derived from phonemes. This lens allows us to examine the relationship between abstract metrics calculated using phoneme data and readability levels.

By exploring the relationship between phonemes and text complexity through these three lenses, we aim to obtain a multi-faceted understanding of how phonemes influence readability. Further details of how we employ these lenses are provided in Section 2 (Methodology).

Our research is a starting point in filling the identified gap by exploring whether phonemes could potentially serve as indicators of text complexity. This exploration is not only needed to better understand the features that contribute to text readability, but also to enhance the accuracy and comprehensiveness of readability estimation models. Incorporating the understanding gained from this research could lead to the development of more accurate and effective tools for assessing and improving the readability of texts.

2 Methodology

In this section, we present the methodology we follow in conducting our experiments. We introduce the dataset and phonetic alphabet of our choice, and elaborate on details regarding our lenses. For each lens, we offer a description, explain its potential relevance to readability, and outline the analytical methods we use to examine its relationship with readability.

2.1 Dataset

We use the WeeBit corpus to conduct our empirical explorations. The WeeBit corpus is a collection of texts at five reading levels covering ages 7-16. Table 1 shows the associated age range and sample size of each level of the corpus. The WeeBit corpus was the most appropriate corpus for our explorations thanks to its large sample size compared to other English readability corpora, as well as its classification of text into a discrete rather than continuous readability scale. We have opted to use only the first 1000 samples from Level 5 to keep the sample size similar across levels. The text files in the corpus initially contained copyright details and technical data at the end of the samples. For accuracy, these unrelated sections have been removed from the samples.

Level	WeeBit Class	Ages	Total Texts
1	Level 2	7-8	807
2	Level 3	8-9	789
3	Level 4	9-10	629
4	KS3	11-14	646
5	GCSE	14-16	7530

Table 1: Sample size of each level in the WeeBit Corpus

2.2 Phonetic Alphabet

Phonemes are defined as the smallest unit of sound that distinguishes one word from another. Ultimately, however, they are approximations of sounds produced during speech. Therefore different phonetic alphabets exist, each offering different nuances in the representation of speech units. For the purposes of this project we have chosen to use the International Phonetic Alphabet (IPA) as it is a standardized representation of phonemes. We have two additional reasons for this: The first is to ensure that our research can be extended to different languages in the future by using an international representation format. The second is that we have chosen to use the CMU Pronouncing Dictionary to transform words into their phonetic representations. This dictionary makes use of the ARPAbet which directly maps to IPA. For the mapping between IPA and ARPAbet notations of phonemes, see Table 2 in Appendix A.

2.3 Individual Phonemes: Unit Frequencies

We begin our exploration at the most basic level of phonemic structure: individual phonemes. At this level, we calculate the normalized frequencies of each phoneme within a text. Normalization involves dividing the raw frequency of each phoneme by the total number of phonemes in the text, producing a measure that reflects the proportion of each phoneme within the text. We then analyze the correlation between the frequency of each phoneme and readability levels. With this analysis, we aim to explore if certain phonemes are more prevalent in texts with different readability levels. If such trends are found, it may imply that the choice of phonemes could potentially influence the readability of a text. However, further studies would be necessary to confirm this hypothesis and understand the implications fully.

2.4 Grouped Phonemes: Collective Phonemic Trends

While the analysis of individual phonemes provides essential initial insights, it primarily captures micro-scale relationships and cannot fully represent the combined influence of phonemic groups. To uncover larger-scale trends that may be indiscernible at the level of individual phonemes, we expand our investigation to a wider lens: Grouped phonemes. At this level, we employ two different strategies to group phonemes.

Age of acquisition. The first grouping is based on the age at which English-speaking children typically acquire each phoneme. A literature review of children's English consonant acquisition in the United States by Crowe and McLeod reports that "the consonants /b, n, m, p, h, w, d/ were acquired by 2;0–2;11; /, k, f, t, , j/ were acquired by 3;0–3;11; /v, , s, , l, , z/ were acquired by 4;0–4;11; /, , / were acquired by 5;0–5;11; and // was acquired by 6;0–6;11 (ordered by mean age of acquisition, 90% criterion)" [4]. Even though all consonants are acquired before the age 7-8 (the lowest readability level in our dataset), children could have more familiarity with or mastery over the consonants acquired earlier, which could mean that texts of lower complexity might rely more heavily on these early-acquired phonemes.

