USING PART OF SPEECH STRUCTURE OF TEXT IN THE PREDICTION OF ITS READABILITY

by

JAGADEESH KONDRU

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2006

ACKNOWLEDGEMENTS

I would like to thank my supervisor Mr. David Levine for his guidance, motivation and support. I am very grateful to Dr. Manfred Huber and Mr. Gil Carrick for their enthusiastic feedback through the course of my thesis. Thanks are also due to Dr. Lionidas Fegaras for serving on my thesis committee.

I would like to thank my family and friends for their encouragement.

November 25, 2006

ABSTRACT

USING PART-OF-SPEECH STRUCTURE OF

TEXT IN THE PREDICTION OF ITS

READABILITY

Publication No. _____

Jagadeesh Kondru, M.S

The University of Texas at Arlington, 2006

Supervising Professor: David Levine

Readability formulas predict the reading difficulty associated with text. They typically output a U.S. school grade level that indicates the reading ability required of a person in order for him to comprehend that text. Ability to predict text readability is useful because it helps educators select appropriate texts for students and authors write texts accessible to the audience they target. Existing readability formulas are based on countable aspects of the text such as average sentence length and average word length. We propose a new readability formula, the Readability Index, which is based on the part-of-speech structure of sentences in a text. We provide experimental results which

show that the Readability Index makes better grade predictions than existing readability formulas.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

"This amount is: **(a)** reported in box 1 if it is a distribution made to you from a nonqualified deferred compensation or nongovernmental section 457(b) plan or **(b)** included in box 3 and/or 5 if it is a prior year deferral under a nonqualified or section 457(b) plan that became taxable for social security and Medicare taxes this year because there is no longer a substantial risk of forfeiture of your right to the deferred amount."- *Instructions for the employee*, Form W-2

"When the vehicle is sold, title holder must assign and furnish this title, current license receipt, and signed application for the title (Form 130- U) indicating sales price to the purchaser who must file application with county tax assessor collector within 20 working days to avoid penalty." – *Texas Certificate of Title*

"Other uses or disclosures of your protected health information will be made only with your written authorization, unless otherwise permitted or required by law. You may revoke an authorization at any time in writing, except to the extent that we have already taken action on the information disclosed or if we are permitted by law to use the information to contest a claim or coverage under a health plan."- *HIPAA Privacy Notice*, El-Dorado County, California

"Cigarette smoking has been identified as the most important source of preventable morbidity and premature mortality worldwide. Smoking-related diseases

claim an estimated 438,000 American lives each year, including those affected indirectly, such as babies born prematurely due to prenatal maternal smoking and victims of "secondhand" exposure to tobacco's carcinogens." – *Smoking 101 Fact Sheet*, American Lung Association

"We may decline to process any full or partial balance transfer request and will not process a balance transfer request from any other account or loan that we or any of our affiliates issued. We may not use your total credit access line or credit line when honoring balance transfers because the total balance transfers and any related fees and finance charges may take your balance over the available credit access line or credit line." – *Two Easy Steps to Transfer Balances*, Chase World MasterCard

"Any changes to the antenna or the device could result in the device exceeding RF exposure requirements and void user authorization to operate the equipment. In addition, this transmitter must not be co-located or operating in conjunction with any other antenna or transmitter. " – *Instruction Manual*, Sony PSP 1001

The text selections we just quoted have two things in common:

- They are all intended for consumption by the average adult in the United States.
- According to popular readability formulas, none of the documents they represent are readable by the average U.S adult. They are all written at college level while the average adult in the U.S reads at 7th grade level [1].

What is readability? What is a readability formula?

A text is considered readable to the extent that the intended readers are able to comprehend it quickly, accept it (i. e persevere in reading it), and understand it clearly [2]. Readability of text and the factors affecting are discussed in detail in the following chapters of this thesis.

A Readability formula is an equation that gives an estimate of readability of a text. The estimate is generally in terms of the number of years of education one needs to have to comprehend that text. We discuss various readability formulas in chapter 2.1.2.

What is this thesis about?

This thesis introduces a novel formula to predict readability of text in terms of the number of years of education (indicated by the US School grade level) one needs to have to comprehend that text. The rest of the thesis is organized into the following sections:

- Background: A thorough discussion of readability, an explanation of popular readability formulas and a discussion of the elements of statistical language modeling that underlie our formula for readability.
- Hypothesis: Statement of our approach to readability prediction and our intuition for this approach.

- Experimentation: Discussion of the process of our experimentation and the software tools used.
- Results and Analysis: Description of the results of experimentation and analysis of those results.
- Conclusion: Putting the observations of our research in perspective.
- Future work: Suggestions for possible future work drawing from this research.

CHAPTER 2

BACKGROUND

This chapter includes a thorough discussion of readability, an explanation of popular readability formulas and a discussion of the elements of statistical language modeling that underlie our formula for readability.

