Syntactic and Lexical Approaches to Reading Comprehension

Henry Lin
esotericbubba@hotmail.com
Hackley School, Tarrytown, New York

Abstract

Teaching reading comprehension in K – 12 faces a number of challenges. Among them are identifying the portions of a text that are difficult for a student, comprehending major critical ideas, and understanding context-dependent polysemous words. We present a simple, unsupervised but robust and accurate syntactic method for achieving the first objective and a modified hierarchical lexical method for the second objective. Focusing on pinpointing troublesome sentences instead of the overall readability and on concepts central to a reading, we believe these methods will greatly facilitate efforts to help students improve reading skills.

1 Introduction

Teaching reading comprehension and readability research are related but also different. Readability research generally focuses on ranking the difficult level of a passage while reading comprehension education more directly aims at helping students read better.

Although readability metrics offer a good indication of a passage's difficulty level, a more useful approach for teaching comprehension is to pick out those difficult sentences for specific, targeted learning. Although vocabulary is an important factor in making a sentence difficult, it also often happens that a sentence, either with no unknown words or after all the words have been looked up, is still difficult to understand. The following is an example from a 6th grade history reading:

"*Nor have legitimate grounds ever failed a prince who wished to show colorable excuse for the non-fulfillment of his promise.*"[1]

Even though the main idea was more or less clear, sentences like this were, in general, difficult for 6th graders.

Sufficient background and vocabulary are two prerequisites of reading success, but beyond these two, what textual features are there that make a sentence hard? This is one question this paper addresses. The second question is how to help students understand all major critical ideas in a reading because in a passage, in addition to the main idea, there are major supporting details that are crucial to comprehension. For example, in Martin Luther King Jr.'s *Beyond Vietnam* speech, the main idea is to oppose the war in Vietnam and there are four major reasons given. Understanding these four reasons is as integral to the passage's comprehension as the main idea. The third question we address is how to help students understand in-context polysemous words. Together, this paper makes the following contributions:

- A set of simple and accurate statistics that identifies, within a passage, the sentences that are challenging.
- A set of interesting findings about the standardized reading tests.
- A modified hierarchical lexical clustering method to find critical concepts in a reading.
- A word2vec application for selecting in-context meaning of a word.

2 Previous Work

One focus of the previous NLP work on accessing text difficulties is readability ranking. For example, Lexile (Lennon, 2004), Flesch-Kincaid (Kincaid, 1975), Dale-Chall (Dale, 1948), Coleman-Liau (Coleman, 1975), and SMOG (McLaughlin, 1969) largely rely on words and sentence length. Since one or two long sentences or difficult words do not necessarily make a passage difficult, those systems give rankings for an entire passage or a

[1] Niccolo Machiavelli, *The Prince*, Chapter XVII.

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 11–19
Melbourne, Australia, July 19, 2018. ©2018 Association for Computational Linguistics

book and are not aimed at pinpointing difficult sentences.

Recently, Pitler et. al. (2008), Peterson et. al. (2009), Kate et. al. (2010), Feng (2010), and Dascalu et. al. (2013) addressed the readability problem using supervised data and a richer set of linguistic features. However, their systems still focus on giving a readability score of the overall article, not individual sentences from which students can improve their reading comprehension. Pitler et. al. (2008) and Tanaka-Ishii et. al. (2010) also built comparators to decide relative difficulty between two sentences. Both and Tanaka-Ishii et. al. (2010) especially make heavy use of lexical features. All these models also require supervised data and vocabulary acquisition.

Works by François et. al. (2014), Siddharthan et. al. (2014), and Vajjala et. al. (2014) have focused on sentence simplification instead of sentence selection for the purpose of teaching reading comprehension. This paper provides a simple and robust method for identifying difficult sentences in a reading passage. We incorporate some of the standard features seen in previous work such as tree depth, but we also devise new features such as abstract appositives. While much of the previous research has made use of both lexical and syntactic features, our focus is on an in-depth study on syntax phenomena that contribute to sentence complexity.

In addition to individual sentences that are hard to read, scattered concepts are also challenging to a reader. An author often develops a critical idea in several paragraphs using paraphrases, synonyms, and related ideas. When a reader cannot see the relation among these words and phrases, he will have difficulty grasping that concept. For this problem, we propose a word2vec-based (Mikolov, 2013) modified hierarchical clustering model to find clusters of concepts in a reading passage.