Manner of Articulation and Placement. Our second grouping strategy classifies phonemes by their manner of articulation, which describes how the speech organs interact to produce each sound. Crowe and McLeod's study indicates that certain manners of articulation are mastered earlier than others, reporting that "on average, all plosives, nasals, and glides were acquired by 3;11; all affricates were acquired by 4;11; all liquids were acquired by 5;11; and all fricatives were acquired by 6;11 (90% criterion).". In this grouping, consonants belong to one of 6 main categories based on their manner of articulation: plosives (also known as stop consonants), nasals, fricatives, affricates, glides and liquids [10]. Vowels, on the other hand, are grouped as front, central, and back, also by their manner of articulation.

See Table 2 in Appendix A for the grouping of phonemes in the International Phonetic Alphabet based on their manner

of articulation.

For both of the grouping strategies, the process by which we analyze and interpret the results are the same. We calculate the normalized frequencies of each group by summing up the normalized frequencies of their constituent phonemes. We then examine the changes in the frequencies of each group with respect to readability. In a similar manner to the first lens, the analysis at this level aims to investigate whether the relative prevalence of these phonemic groups correlates with readability levels, potentially serving as an additional signal for text complexity.

2.5 Phoneme-based Features: Derived Phonemic Properties

The first two lenses we employed mainly considered the direct, observable properties of phonemes: Respectively, their individual occurrences and their occurrences in specific groups. Our last lens concerns higher-level or derived properties that are not directly observable simply by examining the phonemes themselves. They are abstract features that require some calculation or inference based on the phonemes, sometimes also relating them to other features of the text.

Whereas our exploration of individual and grouped phonemes dealt primarily with direct occurrences, our final lens zooms out to consider the more abstract, derived properties of phonemes. These phoneme-based features can't be directly inferred from the phonemes themselves. Instead, they are computed by applying additional analysis or inference to the phonemes, often in relation to other linguistic features of the text. Through this lens, we aim to uncover more complex, perhaps less obvious, ways that phonemes might influence text complexity.

For each of the features that are proposed, we first compute it for each sample, and then examine how these measures vary across readability levels. This is followed by an analysis of the correlation of that features to readability.

Grapheme-Phoneme Cohesion. Grapheme-Phoneme Cohesion (GPC) refers to the degree of correspondence between the spelling and pronunciation of words. The former concerns the graphemes (letters) of words, whereas the latter concerns the phonemes (sounds) of words.

The investigation of GPC in the context of readability is motivated by the assumption that a stronger grapheme-phoneme correspondence can facilitate easier word recognition and reading comprehension. For instance, it has been hypothesized that words with a higher GPC will take less time for dyslexics to comprehend [14].

While it is difficult to capture the essence of this feature numerically, researchers have previously estimated the GPC of a word *w* by the following formula [14]:

$$GPC(w) = phonemes(w)/letters(w)$$

Phonemic Diversity. Phonemic diversity refers to the variety of different phonemes present in a text.

Phonemic diversity is an important feature to consider in readability research as it reflects the richness and complexity of the phonemic composition of a text. Since higher phonemic diversity suggests that a greater variety of phonemes are being used, it could also affect the complexity of the text, impacting the ease of understanding for readers.

We can measure phonemic diversity by calculating the number of unique phonemes in a text and dividing this value by the total count of phonemes in that text. The phonetic diversity of a text sample t is then given by:

$$Phon.Div(t) = uniquePhonemes(t)/phonemes(t)$$

Throughout this chapter, we have proposed a multi-dimensional exploration into phonemic influences on readability. By introducing three distinct lenses—individual phonemes, grouped phonemes, and phoneme-based features—we have outlined a robust methodology to examine the phonemic complexity of a text and how it could potentially impact readability.

3 Results

In this section, we present the outcomes of the data experiments carried out as part of the explorations proposed in the preceding chapter. The presentation of the results follow the three lens structure. Although we interpret the meaning of the results in the following subsections, we discuss the implications of these findings to our research question subsequently in Section 4 (Discussion).

3.1 Individual Phonemes

Our first lens concerned the examination of links between individual phonemes and readability. The aim of this lens was to discern if certain phonemes are more prevalent in texts with different readability levels. To that end, we examine the correlation matrix in Figure 1, which shows the correlation of each phoneme to the five readability levels. Here, we report our findings from this analysis, with focus on phonemes that show significant patterns or correlations with readability levels. In the paragraphs below, we explain how to interpret the patterns exhibited by our results.