2.1 Readability: Definitions, factors, measurement and prediction

In this section, we note the popular definitions of readability, discuss the factors affecting readability and some of the different ways of measuring and predicting readability.

2.1.1. Definitions of readability

Edgar Dale and Jeane Chall define readability as "The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting." [3]

A similar, more concise definition is due to G. Harry McLaughlin reads "the degree to which a given class of people find certain reading matter compelling and comprehensible." [4]

2.1.2. Factors influencing readability

Many factors influence the extent to which a given class of people find certain reading matter compelling and comprehensible. It can be expected that subject matter of the text would be a very important factor. Legibility of text would be important as well as the choice of words. They are indeed shown to affect readability. In-fact, in an influential research [5], William S. Gray and Bernice Leary has identified as many as 228 different variables that affect readability. They also classified those 228 variables into the following four different classes:

(a) Variables associated with Content

Content refers to the subject matter of the text.

(b) Variables associated with Style

Style has to do with the types of sentences and words used in the text.

(c) Variables associated with Format

Format relates to aspects of visual presentation like typography and page layout.

(d) Variables associated with Features of Organization

Such features as headings, paragraphs used to organize the ideas of the text.

Gray and Leary (1935) found that Content was the most important factor affecting readability, followed closely by Style.

2.1.3. Measurement and prediction of readability

A reader-text mismatch (For example, assigning a selection from the unabridged "Othello" for a 3rd grade reading exercise) can result in the user failing to use or ignoring the text [5]. To avoid mismatch, educators would like a tool to check if a given text would be readable by its intended audience. Inventing such tools has been the primary focus of readability research for the past 90 years.

2.1.3.1 Measuring readability

There are three widely used methods to measure readability. They are:

(a) Judgments

(b) Comprehension tests

(c) The Cloze Procedure

Expert judgments were the earliest way of matching readers to texts. Judgments have been used for ranking readability of texts as well as providing estimates for different readability factors. Readability rank given to the text is some representation of the minimum education a person is required to have to comprehend that text, which is typically the U.S school grade. Criticisms of this method of measuring readability

include the difficulty in selecting a sufficiently large group of raters and concerns about reliability and generalizability of results [6].

Comprehension tests have also been a popular method of choice to measure subject's understanding of passages. A typical comprehension test is a multiple-choice or other kind of objective test. Readability of text is measured by the how well a certain group of people performed on the comprehension test. As an example, a rule of readability measurement can be: "if 50% of 3rd grade subjects got 50% of the answers on the test right, then the test is considered to be at 3rd grade level". The use of comprehension tests for measuring readability has been criticized because of possible biasing effects of question formulation and also because of the high costs of developing and validating the tests [6].

The Cloze Procedure is a highly reliable, easily constructible measure of readability. This procedure consists of subjects guessing words intentionally omitted from the text whose readability is being measured. Words are omitted according to a rule, such as "every nth word" rule. Readability of the text is measured by the average number of correct guesses the subjects made on the Cloze test. Results of the Cloze Procedure agreed well with the results from comprehension tests and expert ratings [7].

2.1.3.2 Predicting readability

Using the readability variables identified by Gray and Leary [4], researchers have tried to formulate an estimate of text readability as a function of those variables. In this effort, they chose to ignore variables of Content, Format, Features of

Organization while emphasizing the variables of Style. This is due to the supposed importance of Style variables and lack of proper statistical methods to count the variables of other factors of readability [6].

Among the style variables, "vocabulary load" is considered the most important indicator of reading difficulty [7]. Vocabulary load is usually measured by word length or frequency of the words. Vocabulary load is also referred to as "lexical complexity". Next to vocabulary load, sentence structure is the best indicator of reading difficulty [4]. Sentence structure is usually measured by the average sentence length.

Indeed, most of the popular readability formulas estimate readability as a function of variables denoting semantic difficulty (i.e vocabulary load) and syntactic difficulty (i.e difficulty due to sentence structure).

2.1.4. Readability formulas

Readability formulas are an analytical way to predict readability. Popular readability formulas are based on extensive research and their predictions correlate very well with the results of the actual readability measurements of expert judgments, comprehension tests and the Cloze Procedure. We discuss some of the popular readability formulas in this section.

2.1.4.1 The Flesch Reading Ease formula

Flesch Reading Ease scores readability on a scale on 0 to 100 using the following equation:

Score = $206.835 - 1.015 \times$ Average Sentence Length $- 84.6 \times$ Average Syllables per word

Where Average Sentence Length = Number of words in the text/Number of sentences in the text

Average Syllables per word = Number of Syllables in the text/Number of words in the text

Score ranges from 0 to 100, with 0 corresponding to greatest reading difficulty and 100 corresponding to least reading difficulty. Table1 provides interpretation of the Flesch Reading Ease Score.