3 The Syntactic Features

We present a set of simple and robust features able to identify the difficult sentences in a reading. We show the efficacy of these features in a series of tests on grade-level readings.

3.1 The Features

Figures 1a – 1f depict each feature in action. In the figure, each rectangular box describes what the feature is and how the feature is determined.

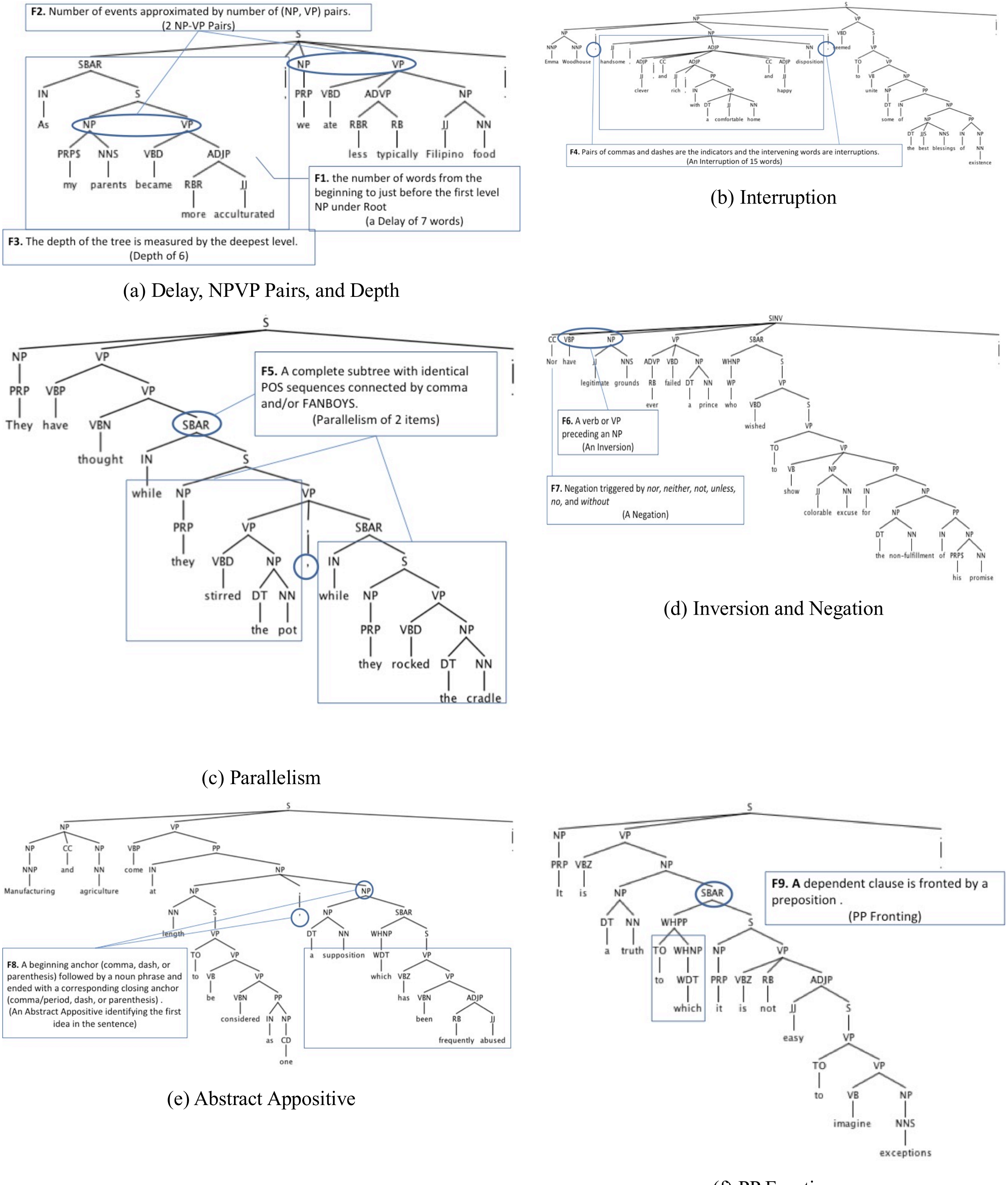
3.2 Feature Performance

Our goal is to find candidate sentences that are challenging for a young reader. This task is difficult to evaluate for two reasons: the lack of labeled data at sentence level and probably more importantly, the lack of a methodology for creating such a dataset. The creation of supervised data involves judgment from a young reader (under 16 years of age). First, young children often cannot articulate what they find difficult. Second, they sometimes think they understand a sentence while they don't. An attempt was made at a local tutoring center for children 11-16. Fifty-two children were given a grade-level passage and an above-grade passage (e.g. a hard SAT passage). They were asked to pick out the sentences they didn't understand. For both passages, more than 80% of the children either said they understood everything or they found the passage hard but couldn't tell where the difficulties were. They were then given multiple-choice questions. Fewer than 5% of the children who claimed they understood everything scored perfectly on the test. For more than 50% of the mistakes made, more than half the children claimed that it was not because they didn't understand the passage but because they were careless. This attempt showed that human judgment from a young reader is hard to obtain. Secondly, an approximation of difficulty via test performance is problematic. Perhaps, a possible approach is to convene expert reading teachers and ask them to, based on their field experiences, rank each sentence's difficulty level for each grade. This would require these teachers to have intimate knowledge of how children process sentences. For these reasons, we first evaluate the features by measuring how well they correspond to the changes in reading levels. We then use the features to rank the difficulty of each sentence and perform a qualitative assessment.

For the first part of the evaluation, we look for data that correlate well with grade levels. Representative grade-level readings are not easy to collect because readers in each grade vary greatly in their reading abilities[2]. We thus use passages in standardized tests. In this section, we present data from passages on the New York State ELA tests, which are annual tests given to students from grades 3 to 8. For high school reading data, we