A positive correlation between an arbitrary phoneme p and an arbitrary readability level l implies that phoneme p occurs more frequently in level l than the others. A negative correlation between p and l implies that p occurs less frequently in level l than the others. The magnitude of the correlation indicates the strength of the relationship between the phoneme and the readability level, with a larger absolute correlation indicating a more pronounced variation in the phoneme's frequency across readability levels.

Phonemes which correlate more to lower readability levels and correlate less to higher levels are said to exhibit a decreasing trend in correlation. Such phonemes are of interest, as this means they are more prevalent in the lower levels and less prevalent in the higher ones. On the other hand, phonemes which correlate less to lower levels and more to higher levels are said to exhibit an increasing trend in correlation. Such phonemes are less prevalent in lower levels and more prevalent in higher levels. The magnitude of change in correlation across different readability levels also indicates the strength of the relationship. A larger change in correlation suggests a stronger differentiation in phoneme frequency between lower and higher readability levels.

In the subsequent subsubsections, we present our findings from this analysis, highlighting phonemes that demonstrate notable patterns or correlations with readability levels as seen in Figure 1. Given the extensive range of phonemes under consideration, we've organized their analysis into two categories: vowels and consonants, to facilitate a more digestible presentation of the results. The phonemes are presented with their ARPAbet representation. See table 2 in Appendix A for their equivalent representation in IPA.

Vowels

Out of the twelve vowel phonemes investigated, five showed noteworthy correlations with readability levels. These are:

- The 'IH' phoneme displayed an increasing pattern with correlations as readability level increased. Initially, it showed a negative correlation with readability levels, going from -0.18 at level 1, to -0.09 at level 2 and -0.05 at level 3. The correlations turned positive at level 4 and 5 with correlations of 0.01 and 0.27 respectively.
- In contrast, the 'AW' phoneme exhibited a decreasing pattern with weaker correlations, starting with 0.13 at level 1, ending with a negative correlation of -0.17 at level 5.
- The 'AE' phoneme showed a positive correlation of 0.17 with level 4, but then showed a negative correlation of -0.16 with level 5.
- The 'OW' phoneme showed positive correlations of 0.08 and 0.10 at levels 1 and 2 respectively, but showed a negative correlation of -0.18 at level 5.
- The 'UW' phoneme showed a positive correlation of 0.22 with level 4, even though its correlations to previous levels were negative.

Consonants

Out of the sixteen consonant phonemes investigated, twelve exhibited interesting patterns, though at varying levels of importance. We saw six consonants showing an increasing trend and six exhibiting a decreasing trend.

The consonants that displayed an increasing correlation trend from lower to higher readability levels were:

- The 'D' phoneme's correlation increased from -0.10 at lower readability levels to 0.15 at higher levels.
- The 'JH' phoneme transitioned from a correlation of -0.16 at lower levels to 0.12 at higher levels.
- The 'K' phoneme showed a correlation pattern transitioning from -0.08 at lower levels (1, 2, and 3 being negative) to 0.16 at higher levels (4 and 5 being positive).
- The 'V' phoneme's correlation pattern progressed from -0.14 to 0.13 as readability levels increased.
- The 'Y' phoneme demonstrated a similar pattern, with a negative correlation at levels 1, 2, and 3 (-0.14), transitioning to a positive correlation at levels 4 and 5 (0.21).
- The 'ZH' phoneme exhibited a pattern of -0.15 at levels 1 and 2, -0.14 at level 3, increasing significantly to 0.41 at level 5.

Conversely, the consonants showing a decreasing correlation trend with increasing readability levels included:

- The 'HH' phoneme showed a positive correlation up to level 4 (0.06 at level 4), which turned strongly negative (-0.30) at level 5.
- The 'L' phoneme showed a positive correlation for readability levels 1, 2, and 3, which then became negative for levels 4 and 5, decreasing to -0.16.
- The 'W' phoneme's correlation decreased from 0.15 to -0.11 with increasing readability.
- The 'Z' phoneme showed a decrease in correlation from 0.16 to -0.12 with increasing readability levels.
- The 'DH' phoneme also exhibited a less consistent pattern, with correlation decreasing from 0.15 to -0.28.
- The 'CH' phoneme showed a less consistent pattern, with correlation decreasing from 0.17 to -0.03 as readability increased.