Table 1: Interpreting the Flesch Reading Ease score

Reading Ease Score	Description	Predicted Reading Grade	Estimated Percentage of U.S Adults
0-30	Very difficult	College graduate	4.5
30-40	difficult	$13^{th} - 16^{th}$ grade	33
50-60	fairly difficult	$10^{th} - 12^{th}$ grade	54
60-70	standard	$9^{th} - 8^{th}$ grade	83
70-80	fairly easy	7^{th} grade	88
80-90	easy	6^{th} grade	91
90-100	very easy	5^{th} grade	93

2.1.4.2 The Flesch Kincaid Grade Level formula

The Flesch Kincaid Grade Level formula predicts readability using an equation similar to that used to calculate the Flesch Reading Ease score. However, the Flesh Kincaid Grade Level function ouputs a U.S school grade level that is indicative of the reading difficulty of the text.

Flesch Kincaid Grade Level = $0.39 \times$ Average Sentence Length + $11.8 \times$ Average Syllable per word – 15.59

A Flesch Kincaid Grade Level of 8.1 indicates that the text is readable by an average 8^{th} grader or a person with 8^{th} grade reading ability.

2.1.4.3 The Dale-Chall formula

The Dale-Chall Formula is based on the average sentence length and the percentage of words not appearing in a list of 3,000 words, 80 percent of which are generally known to fourth grade children.

The Dale-Chall "Raw Score" is given by

Raw Score = $0.0496 \times$ Average Sentence Length + $0.1579 \times$ Percent Difficult Words +3.6365

Percentage of difficult words is the percentage of words in the text that do not appear in the Dale-Chall list of 3000 words.

Raw Score is converted to school grade intervals using the conversion scheme shown in Table 2.

Table 2: Dale-Chall Raw Score to Grade Interval conversion

Raw Score	Grade lnterval
4.9 and below	4^{th} grade and below
5.0 – 5.9	5^{th} – 6^{th} grade
6.0 – 6.9	7^{th} – 8^{th} grade
7.0 – 7.9	9^{th} – 10^{th} grade
8.0 – 8.9	11^{th} – 12^{th} grade
9.0 – 9.9	Grades 13 through 15 (college)
10 and above	Grade 16 and above (college graduate)

2.1.4.4 The Gunning Fog Index

The Gunning fog Index also outputs reading grade level of a given text. Grade level is given by the equation

Grade Level = 0.4 × (Average Sentence Length + Number of hard words)

A “hard word” is defined as a word that is more than two syllables long.

2.1.4.5 The Fry Readability Graph

Estimation of text readability using the Fry Readability Graph is described in the following algorithm:

(a) Select samples of 100 words from the text

(b) On the Y axis of the Fry Graph, plot the average sentence length of the samples

(c) On the X axis of the Fry Graph, plot the average word length

(d) The zone on the graph that includes a point (corresponding to a sample) shows the grade score associated with that sample. Take grade scores associated with at least three points on the graph and average them to get the average grade level associated with the entire text

Shown in Figure 1 is a Fry Graph. Scores that appear in the shaded areas are invalid [8].

