

Investigating the Use of Phonemes as Readability Signals

Mark Musa Mitrani Sabah
M.M.MitraniSabah@student.tudelft.nl
Responsible Professor: M.S. Pera

Background

One of the less explored areas in readability assessment is the role of phonemes, the smallest units of sound that can distinguish one word from another. Phonemes encode phonetic/auditory details about words that letters do not capture. Research has suggested that restricting phonemes can influence readability, and according to research, children typically acquire phonemes between ages 2-5, with certain groups of phonemes consistently being learned earlier than others. **Despite these insights, there has been limited empirical investigation into the relationship between phonemes and readability. This research aims to bridge this gap.**

i:	ɪ	ʊ	u:	ɪə	eɪ	e	ə	ɜ:	ɔ:	ʊə
see	sit	book	too	here	day	men	about	word	sort	tour
ɔɪ	əʊ	æ	ʌ	ɑ:	ɒ	eə	aɪ	aʊ	p	b
boy	go	cat	but	part	not	wear	my	how	pig	bed
t	d	tʃ	dʒ	k	g	f	v	θ	ð	s
time	do	church	judge	kilo	go	five	very	think	the	six
z	ʃ	ʒ	m	n	ŋ	h	l	r	j	w
zoo	short	casual	milk	no	sing	hello	live	read	yes	we

English Phonemes

Research Question

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

Method

We conducted experiments on the **WeeBit corpus**, a collection of articles classified into 5 readability levels (age range 7-16).

Our approach made use of three lenses:

1. Individual phonemes

We converted each text sample to its phonemes, computed the normalized frequency of each phoneme, and identified the strongest correlations between phonemes and readability levels.

2. Grouped phonemes

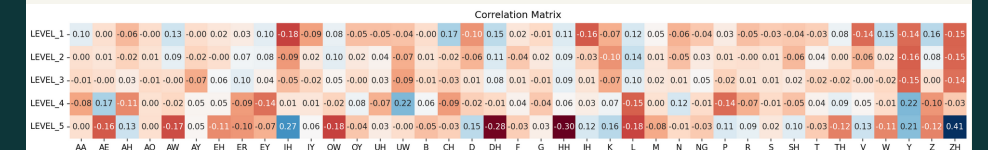
We grouped phonemes in text samples based on two classifications: Age of acquisition, manner of pronunciation; then investigated the frequencies of each group across readability levels for meaningful trends.

3. Phoneme-based features

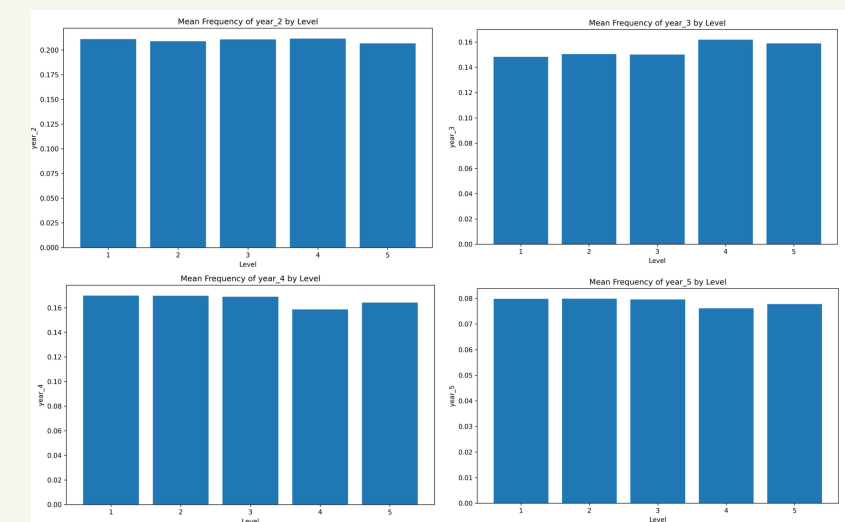
We calculated grapheme-phoneme cohesion (approximate measure of how much a word's spelling matches its pronunciation) and phoneme diversity for each sample, then performed correlation analysis between these features and readability levels.

Results

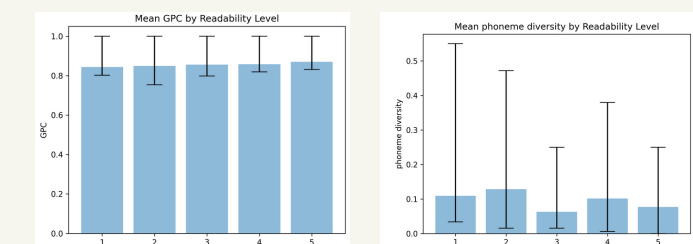
1. Individual Phonemes



2. Grouped Phonemes (by Age of Acquisition)



3. Phoneme-based Features



Conclusion

Utility of phonemes as standalone features for readability assessment is limited. Out of the three lenses, individual phonemes proved most indicative of text complexity, despite weak correlations. **Phonemes should be paired with other linguistic features if they are to be used in readability assessment.**

References:

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.
E. B. Coleman. Experimental studies of readability part I. *Elementary English*, 45(2):166-178, 1968.
English Phonemes: <https://englishlive.ef.com/blog/study-tips/phonemic-chart-will-help-english/>