Summary of Individual Phonemes

While many phonemes didn't show meaningful patterns, others showed clear transitions from positive to negative correlations as readability levels increased (or vice versa). It is important to note, however, that despite there being noticeable trends in correlation, these values were still quite low, usually staying between the range of -0.20 and 0.20.

3.2 Grouped Phonemes

Our second lens involved grouping phonemes according to two different methods of grouping: Age of acquisition and manner of pronunciation. This section presents the frequencies of these two categories across readability levels, as well as their correlations.

Age of Acquisition

As seen in Figure 2, when grouped by age of acquisition, the frequency of the groups remained relatively constant. In addition, as seen in Figure 3, the correlation between different age groups and readability levels presented a variety of patterns, which did not seem to add up to visibly meaningful trends. Consonants acquired at age 2 remained at the same frequency throughout the levels. Their correlations to every readability level were weak. Consonants acquired at age 3 showed a negative correlation with the first 3 readability levels, but showed a positive correlation with the last two. Consonants acquired at age 4 exhibited an opposite pattern. Consonants acquired at age 5 also followed the same pattern, but with noticeably weaker correlations. The frequency of the consonants acquired at age 6 was mixed across the levels. The fluctuation of the frequency could be due to the fact that this final group contained only a single consonant. What is interesting is the quantity of each age of acquisition's relative frequency: consonants acquired at age 2 on average make up 21

Manner of Pronunciation

Figures 4 shows the frequency of vowels and consonants grouped by their manner of pronunciation. The frequency of most groups fluctuated around 0.01 percent, while others

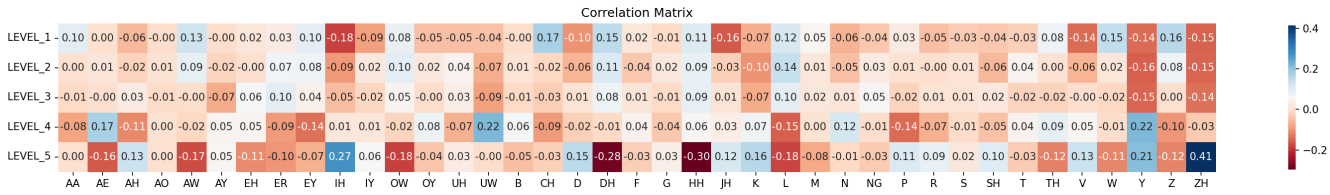
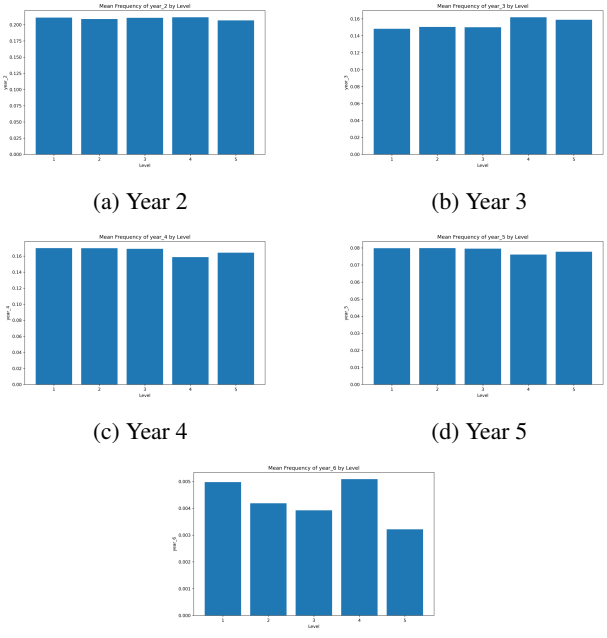


Figure 1: Correlation of individual phonemes to readability levels



(a) Year 2

(b) Year 3

(c) Year 4

(d) Year 5

(e) Year 6

Figure 2: Box plot of age of acquisition group frequencies against readability level

exhibited even less change. Front vowels and plosive consonants exhibited a steady increase in correlation, whereas fricatives showed a steady decrease. The changes in frequency exhibited by these groups imply that they do not serve as good indicators of text complexity across readability levels.