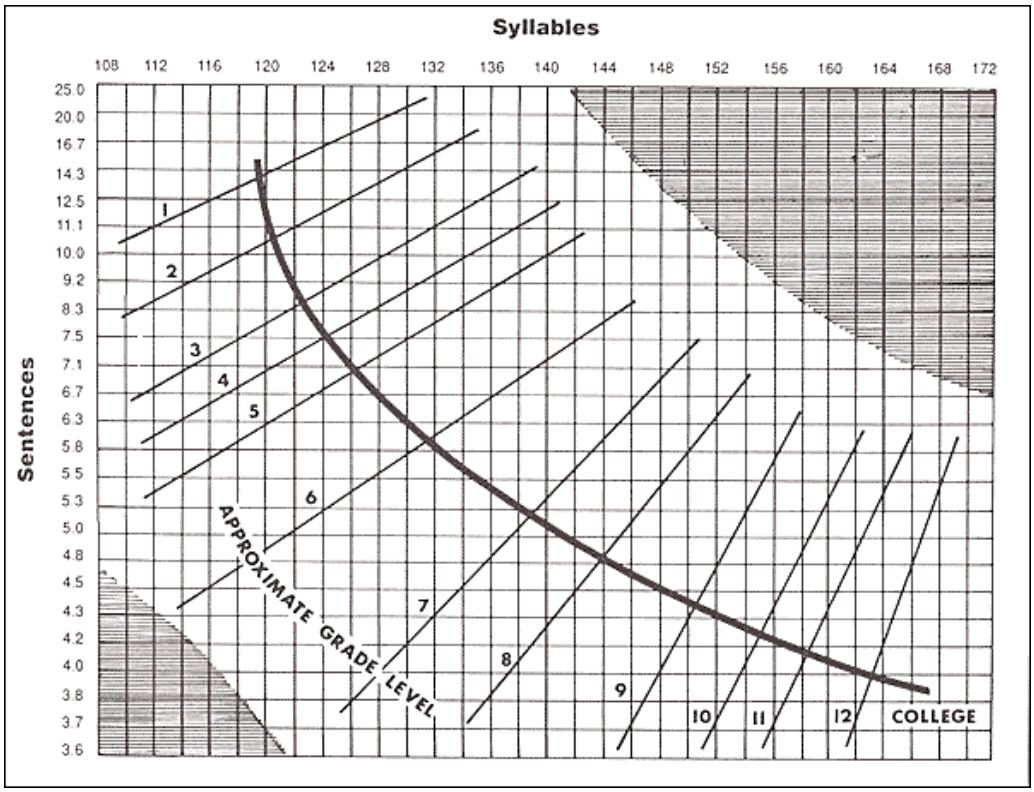

Figure 1: A Fry Graph

2.1.4.6 The SMOG Index

SMOG(Simple Measure Of Gobbledygook) Index outputs the U.S school grade level that is indicative of the number of years of education required to understand the input text.

Grade Level = $\sqrt{total\_complex\_words \times 30 / total\_sentences} + 3$

Complex word is defined as a word that is 3 or more syllables long.

2.2 n-gram models, Entropy

In this section we introduce the elements of statistical language modeling that underlie our new approach to predicting readability of text.

2.2.1 n-gram models

Suppose we have a Language L. We want to know the probability of a string of words $w_1 .. w_z$ ($w_k \in L$) i.e we want to compute $P(w_1 .. w_z)$

By the chain rule of probability,

$$P(w_1 .. w_z) = P(w_1) \times P(w_2 | w_1) \times P(w_3 | w_2\ w_1) \times P(w_4 | w_3\ w_2\ w_1) \times \ldots . P(w_z | w_{z-1} \ldots w_1)$$

However, if the string of words $w_1 .. w_z$ is long, the computation of terms like $P(w_z | w_{z-1} \ldots\ w_1)$ becomes infeasible because it requires a huge number of word sequences from the language L. To get around this problem, we approximate the probability of a word given all previous words in the sequence by the probability of a word given n -1 previous words in the sequence. That is, we approximate $P(w_y | w_{y-1} \ldots w_1)$ by $P(w_y | w_{y-1} \ldots w_{y-(n-1)})$. This statistical model for word prediction is called an n-gram model. n=2 corresponds to a "bigram model", where the context associated with a word in a sequence is only the word that immediately precedes it.

2.2.2 Smoothing

Training n-gram models on a text corpus involves learning probabilities of n-grams from the corpus. However, even a sizable corpus isn't big enough for learning all the possible n-grams. Hence, in a typical n-gram training setting, we see a lot of zero probability n-grams that should really have some non-zero probability. In a process called smoothing, we reevaluate some of the zero probability n-grams and assign them non-zero values [8]. Smoothing is also referred to as Discounting.

The maximum likelihood estimate for the probability of occurrence of an n-gram which occurred r times out of a possible R is

$$P(E) = r/R$$

In a sparse corpus, the estimate for probability is biased high for observed n-grams and low for unseen n-grams. To offset this bias, we redistribute some probability mass from the observed n-grams to the unobserved n-grams, by reducing the counts of observed n-grams by a discount coefficient, dr. The reduced count r* is given by

$$r^* = r \times d_r$$

and the modified probability estimate is

$$P(E) = r^*/R$$

The remaining probability mass that has been discounted from the observed n-grams is allocated to unseen n-grams.

In the following sections, we describe some popular discounting algorithms

2.2.2.1 Absolute discounting

In this type of smoothing, a constant c is subtracted from each of the counts of observed n-grams[8].

$$d_r = (r-c)/r$$

2.2.2.2 Linear discounting

In Linear Discounting, count corresponding to each distinct n-gram is reduced by a value proportional to the count[8]. That is

$$d_r = 1 - \alpha$$

One possible value of α for an n-gram is

$$\alpha = n_1/R$$

Where n_1 is the count of that n-gram and R is the number of words in the training corpus

2.2.2.3 Witten-Bell discounting

This scheme of discounting is based on the idea that we can estimate the probability of unseen events based on the counts of observed events.

Total probability mass assigned to unseen n-grams is

$$P = T/(N+T)$$

Where T is the number of distinct n-grams and N is the total number of n-grams. P gives the maximum likelihood estimate of the occurrence of a new n-gram. [8]

The probability mass p can be distributed equally among all the unseen n-grams. So if m be the number of n-grams with zero probability, then

$p_i = T/(m \times (N+T))$ where p_i is the probability of an unseen n-gram I.

The total probability mass assigned to unseen n-grams should be discounted from the probability mass of observed n-grams. Accordingly, probabilities of observed n-grams are renormalized as

$p_j = c_j/(N+T)$ where c_j is the count of an observed n-gram J.

