5-2013

Analysis of Features Composing an Automated Text Readability Formula.

Michael Schneider
East Tennessee State University

Analysis of Features Composing an Automated Text Readability Formula

by Michael Joseph Schneider

Honors-in-Discipline

East Tennessee State University

15 May 2013

_____	_____
Dr. Jay Jarman, Faculty Advisor	Dr. Christopher Wallace
_____	_____
Dr. Jessica Keup, Thesis Reader	Dr. Joseph Sobol, Thesis Reader

Table of Contents

Abstract

In an effort to make reading more accessible, an automated readability formula can help pair readers with material appropriate for their reading level. This study attempts to discover and analyze a set of possible features that may be included in a future automated readability formula. This set is not a definitive list of readability formula features, but rather an overview of current features being study in the Natural Language Processing field.

Note: This document was designed to be viewed as a Word document on a computer. It has embedded hyperlinks to help the reader switch between the thesis and reference material. The "def" superscript links a word to its definition, while the "note#" superscript links to an applicable note by the author.

Introduction

Reading, a skill that people require to absorb and process information, can be beyond many people's capabilities. For those with low reading abilities, the task of reading important documents such as legal or medical forms can be difficult or even impossible. To help make all information more accessible, complex texts need to be simplified.

Before texts can be simplified, they must be given a reading level. Current processes for automated readability scoring focus on different linguistic features and have varying ranges of effectiveness. The traditional standard for readability, the Flesch-Kinkaid (FK) readability formula, rates a text on its average sentence length and average number of syllables per word. Two problems arise from the assumption that longer sentences with bigger words are uniformly more difficult to read. First, word size does not account for word familiarity, i.e. word frequency. A larger word that is commonly used will be easier to read than a smaller unfamiliar word. Secondly, shorter sentences can dilute information and are not inherently easier to read.

Because of these problems, which can cause FK to unreliably score text readability, FK must be replaced by a new readability formula. Modern readability formulas account for additional semantic and syntactic features that allow for deeper analyses of a text. While different works of research promote different combinations of features, the literature suggests that multiple linguistic features will be needed to accurately score readability. Some of the commonly agreed upon features in the linguistic community are word frequency, sentence length, word choice, being able to target individual reader groups, such as adolescent readers versus adult readers, and reading genres, such as fiction, non-fiction, history, or science fiction.

While these have each shown influence over a text's reading level, it is still not known how each relates to one another. In other words, which feature is most important and under what

conditions might another feature be more important in determining the text's readability? For this reason, focus shall be on defining each feature and finding the best application of each feature, not on their interrelation.

Word Frequency

Word frequency, for instance, can be defined as how often a word appears in a language or corpus[def]. For a readability test, this means that a word with a high frequency is more easily recognized by a reader than a word with a low frequency [Brysb09]. Interpretations of word frequency depend on the frequency norm being used. For the past 40 years, the Kučera and Francis (KF) frequency norms has been the standard choice for psycholinguistic research.

The KF frequency norms, which were created in 1967, are continually cited in psychological research with approximately 215 citations as of January 2009. Table 8 shows that nearly all word frequency articles in a November 2008 issue of *Journal of Experimental Psychology: Learning, Memory, and Cognition* based their word frequencies on KF. The continuing use of KF norms is troubling because KF was shown to be flawed as early as 1998 by [Burge98]. To obtain further data on KF's performance, Brysbaert and New compared Elexicon-project-generated Lexical Decision Times (LDTs) and Elexicon Accuracy rates for over 40,000 words with the Celex, HAL, Zeno, BNC, and SUBTL frequency norms. Tables 12 and 13 show KF falling behind its peer norms on LDTs, Elexicon Accuracy, and Reaction Time (RT).

Table 12

Measure	Acc_{young}	Acc_{old}	Acc_{Elex}
KF	18.0	7.0	22.5
Spoken	16.8	5.5	23.0
Celex	24.2	10.4	26.0
HAL	24.7	8.2	31.3
Zeno	25.5	10.7	29.8
BNC	22.8	9.0	25.4
SUBTL	27.7	12.4	38.3

Note—Multiple regression analysis involved log(freq+1), $\log^2$(freq+1), and word length in number of letters. All stimuli were monosyllabic (N = 2,406). KF, Kučera and Francis (1967); BNC, British National Corpus.

One reason that KF does not effectively rate word frequency is its use of a 1.014 million word corpus, which is small by today's standards. Modern word frequency norms use corpora[def] of 16 million words (Celex), 17 million words (Zeno), 51 million words (SUBTLEX), 130+ million words (HAL), or even 350 million words (MetaMetrics). KF's small corpus size causes errors in estimating rare words, words which should have a low word frequency, by giving these words a high word frequency. The effect of corpus size can be seen in further detail with Table 9, which shows the percentage of variance on different portions of the British National Corpus[def]. As the British National Corpus grows, so does the percentage of variance, showing a clear correlation between corpus size and percentage of variance of LDTs.

Table 9

(Million Words)	R^2 (%)
0.5	48.7
1	51.3
2	53.3
4	55.1
8	55.9
16	56.4
32	56.1
88	56.1

A second problem with KF is that its norms were based on a corpus of adult books. Some frequency norms like KF use written material like books and magazines because these sources are thought to be the most important in representing visual word recognition. Unfortunately, these sources distort word frequencies by "polishing" their language so as not to sound repetitive and by focusing on subjects that are not in a person's everyday life [Brysb09]. To address these issues, groups like [Burge98] turned to Internet resources that reflect a more natural, everyday language.

[Burge98] focused on Internet discussion groups, because of the relative ease of collecting texts and their unsupervised content which is believed to more accurately represent natural language. While [Burge98] focused on Internet discussion groups, [New07] focused on television and movie subtitles. [New07] used subtitles because movies and television shows typically involve social interactions and because modern people watch television more than they read. SUBTLEX, which follows [New07]'s example, assembled 8,388 films and television episodes, which contained 51 million total words, for its corpus.

After determining the source of a corpus, the next step in forming a word frequency is choosing a frequency form. A common frequency form in American research is the simple word form (WF) frequency, or the "frequencies of the words as they appear in the corpus. For instance, there are 18,081 occurrences of the word *play* in the SUBTLEX, 1,521 of the word plays, 2,870 of the word played, and 7,515 of the word playing." [Brysb09]. Each word is given its own frequency but in lemma frequency "the sum of the frequencies of all the inflected forms of a particular noun, verb or adjective. [...] The lemma frequency of the verb *to beg* is the sum of the frequencies of its inflected forms *beg*, *begs*, *begged*, and *begged*" [Brysb09]. The theory behind lemma frequencies is that inflected forms affect one another's processing time. This proved to be false for English, as Table 14 shows little difference between WF frequency and lemma frequency. Further research is necessary to see how a more inflective language would affect lemma frequency.

Table 15

Percentages of Variance Accounted for by the Word Frequency SUBTL Index and the Contextual Diversity SUBTL Index for the Elexicon Project and the Monosyllabic Words Investigated by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004; see Table 15 and 16)

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency						
$SUBTL_{WF}$	22.0	32.9	26.4	49.2	45.2	42.5
$SUBTL_{CD}$	23.4	36.8	28.0	49.5	46.8	43.6
Frequency + word length						
$SUBTL_{WF}$	30.1	40.7	33.6	62.3	45.2	43.5
$SUBTL_{CD}$	31.3	44.0	34.9	62.9	46.8	44.6
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency						
$SUBTL_{WF}$	26.4	12.1	42.1	29.5	9.7	13.6
$SUBTL_{CD}$	29.3	13.9	44.2	31.0	9.4	13.3
Frequency + word length						
$SUBTL_{WF}$	27.7	12.5	42.3	29.6	21.1	22.8
$SUBTL_{CD}$	30.6	14.3	44.3	31.1	21.2	23.0

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming; WF, word form frequency; CD, contextual diversity.

Besides WF and lemma frequencies, Contextual Diversity (CD) can be used to determine word frequency. CD counts the number of texts in a corpus that contain a word. Tables 15, 16 and 17 show that CD outperforms WF by about 1%-3%. A possible reason for CD's greater performance is that many words appear in titles for texts and can be repeated multiple times in a text, which misrepresents a word's natural frequency. For example, a movie entitled "The Wild Sunflower" may excessively use the word *sunflower*. This could cause sunflower to have a higher frequency than it would actually have in the English language. Also some texts can simply outliers, i.e. they use a word more than it is naturally used. For example, fictional characters often have catch phrases. The catch phrases' words, from their constant repetition, could be given higher than normal word frequencies.

Three variables have been determined to affect the quality of a word frequency norm: corpus size, source material for the corpus, and the frequency form used. The optimal size for a corpus has been shown to lie between 16-30 million words. The more natural or the closer to

normal language use a source material is, the more accurate the corpus's word frequencies are. Of the three frequency forms discussed (WF, lemma, and CD), CD showed the greatest performance. All of these variables were used in creating SUBTLEX, which greatly outperformed KF frequency norms. Besides outperforming KF frequency norms, the SUBTLEX frequency norms are freely available to be used in research. This was done to help move research away from the outdated KF frequency norms and move towards modern frequency norms.