Grouped Phonemes Summary

Upon examining phonemes in grouped formats, we discerned patterns linked to age of acquisition and manner of pronunciation. Most notably, consonants acquired at ages 3 and 4 displayed interesting inverse correlation patterns across readability levels. Additionally, front vowels and plosive consonants increased consistently in correlation, while fricatives demonstrated a decreasing trend. The usefulness of these results are to be discussed further in the discussion as the numeric quantities associated with these trends suggest little significance.

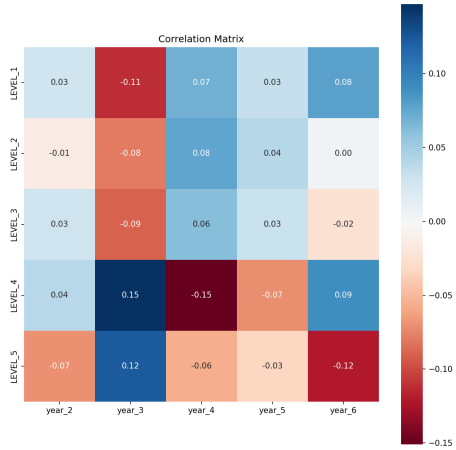


Figure 3: Correlation of age of acquisition groups to readability levels

3.3 Phoneme-based Features

GPC

Figure 6 shows that average GPC increases marginally with readability, and the correlations across levels in figure corr show an increasing trend. However, the increase is minute, and this feature's value varies moderately across samples in a given readability level. Therefore, GPC could not be a strong indicator of text complexity.

Phoneme Diversity

Figure 7 shows that average phoneme diversity fluctuates, not exhibiting a pattern of constant increase or decrease with respect to readability. Furthermore, the feature shows high variance, as demonstrated by the error bars which point to the minimum and maximum value for that feature in the samples belonging to that level. As seen in figure 8, the feature's correlation to levels show neither a clearly increasing or decreasing trend. These results suggest that this feature would serve very poorly as an indicator of readability.

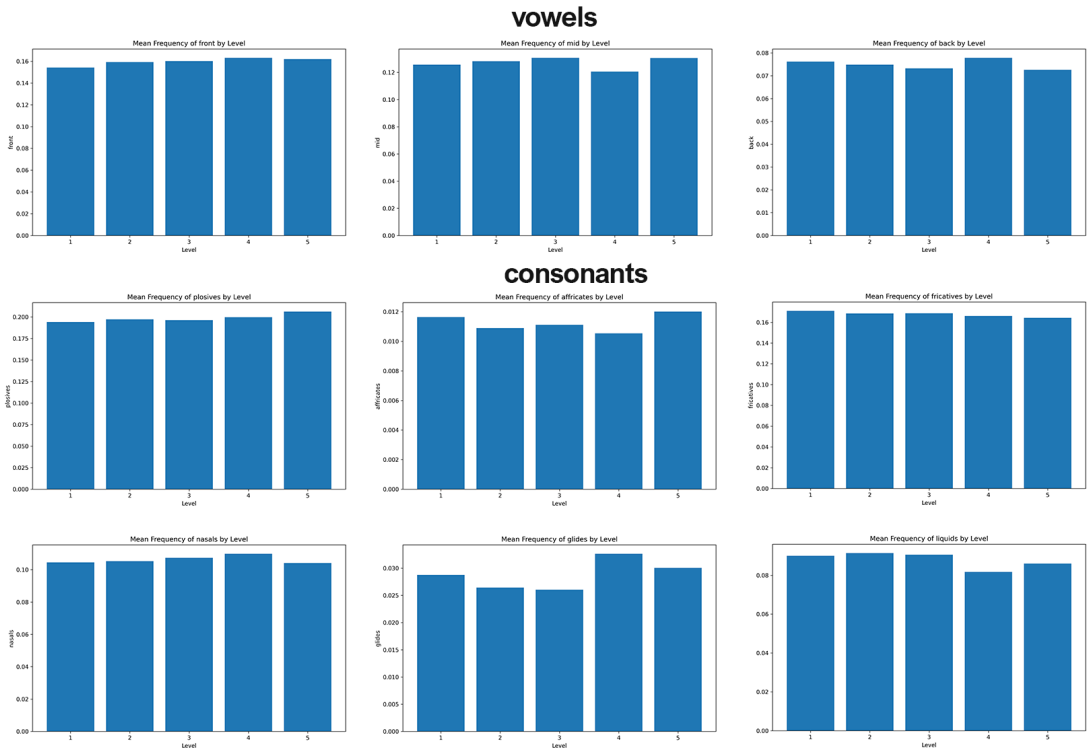


Figure 4: Box plot of pronunciation group frequencies against readability level.