2.2.3 Entropy

Entropy is a measure of information widely used in Information theory and computational linguistics. Intuitively, Entropy is the lower bound on the number of bits used to encode a piece of information in the optimal coding scheme [8]

2.2.3.1 Formal definition of entropy

Considering a language L as a stochastic process that produces a sequence of words, its Entropy Rate (per-word-entropy) is defined [8] as

$$H(L) = \lim_{n\to\infty} 1/n \sum p(w_1 w_2 \ldots w_n) \log p(w_1 w_2 \ldots w_n) \text{ where } w_k \in L$$

Where $p(w_1 w_2 \ldots w_n)$ is the probability of the sequence $w_1 w_2 \ldots w_n$ being generated by L.

According to the Shannon-McMillan-Breiman theorem [8], for a language that is stationary (probabilities associated with word sequences are independent of time) and ergodic (implying a zero probability for the event of no sequence ever recurring),

$$H(L) = \text{Lim}_{n\to\infty} \; -1/n \; \log p(w_1 w_2 \ldots w_n)$$

2.2.3.2 Cross-entropy

Estimation of entropy of a language can provide information about the predictability of that language. A way of estimating the entropy of language is the 'cross-entropy'.

Suppose we have a sequence of words $S = w_1 w_2 \ldots w_n$. We do not know the actual probability distribution p that generated S. For the estimation of the entropy of p, we use some m, which is a model of p. The cross-entropy of m on p is defined [8] by

$$H(p,m) = \text{Lim}_{n\to\infty} 1/n \sum p(w_1 w_2 \ldots w_n) \log m(w_1 w_2 \ldots w_n)$$

Where $p(w_1 w_2 \ldots w_n)$ is the probability of the sequence of words $w_1 w_2 \ldots w_n$ being generated by p and $m(w_1 w_2 \ldots w_n)$ is the probability of $w_1 w_2 \ldots w_n$ being generated by m.

According to the Shannon-McMillan-Breiman theorem [8], for a stochastic process that is stationary and ergodic,

$$H(p,m) = \lim_{n\to\infty} -1/n \log m(w_1 w_2 \ldots w_n)$$

It should be noted that the cross entropy H(p,m) is an upper bound on the entropy H(p).

$$H(p,m) <= H(p)$$

The more accurately m models p, the closer the value of H(p,m) is to H(p).Given 2 models m1 and m2 of p, the one that has the lower cross-entropy with p is the better model of p [8].

2.2.3.3 Perplexity

Perplexity of a language (or any stochastic process) L is defined as 2^H, where H is the entropy rate of L. Intuitively, perplexity is average 'branching factor' i.e the weighed average number of choices the language L has for generating a word in a sequence of words.

Cross-perplexity of a stochastic process p with a process m is defined as $2^{H(p,m)}$ where H(p,m) is the cross-entropy of p with m.Cross-perplexity can be used to measure how well a statistical model matches a test corpus.

Among 2 models m1 and m2 of p, the model that gives the lowest cross perplexity with p is the better model of p [8].

2.3 Part-of-speech tagging

Part-of-Speech [POS] Tagging is the process of associating part of speech tags (like verb, noun etc) with the words of a text. A software that does part of speech tagging is called the POS tagger. POS-tagging involves assigning POS tag to a word based both on the definition of the word and on the relationship of the word with words adjacent to it.

POS taggers use probabilistic models to assign POS tags to words. POS taggers learn the probabilities of different word sequences by training on an annotated corpus of text. This probability information is used to infer the most likely POS tag for a word in a particular context.

For example, once an adjective is seen - and from the corpus probabilities we know the next word is a noun 50% of the time, an adjective 30% of the time, and an adverb 20% of the time - it can be decided that "guess" in "educated guess" is far more likely to be a noun than an adverb or an adjective. This example shows usage of bigram probabilities in predicting the most likely tag for a word in a context. In reality, probabilistic POS taggers learn and use probabilities of much longer word sequences.

2.3.1 Tagset

Different POS taggers use different sets of POS tags. For our experimentation, we used a tagset that is based on the Penn Treebank Tagset [9]. The Penn Tree Bank Tagset is listed in Table 3.

Table 3: The Penn Treebank Tagset

Tag	Description	Example
CC	Conjunction, coordinating	and, or
CD	Adjective, cardinal number	3, fifteen
DET	Determiner	this, each, some
EX	Pronoun, existential there	there
FW	Foreign words	gracias
IN	Preposition / Conjunction	for, of, although, that
JJ	Adjective	happy, bad
JJR	Adjective, comparative	happier, worse
JJS	Adjective, superlative	happiest, worst
LS	Symbol, list item	A, A.
MD	Verb, modal	can, could, 'll
NN	Noun	aircraft, data
NNP	Noun, proper	London, Michael
NNPS	Noun, proper, plural	Australians