Word Frequency: Conclusion & Implementation

SUBTLEX has been proven to be accurate, but this accuracy should be confirmed by another research group to validate SUBTLEX's creators' results. Currently, there is no known study that has used SUBTLEX in its research. While SUBTLEX may require further testing, it does have a significant edge, besides accuracy, over other word frequency norms. Simply put, it's free. Funding was provided for the creation of SUBTLEX, which allows it to be freely offered at no licensing costs. Since development, validation, and implementation costs for creating a readability formula could prove to be substantial, avoiding large licensing costs from other word frequency norms by using SUBTLEX would help reduce project costs. This point does not pertain to readability, but is more of a practical point to be considered for future research. While SUBTLEX word frequency norms are not the only WF norms available to be used in a readability formula, they may prove a good choice because of their accuracy and lack of licensing costs.

No matter the word frequency norm chosen, word frequency will still need to be incorporated into the readability formula. This should be a fairly easy process by using a natural language processing suite, such as GATE[def.] (General Architecture for Text Engineering), to tokenize[def] the text into individual words and lexemes[Note1] such as "kick the bucket". The word frequency for each lexeme set can then be found in the word frequency norms. The word frequencies can then be processed in the readability formula to help determine the texts readability.

Genre - Young Readers

While Flesch-Kinkaid was discredited earlier, it was not discredited for its use of sentence length, but rather its sole reliance on sentence length and applying sentence length to all readability levels. Sentence length cannot be used for determining all readability levels. It has been found to be most effective with specific readers, chiefly young adolescent readers.

Hiebert and Pearson performed a study on how two modern readability formulas would be able to gauge adolescent reading levels. Lexile and Coh-Metrix were chosen to compare different methods for judging reading levels ranging from kindergarten to second grade (American education system). Lexile and Coh-Metrix indices differ in that Lexile uses sentence length and mean frequency of words [Def.] to determine text difficulty, while Coh-Metrix determines text difficulty with five key variables: non-narrativity [Def.], referential cohesion [Def.], situation model cohesion [Def.], syntactic simplicity [Def.], and word concreteness [Def.].

Lexile and Coh-Metrix rated the texts from two categories, reading level and corpus source. The test consisted of corpora that had been professionally rated for young readers and classified into seven specific reading levels: one for kindergarten, two - six for first grade, and

seven for second grade as well as by one of the following corpus sources: trade [7], trade instructional[8], textbook core[9], textbook ancillaries [10], and tests[11]. The results of the Lexile index were then compared with traditional text difficulty indices: Degree of Reading Power[12], Fry readability formula[13], and Spache readability formula[14], while an intercomparison was done between the Coh-Metrix variables.

Table 4
Readability Measures (Means) by Text Levels

Text Levels	DRP	Fry	Spache	Lexile	MSL[1]	MLF[1]
1	1.6	1.3	1.9	86.9	4.9	3.8
2	1.6	1.1	1.8	140.0	5.0	3.6
3	1.6	1.1	1.8	238.0	6.1	3.7
4	1.6	1.3	1.8	238.2	6.4	3.8
5	1.8	1.6	2.0	346.0	7.2	3.7
6	2.0	2.0	2.2	420.6	8.0	3.7
7	2.2	2.6	2.3	489.1	8.8	3.7

[1] MSL (mean sentence length) and MLF (mean lexical frequency) were provided as part of the Lexile analysis of the texts. Although they are not defined as readability measures, they are included here as supplementary information.

The results, in Table 4, show that mean sentence length, one of Lexile's component measurements, showed a clear progression of difficulty between the seven reading levels, while the traditional indices could not clearly distinguish between the reading levels. Mean lexical frequency, on the other hand, stayed relatively the same throughout the seven reading levels, which shows that texts had similar vocabularies. For these texts, syntactic measurements were more accurate in differentiating between the seven reading levels than semantic measurements.

The results of the Coh-Metrix indices, though not as clear as Lexile, show some recognizable patterns. Table 6 shows that some indices yielded inaccurate difficulty ratings for the seven reading levels.

Table 6
Means for Coh-Metrix Indices by Text Level

Text Levels	Non-narrativity	Referential cohesion	Syntactic complexity	Word abstractness	Situation model cohesion	Familiarity	Type/ Token
1	20.9	9.3	4.4	35.2	78.5	1.9	.6
2	18.9	10.5	10.7	34.9	79.9	2.3	.5
3	19.8	14.6	7.3	42.7	62.7	2.2	.5
4	14.6	20.5	7.7	45.8	64.7	2.1	.5
5	17.5	32.0	10.2	37.2	54.7	2.2	.5
6	19.7	39.8	12.7	37.4	52.6	2.2	.6
7	18.6	46.2	16.5	37.4	53.4	2.2	.5

Each index's results are as follows:

- **Non-narrativity -** remains relatively low for all reading levels, indicating narrative elements are easily identifiable in all reading levels
- **Referential cohesion -** only index to show a clear progression between the different reading levels; indicates cohesion between vocabulary and ideas is stronger in beginning levels than higher levels
- **Syntactic complexity** - inconsistent from possible outlier between reading level one and two; inconsistency may be due to corpus's samples or in how material is written for young readers transitioning to independent reading
- **Word abstractness -** relatively low for all reading levels; demonstrates lack of change in vocabulary between the reading levels

- **Situational model cohesion scores-** reversed, incorrectly rates the lower reading levels as being harder to read than the higher reading levels; reversal may be from lower reading level texts which do not use "the casual and temporal links that support comprehension" [Heibe10] or because these links would not be appropriate for simple texts.
- **The Type-Token ratio**- shifted between .5 and .6, did not match the expected results which were to have a high rating in the beginning reading levels and gradually decrease to the higher reading levels.

These results indicate that the Coh-Metrix indices, as a whole, have difficulty differentiating between the seven reading levels.

The Lexile index appeared to outperform all of the readability indices used, due to its use of mean sentence length, which accurately showed the progression in difficulty between the reading levels. While syntactic complexity would appear to be the most important factor in readability for young readers, word frequency and patterns, according to Heibert et al., are the critical factors in readability [Note2]. Heibert et al. counter the results of their study, which shows sentence length to be the greatest determiner of text readability, by referencing a 1986 study done by Allison D. Brennan, Connie A. Bridge, and Peter N. Winograd.

Story Grammar

Brennan et al. take an earlier look at how simple readability tests have affected child reading material, specifically classroom basal stories. While basal stories focus on literary skills not literary quality, [Brenn86] makes the case that educational writing should be written to follow the same plot structure of an actual story, a.k.a. story grammar.

To prove this point, two stories were taken from separate basal anthologies and edited to exhibit story grammar elements. These stories, listed in original and edited versions in Appendix D as A & B, were given to a class of second grade students with average reading test scores. The class was divided into four groups. Each group was given an original story and a revised story to read aloud. Group 1 received unedited story A to read first with the revised version of story B to read second. Group 2 received the same stories but read them in reverse order. Group 3 and 4 read the opposite selection of unedited story B and revised story A, with group 3 reading B first while group 4 read revised A first. The class was divided in this way to ensure that the order of the story being presented first would not affect the study's results. A multivariate analysis of the results showed that the story ordering was deemed insignificant ($p < .05$).

After reading orally the assigned story, each student was asked to recount all they could remember about the story. After telling all that he or she could remember, each student would answer 12 questions to further measure their remembrance of explicit and implicit information from the story. The results of the students' responses were scored and used to determine how the story structure effected free recall and probed recall of implicit and explicit information (Table 24). The story structure significantly improved the students' ability to recall explicit information, both in free and probed recall. Brennan et al. used these results to validate the need for story

structure in basal reading to help improve reader comprehension. In terms of readability, this study contributes two important findings.

Firstly, while both versions of a story had similar readability scores, they had differing reader comprehension scores. This encourages the idea of using story grammar to help children understand the text they are reading. Unfortunately, this study's results can only be applied to fictional children's text. While it is uncertain if story grammar can be applied to other reading genres without further testing, story grammar may prove useful in judging readability of fictional text and some non-fictional text. Story grammar may not be appropriate for non-fictional genres such as science or mathematics, which do not contain story grammar elements (such as a setting, conflict, or resolution). History and biographical information, on the other hand, do sometimes contain story grammar elements, but the questions are to what extent does this genre follow story grammar and would story grammar improve or confuse readability scores? For these reasons, future research will likely focus on fictional genres like science fiction, romance, or horror. These genres typically follow story grammar conventions and they may receive more accurate readability score with a story grammar check.

The second implication that can be drawn is that this study may invalidate Heibert et al.'s results or at least place doubt on the reason for Lexile's accuracy. If Heibert et al.'s basal stories lacked story grammar structure, as Brennan et al. suggests that the majority of basal stories do, then Lexile's accuracy may come from how basal stories are written. If Heibert et al.'s test was redone but with basal material consisting of story grammar structured and unstructured text, the study's results may have been different. Also, if future basal stories shift from being written to follow sentence length guidelines and instead follow story grammar structure, readability scores may alter. Coh-Metrix was unable to accurately score the stories, possibly, because of the lack

of story grammar. If the future basal stories follow story grammar structure, then Coh-Metrix and its five readability variables may prove to be more accurate in discerning different child reading levels.

Story Grammar: Conclusion & Implementation

These statements, while theoretical, do all stem from the proven effect that story grammar structure has on child literature readability. Because each genre can be affected by story grammar structure differently, a readability formula for all reading material would prove impractical and inaccurate. Therefore, specific readability formulas must be made for each genre or a method must be made to alter a base readability formula to handle the separate genres. The incorporation of genre detection or genre specific formulas will be a necessary asset to future readability testing.