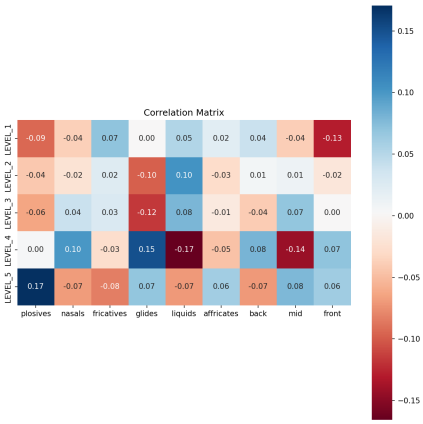


Figure 5: Correlations of phonemes grouped by manner of pronunciation to readability levels

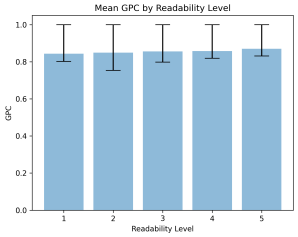


Figure 6: Box plot of GPC frequency against readability level

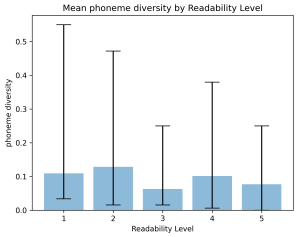


Figure 7: Box plot of phoneme diversity against readability level

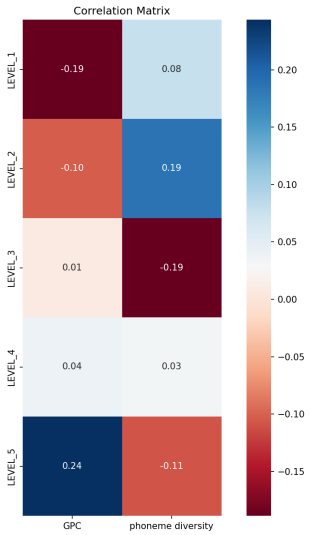


Figure 8: Correlation of phoneme-based features to readability levels

4 Discussion and Limitations

In this section, we discuss our findings and share the limitations of our research.

4.1 Interpretation of Results

Individual Phonemes

Multiple phonemes showed a correlation with lower or higher readability levels as evident in the patterns of increasing or decreasing correlation in Figure 1. Despite the appearance of certain patterns, the magnitude of the correlation levels typically fall between 0 and 0.25. This relatively low correlation range should not be surprising as the readability of a text is affected by many variables. More abstract elements such as vocabulary choice, can be expected to contribute much more significantly to the perception of the readability level of a text.

Grouped Phonemes

When the frequencies of each age group's consonants were compared, the ones that were learnt first appeared more frequently in the text. Although this is an interesting finding, the difference between groups was constant across readability levels, which means it is not beneficial for readability estimation. The analysis showed that the frequency of these groups do not change noticeably across readability levels. This suggests that even if children struggle more with certain phonemes compared to others, this pattern doesn't change as their age increases. Therefore, the frequency of phonemes grouped by their age of acquisition was not a good indicator of text complexity.

Similarly, when we grouped phonemes by manner of their pronunciation, the changes in the frequency of each group was insignificant. This implies that the frequency of these groups do not serve as a good indicator of text complexity either.

Phoneme-based Features

Mean GPC increased marginally across readability levels. This suggests that the average ratio of phonemes to graphemes is relatively constant in texts targeting ages 7-16, regardless of their readability level. On the other hand, the mean phoneme diversity of each text sample showed a high amount of variation in each readability level. Due to this, it could not be used to examine a text's relationship to readability. These result indicates that future work should look into identifying different phoneme-derived features that could prove more useful than phoneme diversity and GPC in readability estimation.