Table 3 - continued

NNS	Noun, plural	women, books
PDT	Determiner, prequalifier	quite, all, half
POS	Possessive	's, '
PRP	Determiner, possessive second	mine, yours
PRPS	Determiner, possessive	their, your
RB	Adverb	often, not, very, here
RBR	Adverb, comparative	faster
RBS	Adverb, superlative	fastest
RP	Adverb, particle	up, off, out
SYM	Symbol	*
TO	Preposition	to
UH	Interjection	oh, yes, mmm
VB	Verb, infinitive	take, live
VBD	Verb, past tense	took, lived
VBG	Verb, gerund	taking, living
VBN	Verb, past/passive participle	taken, lived
VBP	Verb, base present form	take, live
VBZ	Verb, present 3SG -s form	takes, lives
WDT	Determiner, question	which, whatever

Table 3 - continued

WP	Pronoun, question	who, whoever
WPS	Determiner, possessive & question	whose
WRB	Adverb, question	when, how, however
PP	Punctuation, sentence ender	., !, ?
PPC	Punctuation, comma	,
PPD	Punctuation, dollar sign	$
PPL	Punctuation, quotation mark left	``
PPR	Punctuation, quotation mark right	"
PPS	Punctuation, colon, semicolon, ellipsis	:, ..., -
LRB	Punctuation, left bracket	(, {, [
RRB	Punctuation, right bracket	), },]

2.3.2 POS-skeleton

We define the term POS-Skeleton of a text T as the string of POS tags obtained by replacing each word/punctuation symbol in T by its corresponding POS tag. We denote POS-Skeleton of text T by POS-Skeleton(T).

As an illustration, for the text

T = Take the night off from cooking and come out for some free food, drink and music. All residents of BorderTrail are invited.

POS-Skeleton(T) = VB DET NN IN IN NN CC VB IN IN DET JJ NN PPC NN CC NN PP DET NNS IN NNP VBP VBN PP

We define POS-Skeleton corresponding to a text sentence as a POS-sentence.

2.4 The CMU Statistical Language Modeling Toolkit

The CMU Statistical Language Modeling Toolkit (CMU-SLM) is a set of UNIX software tools for the creation and testing of n-gram language models. An exhaustive discussion of the statistical language modeling techniques supported by the CMU-SLM and the tools it encompasses is provided in [10]. In this section, we discuss some of the elements of n-gram modeling relevant to our experiments and the support CMU-SLM provides for those elements, and then proceed to describe a usage scenario involving the CMU-SLM tools.

2.4.1 Elements of n-gram modeling

2.4.1.1 Vocabulary

Vocabulary is the set of words associated with a language. Tasks associated with construction/evaluation of language models require specification of the vocabulary. The CMU-SLM provides different ways to handle out-of-vocabulary

words that may occur during training or testing. Specifically, the CMU-SLM allows the vocabulary to be one of Closed Vocabulary and Open Vocabulary.

Closed Vocabulary means out-of-vocabulary(OOV) words are not allowed. Any such words in the training or test data will cause an error. Closed Vocabulary model is suitable for an environment where OOVs are guaranteed not to occur. Open Vocabulary allows for OOVs. OOVs are all mapped to the same word in the language. That 'same word' can be any word in the language and should be specified.

2.4.1.2 Context cues

Context cues are the markers that indicate events like sentence boundaries. They provide information to the language model and hence aid in word prediction. However they should not be predicted by the model itself. CMU-SLM provides an option to specify a set of context cues.

2.4.1.3 Back-off

In the calculation of the probability of word Y given a context, we may wish to disregard the context before a certain word in vocabulary or a context cue. For instance, in calculating the probability of Y in the context "X EOS Y"(EOS is a context cue indicating end of sentence), we may not wish to calculate the probability of Y based on the full context i.e. P(Y/EOS X). We may wish to "back-off" from the sentence boundary and predict the probability of Y in the given context as

1. P(Y/EOS) or

2. P(Y)

The first way of backing off, in which the word/context cue we back off from is included in the context, is called “inclusive backoff”. The second way of backing off, in which the word/context cue we back off from is excluded from the context “exclusive backoff”. Both kinds of backoff are supported by the CMU-SLM. The CMU-SLM allows optional backing off from context cues. It also allows specification of a set of words in the language from which the language model should always back off.

2.4.1.4 Discounting

Discounting is described in section 2.2.2. CMU-SLM offers support for Linear discounting, Absolute discounting, Witten-Bell discounting and Good-Turing discounting.

2.4.2 Using the CMU-SLM toolkit

In this section, we describe the CMU-SLM tools used in our experimentation and also a typical usage scenario.

2.4.2.1 Tools

In this section, we discuss the CMU-SLM tools (implemented as unix utilities) that we used in our experiments.

Text2idngram

Function: Converts each word into an id (a short integer) enabling storage of more n-grams

Input : Text stream , Vocabulary file, vocabulary type(open/closed), gram size (n value)

Output : File containing list of every id-ngram that occurred in the text along with its frequency of occurrence

Idngram2lm

Function : converts the id-ngram list to language model format

Input : An id-ngram file, vocabulary file, context cues file, gram size, discounting strategy, backoff strategy, vocabulary type, output language model file name

Output : Language model

Evallm-perplexity

Function : Calculates the perplexity of text with respect to a language model

Input : Language Model, test text

Output : Perplexity of the test text with respect to the language model

2.4.2.2 Usage scenario

Figure 2 depicts a usage scenario involving the CMU-SLM tools.