Actual incorporation of story grammar into readability testing will be difficult. This will require the ability to actually "read" the text. It is currently uncertain how this feature would be implemented in an automated readability formula. A story grammar check can be done by hand by checking if the story contains each feature (setting, initiating event, internal response, attempt, consequence, reaction). Since children's stories are relatively incomplex, checking for story grammar should be simple for a trained reader, but the same may not be true of more advanced text. Complex stories can have multiple interwoven plots, which could cause difficulty in determining story grammar for a human reader, much less an automated reader. So while story grammar may prove important to readability, two things must be determined before adding it to a readability formula.

Firstly, the benefits of story grammar to the readability formula's accuracy must be calculated. This could be done initially by having a user determine if story grammar is being followed and fill out the necessary information to be fed to test readability formula. The reason for doing this initial test by hand is because the programming required for an automated story grammar test would be intensive. That is why actual implementation is the second concern. Once the benefits of story grammar structure have been determined then tests can be done to find a way to create an automated test. These simple suggestions do not, unfortunately, shed much light on how an automated story grammar feature would work. That is because there has been little focus in this area outside of the field of education, or at least little that can be found within the Natural Language Processing field. Even though the creation of a story grammar check is currently unknown, the checking and studying of individual sentence structure is not.

Noun Phrase Complexity & Function Word Density

Cowie et al. chose to study how readability of active voice, passive voice, sentential object, and extraposed subject sentence structures are affected by noun phrase complexity and function word density, within the context of medical text. For this study, the fourth sentence structure, extraposed subject sentences are sentential subject sentences with altered word orders. Noun phrase complexity refers to a noun's word frequency, level of occurrence in a language, and presence of compound nouns, which can be difficult to understand because the relationship between the nouns is not always explicit. For example, the phrase "diabetes risk" may "refer to a risk in contracting diabetes or to the risk of having diabetes."[Cowie10] Function words demonstrate grammatical relationships such as prepositions, wh-words, modals, auxiliaries, and determiners. [Leroy08] found patient blogs use twice as many function words as formal medical

documents, which may imply that laypersons find sentences with high function word density to be easier to understand. High function word density sentences may be easier to read because they have a rhythm closer to natural speech and may help space out individual concepts to allow for easier assimilation of information.

To determine how noun phrase complexity and function word density affected readability, Cowie et al. created test sentences for the four sentences structures. Each test sentence modeled one of the following combinations of noun phrase complexity and function word density:

- Simple noun phrase with high function word density
- Simple noun phrase with low function word density.
- Complex noun phrase with high function word density
- Complex noun phrase with low function word density

Table 20
Active Sentence Examples

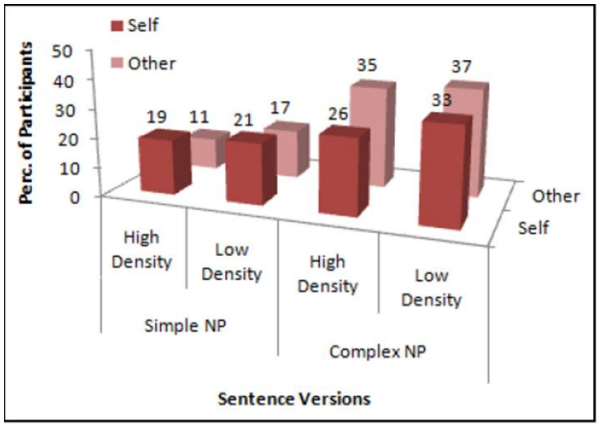
Noun Phrase Complexity	Function Word Density	Example
Simple	Low	Fortunately, changes in personal habits can prevent more damage to arteries supplying the heart.
	High	Fortunately, a few changes in your personal habits can prevent any more damage to the arteries supplying the heart.
Complex	Low	Fortunately, lifestyle changes can prevent further damage to coronary arteries.
	High	Fortunately, a few lifestyle changes can prevent further damage of your coronary arteries.

Each test sentence set, as shown in Table 20, was shown to a study group of 86 people (demographics found in Table 21)[Note3]. Participants were shown the test sentence sets and were asked to choose which sentences he or she believed were the hardest and easiest for him or her to read. Participants were then asked which sentence would be hardest and which would be easiest for someone else to read[Note4]. This was done to see if there was a difference in ratings for the *self*

(layperson opinion) versus ratings for *other* people (expert opinion). A difference was found with other being an extreme of self.

As seen in Figure 2, which displays the results for the active sentence set, the sentence with complex noun phrases and low function word density has the highest percentage of votes for the most difficult sentence for both *self* and *other*, with other having a slightly higher percentage than *self*. The results for easiest sentence were not released because these results were the reverse of those obtained for the difficult sentence. Since the sentence with simple noun phrases and high function word density was chosen the fewest times as most difficult, it can be inferred to be the easiest sentence to read.

Figure 2
Percentage of Participants Selecting Most Difficult Version Among Active Sentences



Comparisons of responses for the four sentence structures yielded two marked sets of contrasts. While active and passive voice sentences showed similar results for choosing the hardest sentence, the easiest sentence for passive voice was tied between both high function word density sentences. Passive voice also showed a greater disparity between *self* and *other* on the most difficult sentence, complex noun phrase with low function word density, with a 10% difference versus active voice, which had a 4% difference. Also, the sets of extrapolated and sentential object sentences have a substantial increase with percentage of participants choosing a complex noun phrase as being the most difficult. They, like passive and active voice sentences,

still show low function word density to be more difficult than high function word density, but the difference between two is not as clear, especially between the two simple noun phrase words.

Noun [...] & Function Word Density: Conclusion & Implementation

These two sets of contrasts, as well as the totality of the results, suggested complex noun phrases to be less readable than simple noun phrases and sentences with low function word density to be less readable than sentences with high function word density. Extrapolated and sentential sentences showed a greater difference between complex and simple noun phrases, while active and passive voice sentences showed a more even distribution, almost like a linear progression. These results are strong indicators of how noun phrase complexity and function word density affect readability. These results, however, merely show that sentence structure influences readability; they do not validate any difficulty rating for specific linguistic features. More research is necessary to validate this study's results[Note5].

Implementation for active and passive voice can be done by using existing features in NLP programs. What needs to be tested is how to use the data collected. A likely option would be to create a ratio of active vs. passive voice. If document X is Y percent passive voice than it is labeled a difficulty level of Z. Future testing would determine what percent of passive voice would go for each reading difficulty level. As for function word density, this could be done in a similar fashion to word frequency. Sentences could be tokenized and the number of function tokens per sentence could be calculated. Future testing would determine the number/percentage of function words per sentence required for each reading difficulty level.

Conclusion

Word frequency, sentence length, genre, story grammar, and sentence structure are the key features that have been covered. Word frequency has the ability to rate the difficulty and level of recognition of a word. With word frequency, a readability formula can rate the vocabulary of a text. But vocabulary may not be the best indicator of readability, as was found in the case of young readers. Within certain genres, such as basal stories, sentence length was a greater determining factor of readability. Therefore, the genre and type of text being rated must be taking into account for what features are used in the readability formula. Especially if story grammar is used. Not all genres follow story grammar, such as science or mathematical reference material. Because of their natural lack of story grammar, these genres would not be appropriate choices for a story grammar based readability formula. Sentence structure could possibly be less genre specific feature with its check of noun phrase complexity and function word density. This can also be seen, to some extent, as an application of word frequency which is used for calculating noun phrase complexity. These features show promise features in forming an accurate readability test. But being that this is not an exhaustive listing of features, further research is required to determine what features are most important to readability.

Future Work

The ranking and weighing of features is necessary to determine which features should be included and which can be left out. It will, most likely, be necessary to leave certain readability features out of the formula because with each feature addition, the performance and/or speed of the formula diminishes. The trade-off of accuracy versus performance will need to be calculated

so that a practical formula can be engineered. For instance, the incorporation of feature X results in the formula doubling its processing time but only increases a fraction of a percent in accuracy. Feature X would then be considered an undesirable feature and be left out of the formula.

Multiple tests are required for determining different benefits for each feature. For example, feature Y may require multiple calculations. Feature Y would run efficiently on a mainframe or computer with multiple processors, but run poorly on a personal computer with low RAM and low processing resources. When considering books, a massive computer system may be reasonable for a publishing company to use to determine their texts' readability. On the other hand, if one is considering a readability formula for a teacher to use on a personal computer for determining the readability of web content for students, a slimmer, less computational formula would be required. This is why multiple features are a benefit and a curse. They can make it difficult to pick which features should be included, but at the same time they allow for some flexibility and versatility in creating readability formulas.

While I consider my undergraduate thesis to be fairly thorough, it is still only a foundation. I say "foundation" because I have only covered small sections of readability formulas being researched which belong in a small section of the vast field of Natural Language Processing. I plan to fortify this foundation with further research in how to implement the readability features I have discussed and determine a way to quantify and rate their performance. Currently, the performance of a feature is rated by how accurate it can calculate a text's readability. I wish to take this a step further and define the performance of a feature by how efficient it operates. Efficiency could include the average processing time for a function, it's Big O notation, or the number of resources required to run the function (such as a database, hard

drive space, or even number of processors). My goal for this efficiency test is not to gauge every readability feature but rather to define a guideline for determining efficiency.