[2] For example, according to Lexile, the range for 7th grade reading is 300L to 1330L, a difference between *Three Billy-Goats Gruff* (340L) and *Understanding Hume* (1290L).

use the SAT test, a national test for high school students. Thus, the data represent standard reading levels of grades 3 to high school. We first run the Stanford parser (Manning et. al.,2014). We then collect statistics of the nine features on each sentence. The data statistics and feature performance are presented in Table 1 and Figures 2a-2c. *p*-values of *t*-test at α=0.05 are shown in Tables 2a – 2c. For example, the increase in Delay from Grade 5 to Grade 6 is 95% statistically significant (p-value 0.003 < 0.05 in Table 2a). All significant changes are in bold. While the general trend is increasing


S
SBAR
IN
As
NP
PRP$
my
NNS
parents
VP
VBD
became
ADJP
RBR
more
JJ
acculturated
,
NP
PRP
we
VP
VBD
ate
ADVP
RBR
less
RB
typically
NP
JJ
Filipino
NN
food
.
Emma Woodhouse , handsome , clever , and rich , with a comfortable home and happy disposition , seemed to unite some of the best blessings of existence
They have thought while they stirred the pot , while they rocked the cradle
SINV
Nor have legitimate grounds ever failed a prince who wished to show colorable excuse for the non-fulfillment of his promise
Manufacturing and agriculture come at length to be considered as one , a supposition which has been frequently abused
It is a truth to which it is not easy to imagine exceptions

F2. Number of events approximated by number of (NP, VP) pairs. (2 NP-VP Pairs)

F1. the number of words from the beginning to just before the first level NP under Root (a Delay of 7 words)

F3. The depth of the tree is measured by the deepest level. (Depth of 6)

(a) Delay, NPVP Pairs, and Depth

F4. Pairs of commas and dashes are the indicators and the intervening words are interruptions. (An Interruption of 15 words)

(b) Interruption

F5. A complete subtree with identical POS sequences connected by comma and/or FANBOYS. (Parallelism of 2 items)

(c) Parallelism

F6. A verb or VP preceding an NP (An Inversion)

F7. Negation triggered by *nor, neither, not, unless, no,* and *without* (A Negation)

(d) Inversion and Negation

F8. A beginning anchor (comma, dash, or parenthesis) followed by a noun phrase and ended with a corresponding closing anchor (comma/period, dash, or parenthesis) . (An Abstract Appositive identifying the first idea in the sentence)

(e) Abstract Appositive

F9. A dependent clause is fronted by a preposition . (PP Fronting)

(f) PP Fronting

Figure 1. Syntactic Features

through grades, sometimes decreases are observed in two adjacent grades. Many of the decreases are statistically insignificant such as the decrease in Delay from G3 to G4 with *p*-value of 0.13.