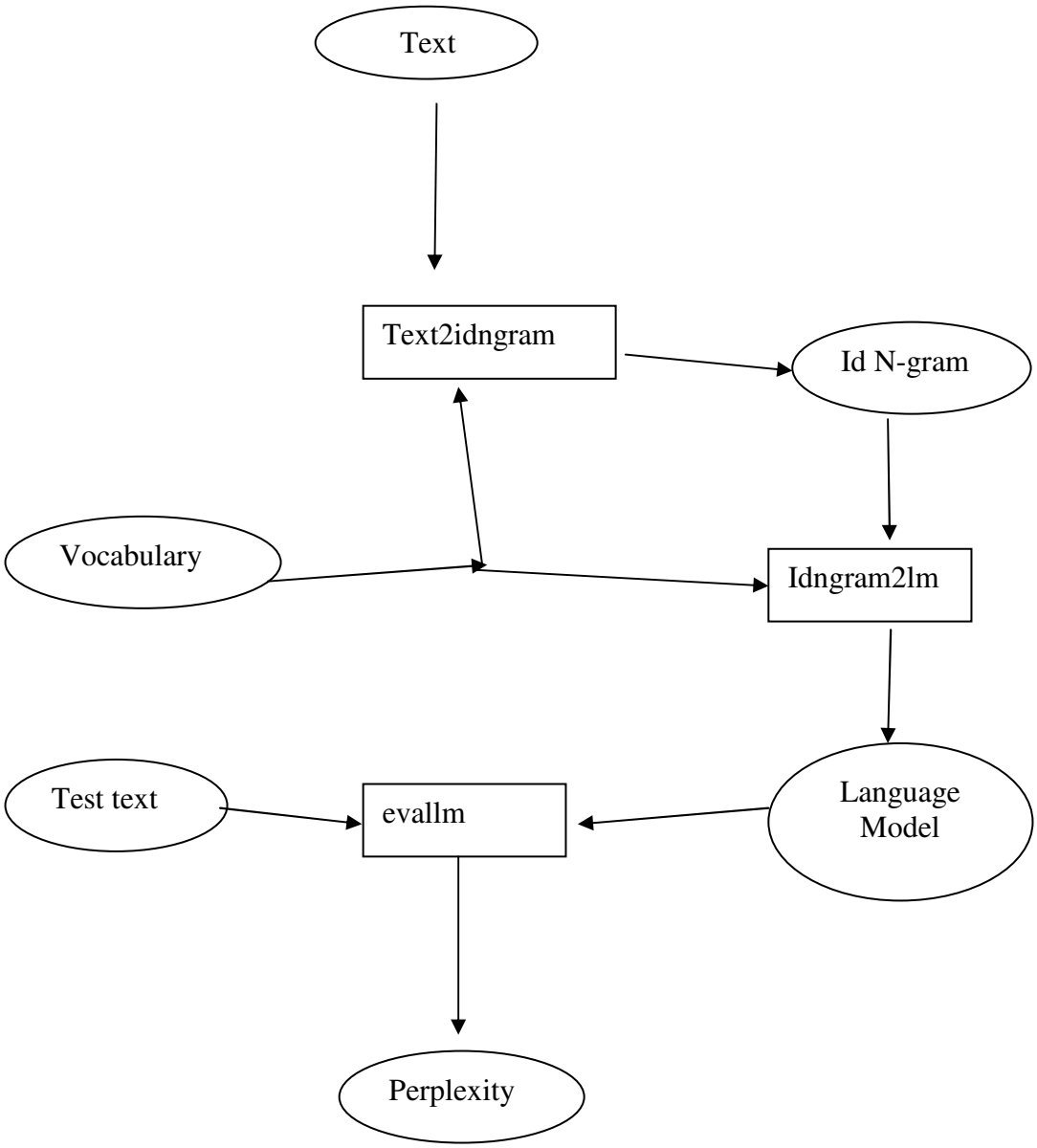

Figure 2: Usage scenario involving the CMU-SLM tools

CHAPTER 3

HYPOTHESIS

In this section, we advance our hypothesis for a new metric for syntactic complexity and a new readability formula.

3.1 A new metric of syntactic complexity

Average sentence length is the most widely used measure of syntactic complexity i.e the reading difficulty due to syntax. However, we argue that average sentence length, as a measure of syntactic complexity, has the following short comings:

1) It is not always true that longer sentences are harder to read.

The following example [10] that contrasts a longer but more readable sentence with a couple of shorter ones:

(a) He is the defendant. He is 15 years old. Someone says he stole from a shop.

(b) The defendant is a 15-year-old accused of shoplifting.