More than likely there is no magic bullet or singular formula that can accurately rate all texts. But once readability features are properly categorized and weighed, a logic decision can be made on how to make specific readability formulas. In the future, key features may be enabled or disabled by a user to help determine what level of accuracy or what features they consider most important. So in the end all of these features are not a hindrance, but a way to enable custom readability formulas for specific to generic readability testing needs.

Works Cited

[Arms13] Arms, William Y. "Digital Libraries". MIT Press, 2000. <http://www.cs.cornell.edu/wya/diglib/ms1999/glossary.html>. Retrieved on 3/4/13.

[Balot07] Balota, David A., et. al. "The English Lexicon Project." < http://elexicon.wustl.edu/userguide.pdf>. Retrieved on 3/4/13.

[Brit13] British National Corpus. < http://www.natcorp.ox.ac.uk/>. Retrieved 2/28/13.

[Brenn86] Brennan, Allison D, Connie A. Bridge and Peter N. Winograd. "The Effects of Structural Variation on Children's Recall of Basal Reader Stories." Reading Research Quarterly, Vol. 21 No. 1, 1986. pp. 91-104

[Brysb09] Byrsbaert, Marc and Boris New."Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English."Behavioral Research Methods, Vol. 4 No. 4.Psychonomic Society, 2009.

[Cowie10] Cowie, James R., Stephen Helmreich, and Gondy Leroy. "The Effects of Linguistic Features and Evaluation Perspective on Perceived Difficulty of Medical Text." *Proceedings of the 43rd Hawaii International Conference on System Sciences*. IEEE, 2010.

[Cryst03] Crystal, David. Cambridge Encyclopedia of Language. (excerpts). 2nd Ed. Cambridge: Cambridge University Press, 2003

[Desch09] Deschacht, Koen and Marie-Francine Moens. "The Latent Words Language Model." *Proceedings of the 19th Annual Belgian-Dutch Conference on*

Machine Learning (Benelearn 09). Tilburg, 2009; also as < http://class.inrialpes.fr/pub/deschacht-benelearn09.pdf>

[Gate12] "GATE: a full-lifecycle open source solution for text processing." <http://gate.ac.uk/overview.html> Retrieved January 1st, 2013

[Hiebe10] Hiebert, Elfrieda H. and P. David Pearson."An Examination of Current Text Difficulty Indices with Early Reading Texts."*Reading Research Report #10-01*.TextProject Inc., 2010. <http://www.cvedcvt.org/docs/8.TextProject_RRR-10.01_Text-Difficulty-Indices.pdf>

[Token08] "Tokenization." Cambridge University Press, 2008. <http://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html> Retrieved: 2/29/13

Appendix A: Glossary

Basal Stories- Literature created to express and/or teach specific reading skills. Typically used in class rooms, basal stories target specific age or grade levels to assist students in learning to read. [Brenn86]

Corpus- A large collection of writings.

Corpora- Plural form of *corpus*.

British National Corpus - "The British National Corpus (BNC) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English, both spoken and written." [Brit13]

Degree of Reading Power- "The DRP bases its semantic index on the count of characters" [Heibe10]

Fry Readability Formula- The Fry Readability Formula measures the average number of sentences and average number of syllables per word for every 100 words in a text.

GATE- An open source text processing software suite. Gate can perform tokenization, sense passive or active voice, and many other text processing features.

Lexeme- A lexeme is a single unit of meaning which can consist of a single word or multiple words. "A much used example is *kick the bucket* (='die'). Here we have a single unit of meaning, which happens to consist of three words."[Cryst13]

Lexical Decision Task- " In the lexical decision task (LDT), participants are presented with a string of letters (either a word or a nonword, e.g., FLIRP), and are asked to press one button if the string is a word and another button if the string is a nonword." [Balot07]

Lexical Decision Time- Amount of time taken to complete a Lexical Decision Task. [Balot07]

Mean Frequency of Words- "The mean frequency of a word is derived from the rankings of words within a massive databank of well over a billion words that Metametrics has amassed over the past 25 years." [Heibe10]

Natural Language Processing - " Use of computers to interpret and manipulate words as part of a language." [Arms13]

Non-Narrativity- "Narrative text tells a story, with characters, events, places, and things that are familiar to the reader and is closely affiliated with everyday oral conversation. Texts that follow a narrative structure have low percentiles on this scale." [Heibe10]

Reaction Time - Time for a study participant to read and say aloud a word or non-word that the participant is shown. [Balot07]

Referential Cohesion- "High cohesion texts contain words and ideas that overlap across sentences and the entire text, forming threads that explicitly connect the text elements for the reader. [...] a high percentile on referential cohesion indicates that a text is difficult and has few of the threads that support explicitness for readers." [Heibe10]

Semantic- The meaning of a word or sentence within specific context.

Situational Model Cohesion - "Causal, intentional, and temporal connectives help the reader to form a more coherent and deeper understanding of the text. A high percentile on situation model cohesion means lower levels of this feature and, consequently, more obstacles for comprehension for readers. Thus, a high percentile on this variable indicates a more difficult text." [Heibe10]

Spache Readability Formula- The Spache formula measures how many unfamiliar words are found in a text. Unfamiliar words are words that are not included in the Spache list of 1,036 words which are considered appropriate for readers below 3rd grade.

Stop List- A set of words that are excluded from text processing because they could slow down processing or askew results.

Syntactic - Relating to syntax (grammar guidelines for a language).

Syntactic Simplicity- "Sentences with few words and simple, familiar syntactic structures are easier to process and understand. When texts have high percentiles on this dimension, they have complex syntactic structures, which suggest that processing will also be complex" [Heibe10]

Tokenization - "Given a character sequence and a defined document unit, tokenization is the task of chopping it up into pieces, called *tokens* , perhaps at the same time throwing away certain characters, such as punctuation. Here is an example of tokenization" [Token08]

Token- A token is a grouping of characters that is in some way significant or identifiable. Tokenization can create simple word tokens, but tokens can also be parts of a word such as the "ed" postfix or the contraction "n't". A token can also include symbols or punctation, such as "@email" or ".com". [Token08]

Type-Token Ration- A ratio between the number of different words to the number of total words in a text. [Heibe10]

Word Concreteness- "Concrete words evoke mental images and are more meaningful to the reader than abstract words. High percentiles on this dimension mean that texts have a substantial number of abstract words. Higher portions of abstract words, in turn, make texts more difficult to comprehend." [Heibe10]

Appendix B: Tables

Table 1
Excerpts From Each Text

Text Type	Excerpt
Trade	A Duckling came out of the shell. "I am out!" he said. "Me too," said the Chick. "I am taking a walk," said the Duckling. "Me, too," said the Chick." "I am digging a hole," said the Duckling. "Me too," said the Chick. "I found a worm," said the Duckling.
Trade Instructional	Time for a bath, Biscuit. Woof, woof. Biscuit wants to play. Time for a bath, Biscuit. Woof, woof. Biscuit wants to dig. Time for a bath, Biscuit. Woof, woof. Biscuit wants to roll. Time for a bath, Biscuit. Time to get nice and clean. Woof, woof. In you go. Woof.
Textbook Core–Current	Pig in a wig is big, you see. Tick, tick, tick. It is three. Pig can mix. Mix it up. Pig can dip. Dip it up. Pig can lick. Lick it up. It is six. Tick, tick, tick. Pig is sad. She is sick. Fix that pig. Take a sip.
Textbook Core–Historical	Look, Dick. Dick! Dick! Help Jane. Go help Jane. Go, Jane. Go, Jane, go.
Text Ancillary–Decodable	Nan's Family On the Mat Sam sits on his mat. Pam sits on Sam. I am on Sam! Tim sits on Pat. Nan sits on Tim. Tip sits on Nan. Tip.
Text Ancillary–Guided	Funny Faces Look at the fish face. Look at the fox face. Look at the dog face. Look at the frog face. Look at the cat face. Look at the flower face. Funny faces!
Test (GORT-4)	See Father. Father is here. We want to play. Can you play, Mother? We can play here.

Source: [Hiebe10]

Table 2
Criteria for Guided Reading and DRA by Text Level

	K	1	2	3	4	5	6	7
Guided Reading[1]	A	B–C	D	E	F–G	H–I	J–K	L–M
DRA[2]	A–2	3–4	6–7	8–9	10–12	14–16	20–24	25–28

[1] Fountas & Pinnell, 1996, 1999
[2] Developmental Reading Assessment

Source : [Heibe10]

Table 3
Number of Texts by Text Type and Text Level

Text Type	Source of Texts	# of Texts	# of Texts by Text Level						
			1 (K)	2 (1.1)	3 (1.2)	4 (1.3)	5 (1.4)	6 (1.5)	7 (2)
Trade	Various sources	42	1	4	4	3	7	11	12
Trade Instructional	I Can Read series	72	6	6	12	12	12	12	12
Textbook Core–Current	Scott Foresman (2007)	42	6	6	6	6	6	6	6
Textbook Core–Historical	Scott Foresman (1962)	36	0	6	6	6	6	6	6
Text Ancillary–Decodable	Open Court (2000), Reading Mastery (1995)	84	12	12	12	12	12	12	12
Text Ancillary–Guided	Ready Readers (1997), Wright Group (1996)	84	12	12	12	12	12	12	12
Tests	BRI, DIBELS, DRA, GORT, QRI)	84	5	5	4	16	17	20	17
Totals		444	42	51	56	67	72	79	77

Source : [Heibe10]

Table 4
Readability Measures (Means) by Text Levels

Text Levels	DRP	Fry	Spache	Lexile	MSL[1]	MLF[1]
1	1.6	1.3	1.9	86.9	4.9	3.8
2	1.6	1.1	1.8	140.0	5.0	3.6
3	1.6	1.1	1.8	238.0	6.1	3.7
4	1.6	1.3	1.8	238.2	6.4	3.8
5	1.8	1.6	2.0	346.0	7.2	3.7
6	2.0	2.0	2.2	420.6	8.0	3.7
7	2.2	2.6	2.3	489.1	8.8	3.7

[1] MSL (mean sentence length) and MLF (mean lexical frequency) were provided as part of the Lexile analysis of the texts. Although they are not defined as readability measures, they are included here as supplementary information.