It is noticeable that in grades 3 – 12, standard readings contain virtually none of the more specialized features of 1c-1f. These features are more prominent in older and more mature readings such as those in 19th-century literature. In section 5, we use only features in 1a and 1b.

Grade	Test Year	#Sentences	#Tokens
3	2006 – 10	975	9,967
4	2006 – 10	1,729	20,533
5	2006 – 10	1,131	14,972
6	2006 – 10	1,145	17,306
7	2006 – 10	1.296	20,256
8	2006 – 10	1,636	26,812
9+	2009,12, 16	1,397	35,415

Table 1. Data Statistics

Grade	Delay	Pair NP-VP	Depth
3→4	0.13	**1.76e-11**	**3.47e-11**
4→5	0.48	**0.035**	**1.48e-7**
5→6	**0.003**	**0.002**	**0.002**
6→7	0.38	**0.011**	**0.011**
7→8	0.59	0.68	0.68
8→9+	**2.64e-9**	**1.09e-38**	**2.26e-55**

Table 2a. *p*-values

Grade	Inversion	Parallel	Interruption
3→4	0.10	0.61	0.20
4→5	0.25	**0.015**	**0.008**
5→6	0.31	0.31	0.04
6→7	0.08	0.08	0.58
7→8	0.83	0.83	0.05
8→9+	**1.80e-14**	**1.80e-14**	**3.10e-6**

Table 2b. *p-values*

Grade	Negation	Abstract Appositive	PP Fronting
3→4	0.07	**0.008**	0.10
4→5	0.45	0.08	0.83
5→6	0.28	0.33	0.75
6→7	0.06	0.76	0.14
7→8	0.30	0.35	0.24
8→9+	**9.52e-12**	0.87	**7.68e-9**