Summary

Our findings suggest that while certain individual phonemes displayed varying degrees of correlation with different readability levels, the magnitude of these correlations were generally quite low, typically ranging between -0.25 and 0.25. Phonemes grouped by age of acquisition or manner of pronunciation also failed to show substantial changes in frequency across readability levels. Lastly, the phoneme-derived features of GPC and Phoneme Diversity were not found to be strong indicators of readability due to their high variance across different levels. These findings indicate that phonemes, in their individual form, grouped by age of acquisition or manner of pronunciation, or even when used to derive more abstract features, do not provide a strong or consistent signal for text complexity. Therefore, in the context of our research question, the evidence suggests that phonemes cannot serve as indicators of the complexity of a text. They cannot independently act as reliable indicators of text complexity, given their limited correlations and the significant variation observed across different readability levels.

4.2 Limitations

Dataset Limitations

The applicability of the findings hinges on the dataset's accuracy and representation of the readability levels. Although the WeeBit corpus is a collection of articles that target different age groups, the assignment of readability levels of articles are motivated by adults' beliefs regarding of how children perceive readability. Therefore, the classifications may not necessarily reflect children's perception of readability. Thus, the real ease or difficulty children experience when reading certain phonemes or phoneme groups may not align with this data.

Furthermore, an obvious limitation of using a single corpus is the validity of results. Since these experiments have been conducted on a single corpus, we cannot say whether the same results would be reproduced had a different corpus been used. Although the exploration of additional suitable corpora could validate our findings further, the availability of such corpora is also a limitation in and of itself.

Linguistic Limitations

As mentioned in Experimental Setup, there were multiple phonetic alphabets that we could've chosen to conduct our explorations in. We have chosen the International Phonetic Alphabet to ensure research on this topic could be extended to other languages, and because phonetic resources that enabled our research were available in this alphabet. It should also be noted that pronunciation of words changes between the many dialects of English. We have taken CMUDict's pronunciation of words as our ground truth, but the existence of accents limits the universality of our results.

Project Scope

The timing of the project limited the potential to experiment with different methodologies. With more time, further literature review could, for instance, uncover alternative ways to investigate the phoneme-related features of texts.

5 Responsible Research

Reproducibility is an integral part of responsible research, as the verifiability of our results hinges on other researchers' ability to reproduce them. Before processing the samples in the corpus, we cut copyright-related and technical information from each file. These sentences can be found under Appendix B. Additionally, only the first 1000 samples have been used from the last reading level in order to keep sample sizes similar across levels. Besides these points, our methodology is described clearly enough such that anyone who wants to

reproduce the results can do so, provided they have access to the WeeBit corpus.

Furthermore, as per the conditions under which the WeeBit corpus was shared with us, we are only allowed to share the results of our research, and not the contents of the corpus itself. We have made sure to comply with these rules by not uploading the resources to openly accessible cloud services during the process of computing our results. Lastly, our research conclusions could potentially exhibit bias if the WeeBit corpus is found not to accurately represent the various readability levels of children's texts.

6 Conclusions and Future Work

The aim of this research project was to explore the relationship between readability and phonemes. We set out to conduct this research based on the premise that phonemes could potentially impact the complexity of reading materials, and the observation that past scientific literature had only explored this relationship to a limited extent.

Our research question was: "Can phonemes serve as indicators of the level of complexity of English texts?". Our investigation involved three lenses, examining phonemes' relationship to readability through individual phonemes, phonemes grouped by age of acquisition and manner of pronunciation, and two phoneme-based features: Grapheme-phoneme cohesion and phonemic diversity.

The results show that the utility of phonemes as standalone features in readability estimation is limited. This emphasizes the fact that readability cannot be defined solely by the use of surface-level features. If phonemes are to be used as indicators of text complexity, they should be used in conjunction with additional, higher-level linguistic features.

Nevertheless, this research contributes to the understanding of the relationship between phonemes and readability. As an initial exploration into the intersection of these fields, it opens avenues for further research that can build upon our findings to either refine existing readability estimation models, or deepen our understanding of the relationship. Further research should explore a wider collection of phoneme-derived features, make use of multiple or more diverse corpora, and relate phonemes to other linguistic factors, exploring their combined influence on readability.

7 Acknowledgements

This work could not have been possible without the guiding comments and advice of our responsible professor, Maria Soledad Pera.

References

[1] Readability.

[2] E. B. Coleman. Experimental studies of readability part i. *Elementary English*, 45(2):166–178, 1968.

[3] Scott A. Crossley, Stephen Skalicky, and Mihai Dascalu. Moving beyond classic readability formulas: new methods and new models. *Journal of Research in Reading*, 42(3-4):541–561, 2019.