Source : [Heibe10]

Table 5
Conventional and Current Readability Indices
for Text Types

Text Types	DRP	Fry	Spache	Lexile	MSL[1]	MLF[1]
Trade	2.4	2.8	2.5	534.6	9.2	3.7
Trade Instructional	1.8	1.6	1.9	276.0	6.4	3.7
Textbook Core–Current	1.9	1.7	2.0	320.7	6.6	3.6
Textbook Core–Historical	1.6	1.3	1.5	185.8	5.9	3.7
Text Ancillary–Decodable	1.6	1.3	2.0	315.7	6.9	3.7
Text Ancillary–Guided	1.8	1.5	1.9	228.4	6.2	3.7
Tests	1.8	1.8	1.9	333.2	7.5	3.8

[1] MSL (mean sentence length) and MLF (mean lexical frequency) were provided as part of the Lexile analysis of the texts. Although they are not defined as readability measures, they are included here as supplementary information.

Source : [Heibe10]

Table 6
Means for Coh-Metrix Indices by Text Level

Text Levels	Non-narrativity	Referential cohesion	Syntactic complexity	Word abstractness	Situation model cohesion	Familiarity	Type/ Token
1	20.9	9.3	4.4	35.2	78.5	1.9	.6
2	18.9	10.5	10.7	34.9	79.9	2.3	.5
3	19.8	14.6	7.3	42.7	62.7	2.2	.5
4	14.6	20.5	7.7	45.8	64.7	2.1	.5
5	17.5	32.0	10.2	37.2	54.7	2.2	.5
6	19.7	39.8	12.7	37.4	52.6	2.2	.6
7	18.6	46.2	16.5	37.4	53.4	2.2	.5

Source : [Heibe10]

Table 7
Means for Coh-Metrix Indices by Text Type

Text Types	Non-narrativity	Referential cohesion	Syntactic complexity	Word abstractness	Situation model cohesion	Familiarity	Type/ Token
Trade	42.8	22.9	21.7	44.8	19.5	2.3	.6
Trade Instructional	25.4	11.2	17.3	37.8	6.0	2.1	.5
Textbook Core–Current	45.9	7.9	32.5	34.0	5.5	2.3	.5
Textbook Core–Historical	12.1	2.3	11.3	19.3	5.3	1.8	.4
Text Ancillary–Decodable	38.0	7.8	16.4	17.7	12.7	2.1	.5
Text Ancillary–Guided	43.1	12.2	16.9	18.6	8.1	2.3	.5
Tests	22.1	28.5	17.1	27.9	14.9	2.1	.6

Source : [Heibe10]

Table 8

Frequency Norms Used in Research on Memory and Language Processing in the November 2008 Issue of the *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Source	Topic	Frequency Norms
Huber, Clark, Curran, & Winkielman (2008)	recognition memory	KF
Szpunar, McDermott, & Roediger (2008)	memory for word lists	KF
O'Malley & Besner (2008)	reading aloud	HAL
Hockley (2008)	recognition memory	KF
McDonough & Gallo (2008)	autobiographical memory	KF
McKay, Davis, Savage, & Castles (2008)	reading aloud	KF
Klepousniotou, Titone, & Romero (2008)	understanding ambiguous words	KF
Drieghe, Pollatsek, Staub, & Rayner (2008)	eye movements in reading	KF

Note—KF, Kučera and Francis (1967).

Source : [Brysb09]

Table 9

Percentage of Variance Accounted for in the Elexicon Lexical Decision Times by Various Portions of the British NationalCoprus (N = 31,201)

Size (Million Words)	R^2 (%)
0.5	48.7
1	51.3
2	53.3
4	55.1
8	55.9
16	56.4
32	56.1
88	56.1

Source : [Brysb09]

Table 10

Percentage of Variance Accounted for in High-Frequency (HF) and Low-Frequency (LF Words of the Elexicon Lexical Decision Times by various portions of the British National Corpus

Size (Million Words)	HF R^2 (%)	LF R^2 (%)
0.5	50.3	38.2
1	51.3	40.8
2	51.1	43.1
4	51.3	45.4
8	51.6	46.7
16	51.3	47.6
32	51.1	47.7
88	51.2	48.0

Note—N = 3,754 and 27,572 for high- and low-frequency words, respectively. HF words had a frequency of >20 per million; LF words had a frequency of <10 per million.

Source : [Brysb09]

Table 11

Portions of Variance Explained by HAL and SUBTLEX for Words of Different Lengths

Word Length	HAL	SUBTLEX
3	.38	.51
4	.47	.53
5	.47	.49
6	.47	.47
7	.45	.44
8	.44	.42
9	.42	.38
10	.41	.36
11	.40	.35
12	.39	.33
13	.39	.30

Source : [Brysb09]

Table 12

Percentages of Variance Explained by the Various Frequency Counts in the Lexical Decision Task Accuracy (Acc) Data Reported by Balota, Corese, Sergent-Marshall, Spieler, and Yap (2004) and the Elexicon Project (Balota et al., 2007)

Measure	Acc_{young}	Acc_{old}	Acc_{Elex}
KF	18.0	7.0	22.5
Spoken	16.8	5.5	23.0
Celex	24.2	10.4	26.0
HAL	24.7	8.2	31.3
Zeno	25.5	10.7	29.8
BNC	22.8	9.0	25.4
SUBTL	27.7	12.4	38.3

Note—Multiple regression analysis involved log(freq+1), $\log^2$(freq+1), and word length in number of letters. All stimuli were monosyllabic (N = 2,406). KF, Kučera and Francis (1967); BNC, British National Corpus.

Source : [Brysb09]

Table 13

Percentages of Variance Explained in the Reaction Time Data Reported by Balota, Cortese, Sergent-Marshall, Spierer, and Yap (2004) and Balota et al. (2007)

	R^2 (%)							
Measure	LDT_{young}	LDT_{old}	LDT_{Elex}	$z_{LDTElex}$	NMG_{young}	NMG_{old}	NMG_{Elex}	$z_{NMGElex}$
KF	31.8	23.8	32.1	38.0	20.0	21.5	22.7	23.9
Spoken	31.1	19.9	31.9	38.5	19.7	20.6	23.0	24.4
Celex	37.0	28.4	33.8	40.8	20.0	21.3	22.3	23.9
HAL	36.7	24.2	37.7	45.6	20.5	21.9	23.9	25.3
Zeno	38.8	30.1	35.6	43.3	20.5	22.2	23.5	24.7
BNC	34.8	26.8	34.4	41.2	19.7	21.5	22.8	24.0
SUBTL	42.2	29.3	40.1	48.6	21.0	22.9	24.1	25.2

Note—Multiple regression analysis involving log(freq + 1), $\log^2$(freq + 1), and word length in number of letters. All stimuli were monosyllabic (N = 2,406). LDT, lexical decision task; NMG, word naming; KF, Kučera and Francis (1967); BNC, British National Corpus.

Source : [Brysb09]

Table 14

A Comparison of the Variance Explained by CELEX Word Form (WF) Frequencies and Lemma Frequencies for Performance in the Experiments Reported in the Elexicon Project (z Scores) and Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004). When Word Length (Number of Letters and Number of Syllables if Applicable) Was Included and When it Was not Included

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency						
Celex WF	21.3	33.9	21.4	36.2	39.4	34.6
Celex lemma	21.9	36.6	21.3	37.9	37.1	32.4
Frequency + word length						
Celex WF	25.2	36.1	25.8	60.7	41.1	37.6
Celex lemma	25.8	37.9	25.4	60.2	40.0	35.9
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency						
Celex WF	23.9	10.3	36.9	27.9	6.4	9.8
Celex lemma	25.3	10.1	36.5	27.3	6.2	9.2
Frequency + word length						
Celex WF	24.2	10.4	37.0	28.4	20.0	21.3
Celex lemma	25.5	10.3	36.7	28.0	20.0	21.2

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming.