Table 2c. *p-values*

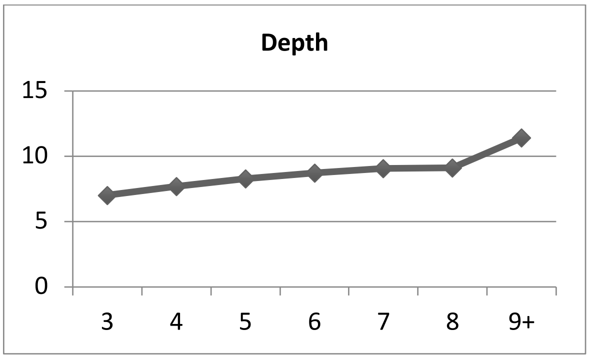


Figure 2a. Depth

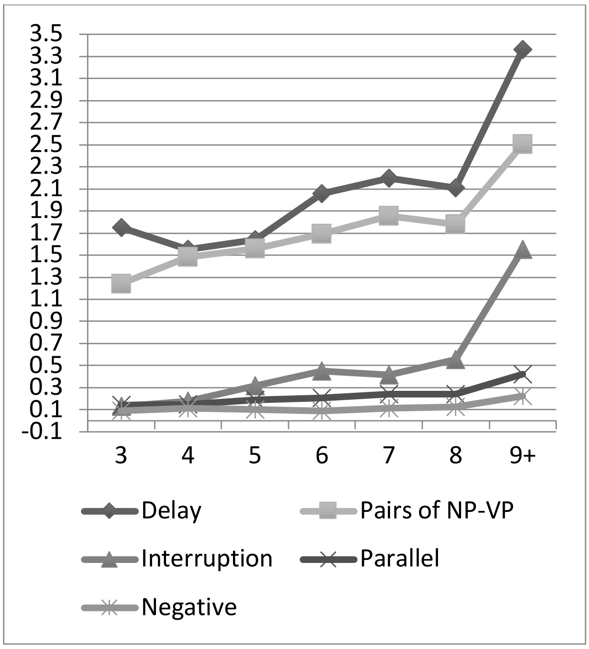


Figure 2b. Delay, NPVP, Interruption, Parallel, and Negation

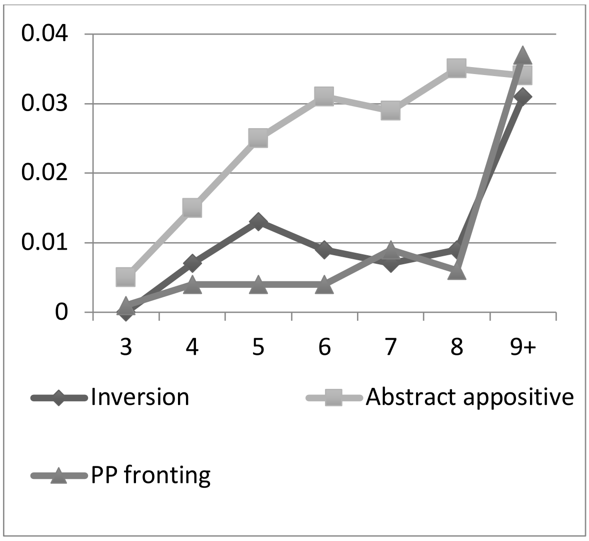


Figure 2c. Inversion, Abstract Appositive, and PP Fronting

Next we rank the sentences. Each sentence has a vector of nine feature scores. Although many different weighing schemes are possibilities, we take the simple approach of uniform weights. We compare the top-3 most difficult sentences ranked by the nine features to those ranked by sentence length and tree depth. For lower-grade texts, there is almost no difference in the order. But for more complex passages, more significant differences start to show. Through this exercise, we also find a qualitative value of the nine features. Even when the rankings by our nine features agree with the length-based rankings, we can point out more specifically what makes these sentences difficult. These specifics are shown as Notes in Table 3. We believe the ability to locate these syntax phenomena for students should be helpful in improving their reading skills.

Rank	Sentence
Top 1 by both	Deeming that a serene and unconscious contemplation of him would best beseem me, and would be most likely to quell his evil mind, I advanced with that expression countenance, and was rather congratulating myself on my success, when suddenly the knees of Trabb's boy smote together, his hair uprose, his cap fell off, he trembled violently in every limb, staggered out into the road, and crying to the populace, "Hold me!"
Notes:	Specifically, in addition to a depth of 17 levels, two long delay (underlined), and a parallel phrase (double underlined).
Top 2 by length and depth	Words cannot state the amount of aggravation and injury wreaked upon me by Trabb's boy, when, passing abreast of me, he pulled up his shirt collar, twined his side-hair, stuck an arm akimbo, and smirked extravagantly by, wriggling his elbows and body, and drawling to his attendants, "Don't know yah, don't know yah, 'pon my soul don't know yah!"
Top 2 by nine features	The disgrace attendant on his immediately afterwards taking so crowing and pursuing me across the bridge with crows, as from an exceedingly dejected fowl who had known me when I was a blacksmith, culminated the disgrace *with which* I left the town, and was, so to speak, ejected by it into the open country.
Notes:	a long interruption of 18 words (underlined), one parallel phrase ("crowing and pursuing", double underline), and one PP fronting ("with which", italicized).
Top 3 by both	One or two of the tradespeople even darted out of their shops, and went a little way down the street before me, that they might turn, as if they had forgotten something, and pass me face to face – *on which* occasions I don't know whether they or I made the worse pretence; they of doing it, or I of not seeing it.
Notes:	Specific features are PP fronting (italicized) and one parallel phrase (underlined).

Table 3. Sentence Ranking Example

4 The Lexical Approach

We now turn to finding critical ideas in a reading. Our concern is to find related and paraphrased words that contribute to the same idea.

4.1 An Example

We distinguish critical ideas from the main idea of a reading. Critical ideas are any ideas that the author develops to some extent. A crude definition is that a critical idea is an idea that the author mentions more than once. They may or may not be the main idea, but they should all contribute to the main idea. In the following short passage, there is one main idea and several critical ideas.

"*Black holes are the most efficient engines of destruction known to humanity. Their intense gravity is a one-way ticket to oblivion, and material spiraling into them can heat up to millions of degrees and glow brightly. Yet, they are not all-powerful. Even supermassive black holes are minuscule by cosmic standards. They typically account for less than one percent of their galaxy's mass. Accordingly, astronomers long assumed that supermassive holes, let alone their smaller cousins, would have little effect beyond their immediate neighborhoods. So it has come as a surprise over the past decade that black hole activity is closely intertwined with star formation occurring farther out in the galaxy.*" (SAT 2009 Practice Test)

The main idea is the last sentence of the passage, but the many critical ideas that the author develops are: "black holes", "destruction", and "intertwined with star formation".