[4] Kathryn Crowe and Sharynne McLeod. Children's english consonant acquisition in the united states: A review. *American Journal of Speech-Language Pathology*, 29(4):2155–2169, 2020.

[5] Félix Desmeules-Trudel and Tania S. Zamuner. Spoken word recognition in a second language: The importance of phonetic details. *Second Language Research*, 39(2):333–362, 2023.

[6] William H. DuBay. The principles of readability., Aug 2004.

[7] Linda B. Gambrell, Robert Mills Wilson, and Walter N. Gantt. Classroom observations of task-attending behaviors of good and poor readers. *Journal of Educational Research*, 74:400–404, 1981.

[8] Fernand Gobet. Vocabulary acquisition. In James D. Wright, editor, *International Encyclopedia of the Social Behavioral Sciences (Second Edition)*, pages 226–231. Elsevier, Oxford, second edition edition, 2015.

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

[10] Juergen Luettin. Visual speech and speaker recognition. 04 2003.

[11] Ion Madrazo. Towards multipurpose readability assessment. Thesis, Boise State University, 12 2016.

[12] Vera Paola E. Reyes. 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, page 284–288, New York, NY, USA, 2019. Association for Computing Machinery.

[13] Neena M. Saha, Laurie Cutting, Stephanie Del Tufo, and Stephen Bailey. Initial validation of a measure of decoding difficulty as a unique predictor of miscues and passage reading fluency. *Reading and writing*, 34:497–527, Feb 2021.

[14] Laurianne Sitbon and Patrice Bellot. A readability measure for an information retrieval process adapted to dyslexics. In *Second international workshop on Adaptive Information Retrieval (AIR 2008 in conjunction with IIiX 2008)*, pages 52–57, 2008.

[15] Sowmya V. and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. pages 163—-173, 01 2012.

Appendix A Grouping of Phonemes by Manner of Pronunciation and Mapping between IPA and ARPAbet Symbols

		IPA Symbol	ARPAbet (SV)	ARPAbet (UV)	Examples
Vowels	Front	i	i	IY	beat
		ɪ	I	IH	bit
		e	e	EY	bait
		ɛ	E	EH	bet
		æ	@	AE	bat
	Back	ɑ	a	AA	Bob
		ɔ	c	AO	bought
		o	o	OW	boat
		ʊ	U	UH	book
		u	u	UW	boot
	Mid	ɝ	R	ER	bird
		ə	x	AX	ago
		ʌ	A	AH	but
Diphthongs		ɑɪ	Y	AY	buy
		ɑʊ	W	AW	down
		ɔɪ	O	OY	boy
		ɨ	X	IX	roses
Stop Consonants	Voiced	b	b	B	bat
		d	d	D	deep
		g	g	G	go
	Unvoiced	p	p	P	pea
		t	t	T	tea
		k	k	K	kick
Fricatives	Voiced	v	v	V	vice
		ð	D	DH	then
		z	z	Z	zebra
		ʒ	Z	ZH	measure
	Unvoiced	f	f	F	five
		θ	T	TH	thing
		s	s	S	so
		ʃ	S	SH	show
Semivowels	Liquids	l	l	L	love
		l̩	L	EL	cattle
		r	r	R	race
	Glides	w	w	W	want
		ʍ	H	WH	when
		j	y	Y	yard
Nasal	Non vocalic	m	m	M	mom
		n	n	N	noon
		ŋ	G	NX	sing
	Vocalic	m̩	M	EM	some
		n̩	N	EN	son
Affricates		tʃ	C	CH	church
		dʒ	J	JH	just
Others	Whisper	h	h	HH	help
	Vocalic	ɾ	F	DX	batter
	Glottal stop	ʔ	Q	Q	

Table 2: Grouping of Phonemes by Manner of Pronunciation and Mapping between IPA and ARPAbet Symbols [10]

Appendix B Preprocessing on Text Samples

For levels 1-3, the following sentence was cut from the end of each file: ”All trademarks and logos are property of Weekly Reader Corporation.”

For levels 4-5, the following sentences were cut from the end of each file: ”The BBC is not responsible for the content of external internet sites. This page is best viewed in an up-to-date web browser with style sheets (CSS) enabled. While you will be able to view the content of this page in your current browser, you will not be able to get the full visual experience. Please consider upgrading your browser software or enabling style sheets (CSS) if you are able to do so.”