Source : [Brysb09]

Table 15

Percentages of Variance Accounted for by the Word Frequency SUBTL Index and the Contextual Diversity SUBTL Index for the Elexicon Project and the Monosyllabic Words Investigated by Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004)

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency						
$SUBTL_{WF}$	22.0	32.9	26.4	49.2	45.2	42.5
$SUBTL_{CD}$	23.4	36.8	28.0	49.5	46.8	43.6
Frequency + word length						
$SUBTL_{WF}$	30.1	40.7	33.6	62.3	45.2	43.5
$SUBTL_{CD}$	31.3	44.0	34.9	62.9	46.8	44.6
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency						
$SUBTL_{WF}$	26.4	12.1	42.1	29.5	9.7	13.6
$SUBTL_{CD}$	29.3	13.9	44.2	31.0	9.4	13.3
Frequency + word length						
$SUBTL_{WF}$	27.7	12.5	42.3	29.6	21.1	22.8
$SUBTL_{CD}$	30.6	14.3	44.3	31.1	21.2	23.0

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming; WF, word form frequency; CD, contextual diversity.

Source : [Brysb09]

Table 16

Percentages of Variance Accounted for When the Word Form (WF) Frequency and Contextual Diversity (CD) Measures Are Based on All the Occurrences of the Words or Only on the Occurrences of the Words Starting with a Lowercase Letter, Separately for the Elexicon Project and the Monosyllabic Words Investigated by Balota, Cortese, SErgent-Marshall, Spieler, and Yap (2004)

	Elexicon					
	$Acc_{all\ words}$ (N = 37,059)	$Acc_{Nsyl=1}$ (N = 5,766)	$Acc_{Nsyl=2}$ (N = 14,306)	$RT_{all\ words}$ (N = 31,201)	$RT_{Nsyl=1}$ (N = 5,042)	$RT_{Nsyl=2}$ (N = 12,039)
Frequency + word length						
$SUBTL_{WF}$	30.1	40.7	33.6	62.3	45.2	43.5
$SUBTL_{CD}$	31.3	44.0	34.9	62.9	46.8	44.6
$Frequency_{lowercase}$ + word length						
$SUBTL_{WF}$	31.1	44.2	34.5	62.7	47.5	44.0
$SUBTL_{CD}$	31.8	46.1	35.2	63.0	47.8	44.4
	Balota et al. (N = 2,406)					
	Acc_{young}	Acc_{old}	LDT_{young}	LDT_{old}	NMG_{young}	NMG_{old}
Frequency + word length						
$SUBTL_{WF}$	27.7	12.5	42.3	29.6	21.1	22.8
$SUBTL_{CD}$	30.6	14.3	44.3	31.1	21.2	23.0
$Frequency_{lowercase}$ + word length						
$SUBTL_{WF}$	31.0	14.3	45.3	32.1	20.9	22.8
$SUBTL_{CD}$	32.0	15.3	45.5	32.1	21.0	22.9

Note—Acc, accuracy; RT, reaction time; LDT, lexical decision task; NMG, word naming.

Source : [Brysb09]

Table 17

Percentages of Variance Accounted for by the Different Frequency Measures for the Elexicon Project When the Analyses Are Limited to the Words That More Often Start With a Lowercase Letter Than With an Uppercase Letter (RT Analyses Limited to Words With an Accuracy Level > .66)

Measure	Acc_{all} (N = 31,246)	$Acc_{Nsyl=1}$ (N = 5,281)	$Acc_{Nsyl=2}$ (N = 12,439)	RT_{all} (N = 27,350)	$RT_{Nsyl=1}$ (N = 4,721)	$RT_{Nsyl=2}$ (N = 10,840)
HAL	29.5	38.3	30.0	59.1	46.4	42.1
Zeno	30.2	40.6	29.9	58.5	44.4	40.7
$SUBTL_{WF}$	28.6	40.7	31.5	58.1	47.4	42.2
$SUBTL_{WFlow}$	28.9	41.3	31.8	58.3	47.6	42.4
$SUBTL_{CD}$	29.7	43.2	32.7	58.7	47.8	42.9
$SUBTL_{CDlow}$	30.0	43.6	32.8	58.7	47.8	42.9
Hal + Zeno + SUBTL	31.1	40.4	31.8	60.0	47.8	43.6

Note—Acc, accuracy; RT, reaction time; WF, word form frequency; CD, contextual diversity.

Source : [Brysb09]

Table 18

Frequencies of Words Used in Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) Overestimated and Underestimated on the Basis of HAL (Relative to $SUBTLEX_{US}$)

Overestimated on the Basis of HAL	Underestimated on the Basis of HAL
thru	swear
null	dad
lisp	bye
pub	staunch
warp	calm
death	breath
node	cinch
text	sir
stilt	slept
mime	beg
spool	shut
Web	knock
sphere	toast
mint	till
vale	sit
and	kid
width	hush
volt	cheer
prompt	drink
strand	wow
strait	drunk
dole	sweet
ram	booze
hind	wake
mode	hang

Source : [Brysb09]

Table 19

Frequencies of Words Used in Balota, Cortese, Sergent-Marshall, Spieler, and Yap (2004) Overestimated and Understimated on the Basis of Kucera and Francis (KF), Relative to $SUBTLEX_{US}$

Overestimated on the Basis of KF	Underestimated on the Basis of KF
chive	hook
whig	sneak
shear	stuff
strode	freak
oust	lab
fig	bet
spire	trash
daunt	fry
and	heck
gaunt	swear
loath	weird
flux	thank
clad	scare
strait	steal
strove	ouch
null	dad
thru	jerk
wry	yeah
scribe	cute
quill	bike
clung	pal
scant	hey
sprawl	ass
sparse	bye
blithe	wow

Source : [Brysb09]

Table 20
Active Sentence Examples

Noun Phrase Complexity	Function Word Density	Example
Simple	Low	Fortunately, changes in personal habits can prevent more damage to arteries supplying the heart.
	High	Fortunately, a few changes in your personal habits can prevent any more damage to the arteries supplying the heart.
Complex	Low	Fortunately, lifestyle changes can prevent further damage to coronary arteries.
	High	Fortunately, a few lifestyle changes can prevent further damage of your coronary arteries.

Source : [Cowie10]

Table 21
Highest Education Achieved by Participants

Highest Education Level Achieved N = 86	Percentage
High School	6
Some Community College	20
Community College Associate Degree	13
Some College	26
Bachelor's Degree	19
Master's Degree	11
Ph.D. Degree	5

Source : [Cowie10]

Table 22

Highest Average Flesch-Kincaid Readability Grade Levels per Condition

Structure	Noun Phrase Complexity	Function Word Density	Average Flesch-Kincaid Grade Level (N=4)
Active	Simple	High	12.9
		Low	12.8
	Complex	High	16.0
		Low	15.7
Passive	Simple	High	12.7
		Low	12.9
	Complex	High	15.0
		Low	14.8
Extraposed Subject	Simple	High	11.3
		Low	11.0
	Complex	High	13.1
		Low	12.1
Sentential Subject	Simple	High	12.2
		Low	11.7
	Complex	High	14.6
		Low	12.8

Source : [Cowie10]

Table 23
Readability of Passages

Table 1 Readability of passages

	A_1	A_2	B_1	B_2
Number of words in complete text	113	112	85	105
Number of words not found on the graded word list—'unique words'	2	2	2	2
Number of sentences	12	12	11	12
Number of different words used in the text	38	41	30	41
Readability level	2.25	2.35	2.02	2.15
Number of new words added		13		20
Percentage of original words used in well-formed passage		87%		77%

Note. A_1 = poorly-structured version of passage 1. A_2 = well-structured version of passage 1. B_1 = poorly structured version of passage 2. B_2 = well-structured version of passage 2.

Source : [Brenn86]

Table 24
Means (and standard deviations) for poorly-structured and well-structured versions of both stories

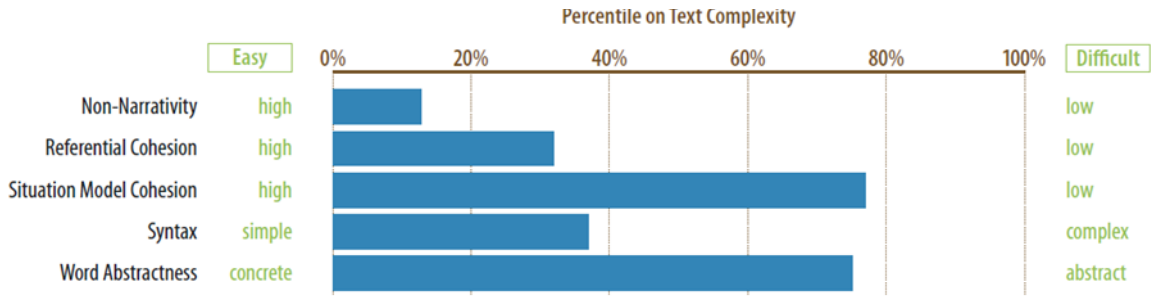
Table 2 Means (and standard deviations) for poorly-structured and well-structured versions of both stories

	Story 1		Story 2	
Variable	Poorly-Structured	Well-Structured	Poorly-Structured	Well-Structured
FREX[a]	.28 (.14)	.41 (.17)	.31 (.12)	.47 (.12)
FRIM[b]	15.75 (12.99)	12.44 (11.25)	16.00 (11.53)	10.69 (6.64)
PREX[c]	12.31 (2.60)	16.88 (1.09)	8.81 (2.71)	16.69 (0.95)
PRIM[d]	3.75 (1.39)	5.19 (1.17)	1.56 (0.89)	5.25 (1.07)
WTSEQ[e]	.13 (.09)	.25 (.17)	.19 (.11)	.32 (.21)

[a]Free recall of explicit information; scores are expressed as proportions.
[b]Free recall of implicit information; scores can range from 0 upward.
[c]Probed recall of explicit information; scores can range from 0 to 18.
[d]Probed recall of implicit information; scores can range from 0 to 6.
[e]Weighted sequence of information recalled; scores are expressed as proportions.