4.2 Finding Critical Ideas

The word2vec model (Mikolov, 2013) has been a widely used statistical model for encoding word meanings. We use a modified hierarchical clustering algorithm using word2vec[3] as a representation of each word. First, cosine distances are computed on every word pair in the passage (after removing stopwords), resulting in an $n \times n$ matrix where n is the number of words. Unlike the traditional hierarchical clustering where the end result is a tree structure, our clustering is more flat and does not build a hierarchy. The linking criteria are two: (1) the distance between two words must exceed a minimum and (2) the distance between a word and an existing cluster must exceed a minimum percentage of the best pair in the cluster. The algorithm is in Figure 3.

```
Make an empty Critical Cluster list
While (1) {
    ($w_i$, $w_j$) = next best word pair in the matrix
    $score_{ij}$ = score of ($w_i$, $w_j$)
    if ($score_{ij}$ < minimum_score_threshold) {
                break;
    }
     if (neither wi nor wj is in any cluster) {
        make a new cluster ($w_i$, $w_j$);
                add to Critical Cluster list;
     }
     else if ($w_i$ or $w_j$ is in a cluster) {
        $cluster_k$ = the cluster $w_i$ or $w_j$ is in
                if ($score_{ij} \geq cluster_k$'s score) {
                    add $w_i$ or $w_j$ to $cluster_k$
     }
     else {
        $cluster_i$ = the cluster wi is in;
                $cluster_j$ = the cluster wj is in;
        if ($score_{ij} \geq cluster_i$'s score or
            $score_{ij} \geq cluster_j$'s score) {
                merge clusteri and clusterj
        }
      }
}
```

Figure 3. Word2Vec Modified Clustering

5 Applications, Experiments and Results

In addition to identifying troublesome sentences, there are many other useful things possible with these features. Interesting experiments include comparing tests across many dimensions such as across geography and across standards.

5.1 State Difference?

The National Assessment of Educational Progress, or NEAP offers reading assessments to 4th and 8th graders nationwide. In 2015, all 52 states participated. A state may score higher than another state for a variety of reasons, economic, political, etc. In this experiment, we're interested in seeing if there might be any meaningful correlation at all between a state's NAEP score and the difficulty level of its state ELA[4] tests. To this end, we select Massachusetts, the top-ranking state whose NAEP score of 235 is considerably higher than the national average of 221, and compare its state ELA passages to those of New York whose score is 223. The data comparison is shown in Table 4a. The metrics are shown in Tables 4b and 4c where *p*-values are at 95% and the bold values indicate statistical significance. Again, the more specialized feature 'Inversion' is not a significant factor in 4th and 8th grade readings[5].

Grade	Sentences	Words
NY 4th	1,729	20,533
MA 4th	1,093	16,593
NY 8th	1,636	26,812
MA 8th	908	17,594

Table 4a. NY and MA ELA Passages

Metric	NY 4th	MA 4th	*p-value*
Delay	1.551	2.083	**9.26e-5**
Interruption	0.180	0.527	**3.54e-7**
Pairs NP VP	1.484	1.765	**7.68e-11**
Depth	7.723	8.662	**1.85e-15**
Inversion	0.002	0.002	0.80

Table 4b. NY and MA 4th grade comparison

Metric	NY 8th	MA 8th	*p-value*
Delay	2.110	2.613	**0.016**
Interruption	0.557	1.116	**1.71e-6**
Pairs NP VP	1.778	2.074	**5.46e-7**
Depth	9.114	9.809	**1.26e-5**
Inversion	0.004	0.007	0.46

Table 4c. NY and MA 8th grade comparison

It's interesting to see that for both 4th and 8th grades, there is a progression of text difficulty from NY's ELA tests to MA's ELA tests. There are many reasons, both educational and non-educational, that come into play to influence one

[3] This is the Google News word2vec at https://github.com/mmihaltz/word2vec-GoogleNews-vectors

[4] English Language Arts

[5] At the time of the paper, only the 4th and 8th grade ELA from Massachusetts tests are publically available online.

state's performance. Perhaps this could be a first step in better understanding the impact of increased level of difficulty on student reading performance.

5.2 SAT or ACT?

The SAT and the ACT are standardized tests college-bound juniors and seniors take. One common section in both tests is the Reading section where students are given passages to read and multiple-choice questions to answer. Students and parents have long wondered which test is easier. A simple online search of "SAT reading vs. ACT reading" yields many comparisons. The question of which test is easier depends on many factors such as timing, question types, and so on. What this paper is concerned with is not necessarily the simple yes/no answer to the question of which test is easier, but rather with comparing the passages on each reading test. From a simple survey at a local test preparation center, students who choose ACT all report that the ACT passages are more straightforward than those on the SAT, and those who take the SAT report that some SAT passages are harder to read, specifically in genres such as pre-1900 fictions and history. This fact does not directly lead to a judgment of which test is easier, simply that the ACT passages are easier to read[6]. To test this hypothesis and to quantify how much easier or harder the reading passages differ on each test, we collect passages from both tests and run the feature analysis on them. The data information is presented in Table 5a.

Test	**Year of Test**	**Number of passages**	**Number of words**
SAT	2015 – 16 Official Practice	40	26,862
ACT	2015 – 17 Official Released Tests	40	28,752

Table 5a. SAT and ACT Passage Data

Feature	**SAT**	**ACT**	***p-value***
Delay	3.364	2.570	**0.0006**
Interruption	1.552	1.214	**0.014**
Pairs NP-VP	2.502	2.068	**2.92e-12**
Depth	11.403	10.264	**1.92e-12**
Inversion	0.009	0.008	0.728

Table 5b. SAT and ACT

Feature	**SAT**	**ACT**
Delay	2.397	1.248
Interruption	1.349	0.841
Pairs NP VP	0.893	0.425
Depth	2.179	1.490
Inversion	0.031	0.021

Table 5c. SAT and ACT Standard Deviation

The results of the analysis are shown in Table 5b. ACT passages score uniformly lower than those on the SAT with majority of the difference being statistically significant. Table 5c shows that the standard deviations of the SAT are higher, indicating that the SAT passages have more variations. The two excerpts from each test in Table 6 give a qualitative view of the phenomenon where * indicates an example of increased complexity.

ACT Humanities	In 2008, the prodigiously gifted bassist, singer, and composer Esperanza Spalding released her major-label debut. Esperanza, which she recorded as a twenty-three-year-old instructor at the Berklee College of Music.
ACT Science	Pikas, a diminutive alpine-dwelling rabbit relative. are unique among alpine mammals in that they gather up vegetation throughout summer—including flowers, grasses, leaves, evergreen needles, and even pine cones – and live off the hay pile throughout winter, rather than hibernating or moving downslope.
* SAT Humanities:	But of all relations, that between men and women, being the nearest and most intimate, and connected with the greatest number of strong emotions, was sure to be the last to throw off the old rule, and receive the new; for, in proportion to the strength of a feeling is the tenacity with which it clings to the forms and circumstances with which it has even accidentally become associated …
SAT Science	Nearly a half-century ago, Peter Higgs and a handful of other physicists were

[6] Independent of the level of the passages, the questions can still be hard. Therefore, the level of passages is but one factor among many that a student takes into account in deciding which test to take.

	trying to understand the origin of a basic physical feature: mass. You can think of mass as an object's heft or, a little more precisely, as the resistance if offers to having its motion changed.

Table 6. SAT and ACT Passage Difference Examples

5.3 Automatic Vocabulary Response

It is labor intensive to manually evaluate the efficacy of the word2vec-based lexical approach. While we annotate data for further research, we meanwhile evaluate the idea on vocabulary questions on the 8 released SAT official tests (CollegeBoard, 2009). These vocabulary questions ask the meaning of a word in the context of a given passage. The majority of the choices consist of one word each. Our baseline approach is to measure the vector cosine score between the word in question and the words in each choice. The choice with the greatest similarity score is chosen as the answer. When a choice has more than one word, we first remove the function words and then take the average of the vector scores.

We then apply a contextual word2vec model to the questions. For each word in a vocabulary question, we locate the sentence that the word occurs in and add up the vectors of all the content words in that sentence. The resultant vector is then compared to each choice in the vocabulary question. Table 7 shows that the context model outperforms baseline significantly. This experiment shows the power of combining context and a computable meaning representation such as the word2vec.