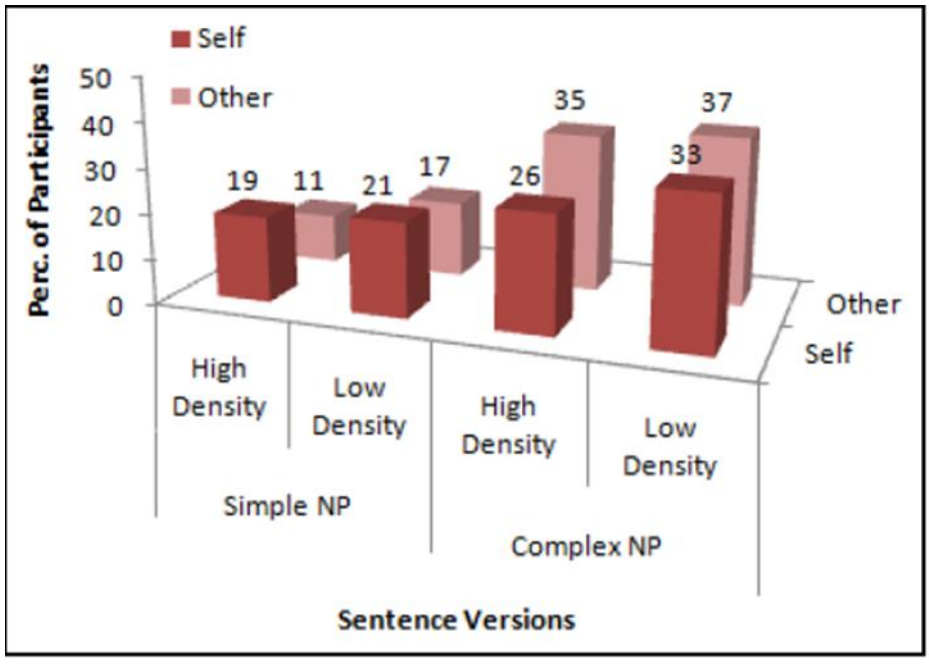
Source : [Brenn86]

Figure 1
Coh-Metrix Variables as Percentiles
for *Morris Goes to School*



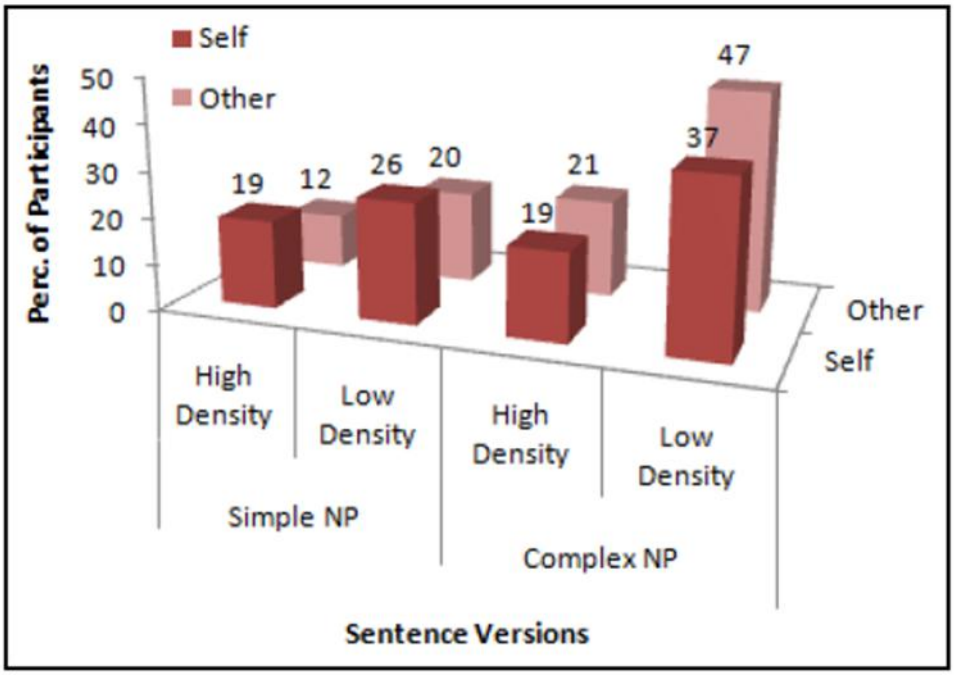
Source : [Heibe10]

Figure 2
Percentage of Participants Selecting Most
Difficult Version Among Active Sentences



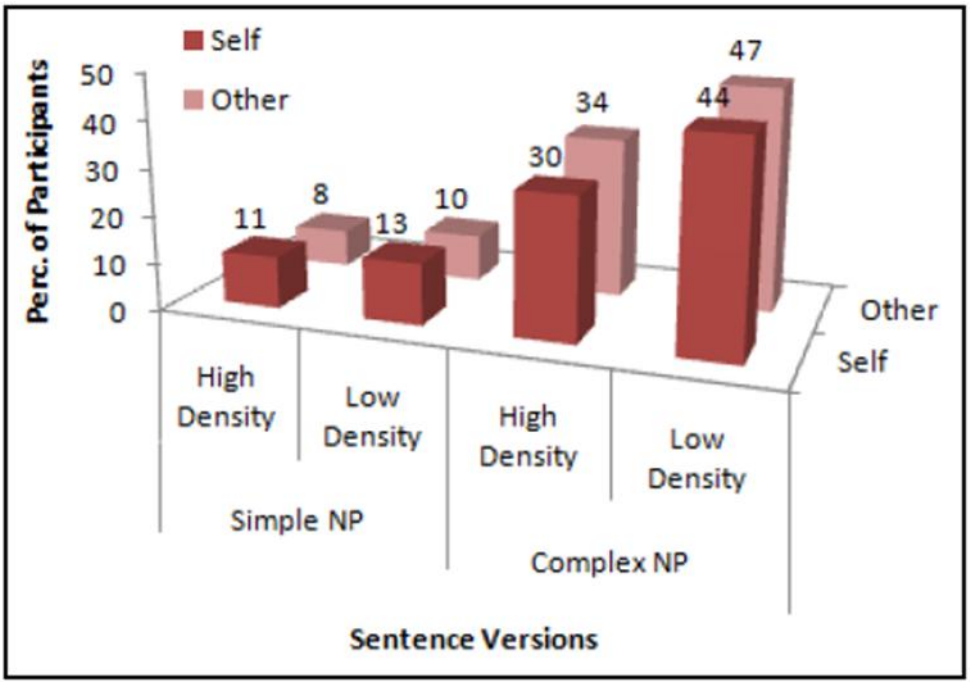
Source : [Cowie10]

Figure 3
Percentage of Participants Selecting Most
Difficult Version Among Passive Sentences



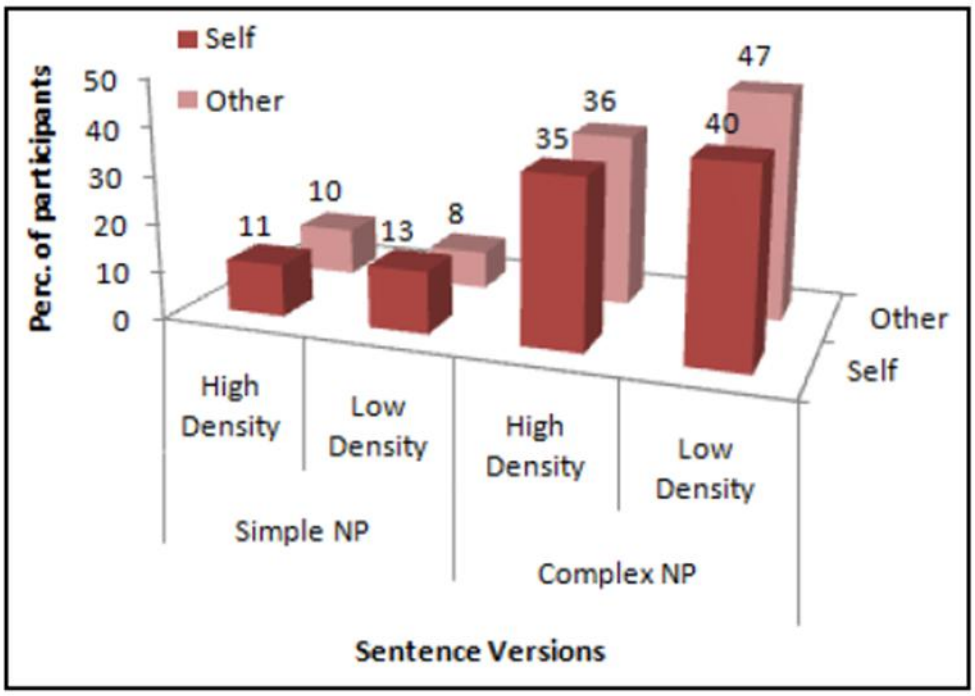
Source : [Cowie10]

Figure 4
Percentage of Participants Selecting Most
Difficult Version Among Extraposed Sentences



Source : [Cowie10]