28 Vocabulary Questions from 8 official SAT tests		
Method	**Num. Correct**	**Accuracy**
Baseline	5	17.86%
Context	20	71.43%

Table 7. Word2Vec-based Vocabulary Performance

One reason the baseline performs poorly is that almost all words tested in the SAT vocabulary questions are polysemous. The word2vec is trained on mostly news data which biases the meaning of a word toward a typical news-oriented meaning. For example, the word ‘consumption’, without context, is most intuitively associated with consumer and commerce. In this question, of the five choices, “destruction”, “viewing”, “erosion”, “purchasing”, and “obsession”, the most likely context-independent choice is “purchasing” and that is what the baseline model chooses. In the given passage, however, the enclosing sentence is “According to [this thesis], television consumption leads above all to moral dangers.” After adding up all the vectors of the contextual words, the correct answer “viewing” surfaces and the context-model is able to answer that question correctly. This model makes concrete what the English teachers have meant when they instruct the students to look at the context. It also represents nicely the idea that the meaning of a word is *selected* by its surrounding words (the context).

6 Conclusion and Future Work

We present a set of straightforward and novel features to identify difficult sentences in a reading passage. In our experiments, the features correlate well with the actual grade of each text. We are also able to quantify and make more concrete of the differences between Common Core and pre-Common Core standards, and between different states. In the future, we hope to not only put all in an application for real use but also to incorporate general-purpose lexical features to further enhance reading comprehension education. Secondly, we intend to continue to investigate using word2vec as a stepping stone to distributed meaning representation. For example, extend critical ideas to multi-word phrases and tackle reading comprehension questions such as those on the SAT.

Acknowledgments

This work would not have been possible without the patient and generous help from my mentor Dr. Scott McCarley from IBM Research.

References

Coleman, Meri; and Liau, T. L. 1975. *A computer readability formula designed for machine scoring*, Journal of Applied Psychology, Vol. 60, pp. 283–284.

CollegeBoard, 2009. *The Official SAT Study Guide*. 2nd edition, Macmillan Publishing

Common Core. *Common Core State Standards* 2010 National Governors Association Center for Best Practices, Council of Chief State School Officers, Washington D.C.

Dale E, Chall J. 1948. *A Formula for Predicting Readability*. Educational Research Bulletin. 27. pp. 11–20.

Dascalu, M., P. Dessus, S. Trausan-Matu, M. Bianco, and A. Nardy. 2013. *Readerbench, an environment for analyzing text complexity and reading strategies*. In Artificial Intelligence in Education, pages 379–388. Springer

Feng, Lijun. 2010. *Automatic Readability Assessment*. Dissertation Thesis, City University of New York, NY.

François, T. et Bernhard, D. (eds.), 2014. *Recent Advances in Automatic Readability Assessment and Text Simplification.* In International Journal of Applied Linguistics (Special issue), 165:2

Kate, Rohit. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Welty, C. 2010. *Learning to predict readability using diverse linguistic features*. In the 23rd International Conference on Computational Linguistics, p. 546–554

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S. 1975. *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel.* Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.

Lennon, Colleen and Hal Burdick. 2004. *The Lexile Framework as an Approach to Reading Measurement and Success*. https://cdn.lexile.com/cms_page_media/135/The%20Lexile%20Framework%20for%20Reading.pdf

Manning, Christopher D., Mihai S., John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit.* In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55-60.

McLaughlin, G. Harry. 1969. *SMOG Grading — a New Readability Formula*. Journal of Reading. 12 (8): 639–646.

Mikolov T., Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013 *Distributed Representations of Words and Phrases and their Compositionality*. In NIPS, pages 3111–3119

Petersen, S.E., Ostendorf, M. 2009. *A machine learning approach to reading level assessment.* Computer Speech and Language, 23: 89-106.

Pitler, Emily and Ani Nenkova. 2008. *Revising Readability: A Unified Framework for Predicting Text Quality.* Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii.

Siddharthan A., Angrosh Mandya. 2014. *Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules*. In Proceedings of the 14th Conference of the European Chapter of the ACL, Gothenburg, Sweden.

Tanaka-Ishii, K., Tezuka, S., and Terada, H. 2010. *Sorting texts by readability*. Computational Linguistics, 36(2), pp. 203-227

Vajjala Sowmya, Detmar Meurers 2014. *Assessing the relative reading level of sentence pairs for text simplification*. In Proceedings of the 14th Conference of the European Chapter of the ACL, Gothenburg, Swede.