Figure 5
Percentage of Participants Selecting Most
Difficult Version Among Sentential Sentences



Source : [Cowie10]

Appendix C: Corpora Sources for [Heibel10]

Test texts are created by following readability formulas. This means that text passages in tests are typically not from genuine reading sources such as magazines or books. The test materials used were the following resources:

"(a) the DevelopmentalReading Assessment (DRA) (Beaver, 1997)—an assessment based on a set ofguided reading levels;

(b) the Gray Oral Reading Test (GORT-4) (Wiederholt& Bryant, 2001);

(c) two informal reading inventories—the Qualitative ReadingInventory (QRI) (Leslie &Caldwell, 2001) and the Basic Reading Inventory(BRI) (Johns, 1997);

(d) the benchmark oral reading fluency assessments of the Dynamic Indicators of Basic Essential Literacy Skills (DIBELS) (Good &Kaminski, 2002)." [Heibe10]

Textbook Ancillaries-

These consist of guided reading texts and decodables. Guided reading texts consist" of individual books, 8–32 pages in length, that are clustered in levels that vary in difficulty." [Hebe10] Two guided reading groups were used, one from Australia by the Wright Group publisher and one American by the Ready Reader publisher. Decodables are "numerous sets of stand-alone programs of decodable texts. Similar to guided reading texts, the decodables are small books. Unlike the guided reading programs where the difficulty levels of books are determined on the basis of book and print features, content, text structure, and literary elements, the difficulty levels of decodables are typically a function of the phonics content represented in the texts. Those phoneme-grapheme patterns that have one-to-one correspondences (e.g., short a in cat) are typically viewed as less difficult and appear in earlier levels. Patterns where a phoneme is represented by

more than one grapheme (e.g., long a in gate) are considered more difficult and come later in the sequence of texts. Texts from two programs of decodables were used in this analysis: (a) the Open Court Reading Program (Adams et al., 2000) and (b) Reading Mastery (Englemann & Brunner, 1995)." [Heibe10]

Textbook Core-

"We used two textbook programs in this analysis: a program currently inuse—Scott Foresman's *Reading Street* (Afflerbach et al., 2007)—and a historical copyright of this program—Scott, Foresman, & Company's *The New Basic Readers* (Robinson, Monroe, & Artley, 1962). We used the Scott Foresman programs for several reasons, the most prominent of which is that this is the only program still published which Chall (1967/1983) reviewed. In addition, Scott Foresman's *Reading Street* showed the greatest percentage of market share during the 2008–09 school year (Education Market Research, 2010)." [Heibe10]

Trade-

"The sample of trade books for this study came from three sources: (a) Caldecott award-winning picture books, (b) picture books listed in the Read-Aloud Handbook (Trelease, 2006), and (c) the trade books on a list of grade-one literature from Accelerated Reader (Renaissance Learning, 2010). For the books on the Accelerated Reader list, presence in a public library collection was regarded as an indication that a book was of trade quality and not a textbook. Those books that appeared in the public library collection were included in the sample; those books that didn't appear weren't. The books from the other two sources were reviewed by two raters, both with teaching experience in the primary grades and knowledge of children's literature. Those books that both raters identified as

appropriate for independent reading by primary level students were included in the sample." [Heibe10]

Trade Instructional-

These are books intended to teach reading skills to young readers. The selection used for this study was the *I Can Read* series by HarperCollins.

Appendix D: Basal Reading Passages

Passage A: Original Basal Reader Version

In the grass was a little hill.
On the little hill was a little house.
In the little house was a little witch.
On the little witch was a big hat.
It was a hat that a big witch had lost.
The big had looked funny.
But the little witch was happy with the big hat.
The big had had big magic.
With the hat on, the little witch jumped with grasshoppers, swam with ducks, and ran fast with rabbits.
Morning after morning, the little witch went down the hill.
The little witch went to the pond to sing songs with the turtle.

Passage A: Revised Version

Setting:	A little witch lived in a house on the hill. She had a big hat. The hat was magic.
Initiating Event:	Every morning the little witch went down the hill to the pond. When she had the big had on, she could jump with the grasshoppers and run fast with the rabbits.
Internal response:	One morning the little witch wanted to swim with the ducks.
Attempt:	She had the big hat on when she went in the pond to swim with the ducks.
Consequence:	The big hat fell in the pond. She lost the big magic hat. The little witch was not magic without the big magic hat.
Reaction:	She looked for the magic hat, but it was lost. The little witch was not happy.

Passage B: Original Basal Reader Version

The afternoon sun was on the grass.
A big bee went up and up in the afternoon sun.
The bee went on and was lost in the shadows.
In the shadows was a house.
It was a very, very little house.
The big bee went over the house.
The bee looked down but did not see the little house.
The house in the shadows was Mr. Fig's house.
Mr. Fig was very little.
Little Mr. Fig was happy in the little house.

Passage B: Revised Version

Setting:	Mr. Fig was a happy little man. He liked to play in the grass.
Initiating event:	One afternoon Mr. Fig was in the grass by the house. He saw a big bee fly up and over house.
Internal response:	Mr. Fig wanted to play with the bee but he could not see the bee. Mr. Fig looked and looked for the bee. The bee was down in the grass.
Attempt:	Mr. Fig saw the bee. He ran over to play with the bee.
Consequence:	The bee stung Mr. Fig on the nose.
Reaction:	That did not make Mr. Fig very happy. He went in the house and the bee flew away.

Appendix E: Notes

Note1: Word frequency norms, currently, focus on individual words and not lexemes. To clarify, lexemes being basic lexical units can consists of individual words. Lexemes also consist of multiple words being used together. These multi-word lexemes could prove a crux for word frequency readability scores. For instance, lexemes such as "kick the bucket", an English euphemism for death, would be given a rating based on its individual words, not on the frequency the phrase is actually used. While "kick the bucket" may seem a common phrase, but non-native speakers or even non-American speakers may not grasp the death reference as quickly as their native English speaking counterparts. Lexeme frequency rating could have considerable effect on fictional text that contain flowery and illusionary language. For example, Homeric similes often use normal words, such as "Herculean strength" or "Achilles heel", but what these lexemes refer to is not inherently known from the words that make up the lexeme.

Examples of multi-word lexemes could be written for many more pages, but the bottom line is that these lexemes have the potential of altering the readability of a text and singular word frequency norms are unable to handle this potential readability issue. A lexeme frequency norm will need to be created that would be able to handle not only individual words but be used for multi-word lexemes. The closest to a lexeme frequency norm can be found in Google's N-Gram data set which provides frequencies for n-gram sets that they have collected from millions of public web pages. This is not an actual lexeme frequency but merely the frequency of word sets such as "serve as the info" or "ceramics consist of" [Franz06]. Also, while the n-gram data would contain many lexemes, it may not give ratings for lexemes but the word set that makes them up. Going

back to the "kick the bucket" example, a n-gram frequency may include the times that "kick the bucket" was used and differentiate when it was actually used as a euphemism. This means that the sentences "Little Jonny kicked the bucket down the street" and "Old farmer John had a heart attack and kicked the bucket" would both contribute to "kick the bucket" 's frequency. Therefore, an actual lexeme corpus must be built and used for determining lexeme frequencies.

Note2: It is uncertain how the authors concluded that vocabulary and word patterns played a greater role in text difficulty than syntactic features like sentence length. This seems to run counter to their results. The cited source, [Brenn86], was a study to see if story grammar could be used to improve a child's memory recall of a story. While this study did show improvement in memory recall by altering textbook stories to follow a story grammar, it only tested this theory on second graders. Since Heibert and Pearson are working to show the differences between the text difficulty of kindergarten through second grade it does not seem appropriate to attach the results of the [Brenn86] study to all young readers. For the authors to refute their own results and, instead, promote the results of [Brenn86] gives the impression that this study was an initial study that did not follow their hypothesis.

Note3: The educational background of the participants, as shown in Table 2, raises concerns about the study's results. The average education level for the group is slightly above the one for the U.S. populace, as determined from the 2008 U.S. census. While Cowie et al. acknowledged this point, they failed to discuss how well their results would generalize to the U.S. population. Since this research is intended to make medical text more accessible

for laypersons, a study like this should presumably include more participants at the lower end of the bell curve. While their responses might mimic their counterparts who have higher levels of education, a follow-up study is needed to verify this claim.

Note4: Since each participant was shown the same test sentence set with all test sentences in the same order, sentence order could have affected results. Cowie et al., however, argued that the lack of differences in the results from the various test sets suggests that sentence order only played a marginal role in the results. They added that future research will randomize test sentences.

Author's note: This study appears to have been an initial study or a study from a larger, ongoing study. This would explain why the results were released without testing randomized sentence order. While the authors' results suggest that low function word density and complex noun phrase sentences correlate with low readability, sentence order affects decisions. The researchers appear to have minimized the influence of sentence order, which does not invalidate their results but does lower the confidence in the results.

Note5: *Author's note*: While this study provides compelling evidence that noun phrase complexity and low function word density decrease readability, this study is inconclusive, since the educational background of the participants neither matches populace's bell curve nor the targeted audience which has a lower education background. Cowrie et al. also failed to demonstrate that their results were independent of the order in which they presented their test sentence. While the study seems inconclusive, its hypotheses warrant further research. These data suggest that the linguistic features that

the authors have identified affect a medical text's readability. What remains to be determined is the extent to which readability is affected.