Note that Average sentence length of (a) is lesser than that of (b).

2) Sentences in passive voice are generally, though not always, considered harder to read.

(a) The entrance exam was failed by over one third of the applicants to the school

(b) Over one third of all the applicants to the school have failed the entrance exam

Sentence (b) is easier to read than sentence (a), but Sentence Length metric cannot capture this distinction in reading difficulty.

3) In compound sentences, embedding a subordinate clause within the main clause can make a sentence harder to read

(a) Industrial spying, *owing to the growing use of computers to store and process corporate information,* is increasing rapidly.

(b) Industrial spying is increasing rapidly *owing to the growing use of computers to store and process corporate information.*

Sentence (a), with an embedded subordinate clause, is harder to read than sentence (b). Again, the metric of average sentence length cannot account this difference in reading difficulty.

The inadequacy of average sentence length as a measure of syntactic complexity calls for an alternative metric that can account for the perceived difference in syntactic complexity between different types of sentences. Posited on the belief that part-of-speech structure of text represents syntactic complexity of the text more accurately than

average sentence length, we propose a new index of syntactic complexity in our hypothesis:

- POS-skeleton corresponding to a text represents the syntax of that text. Different texts having the same POS-skeleton present the same level of difficulty-due-to-syntax for reading.
- Syntax of text is associated with the school grade-level. POS skeletons corresponding to grade-level text corpora are models for grade-level syntax. We call these models 'grade-level syntax models'.
- We define a new index of syntactic complexity of text: the Syntax Grade. Syntax Grade of a text T is the grade corresponding to the grade-level syntax model that has the highest probability of generating the syntax of T. In other words, Syntax Grade of a text indicates the school grade-level the syntax of that text is most likely at.

.

3.2 A new readability formula

We propose a new readability formula based on the following premises:

- Reading difficulty (or readability) correlates with school grade level.
- Readability is a function of syntax load and vocabulary load. Our readability formula estimates reading grade based on the Syntax Grade (measure of syntax load) and some measure of vocabulary load.

CHAPTER 4

EXPERIMENTATION

To implement a system that computes Syntax Grade of a text as stated in our hypothesis in section 3.1, we need to build grade-level syntax models from grade-level text corpora. The first step of our experimentation is to collect grade-level text corpora.

4.1 Grade-level text corpora

For our task of building Grade-level syntax models, we need texts that are known to be at a particular Grade-level. We collected texts that appeared as passages on reading comprehension tests conducted by various U.S state education boards for various school grades [All of the texts we used for training/testing are released for public use by state education boards of different U.S states]. However, since enough data is not available to build a syntax model for each school grade, we clubbed grades in twos so a model could be built for each such 'Grade couple'. Table 4 shows the different Grade couples and the associated individual school grade pairs.

Table 4 : grades – Grade couple mapping

School grades	Grade couple label
3,4	4
5,6	6
7,8	8
9,10	10
11,12	12

Henceforth, we will refer to Grade couple label simply as Grade.

Sets of 20 texts each corresponding to a Grade are used in building our Grade-level syntax models. Each such set is called "Grade-level training corpus". Table 5 shows the average size of an individual text of different Grade-level training Corpora.

Table 5: Average size of Grade-level training texts

Grade	Average size(in KB) of a text used in training
4	4.0
5,6	4.6
7,8	5.0
9,10	5.9
11,12	6.8

4.2 Building Grade-level syntax models

POS-Skeletons are extracted from Grade-level text corpora to form the Grade-level syntax corpora. Grade-level syntax models are built from Grade-level syntax corpora using the CMU-SLM.

4.2.1 Syntax extraction

POS-skeletons are extracted using the Perl module Lingua::EN::tagger-0.13. The tagger uses the Penn Treebank tagset to POS-tag text. The Penn Treebank tagset was described in Section 2.3.1. However, the Penn Treebank tagset has tags representing variations of basic part of speech elements that are too fine for readability purposes. So we condensed the Penn Treebank tagset to 24 tags and programmed the tagger to mark up text with those 24 tags. We call those 24 tags " The Readability POS tagset ". Tags from the Readability POS tagset are listed in Table 6.

Table 6: The Readability POS tagset

Tag	Description	Example
ADJ	Adjective	happy,worst,half
ADV	Adverb	often, very, fastest, off
CD	Adjective, cardinal number	3, fifteen
DET	Determiner	this, each, some

Table 6 - continued

EX	Pronoun, existential there	there
FW	Foreign words	merci
IN	Preposition / Conjunction	for, of, although, that
LS	Symbol, list item	A, A.
MD	Verb, modal	can, could, 'll
NN	Noun- singular, pl	aircraft, women
NNP	Proper Noun – singular, pl	London, Americans
PSV	Possessive	‘s, ‘, mine, your, their
SYM	Symbol	*
UH	Interjection	oh, yes, mmm
VRB	verb, all forms	take, took, taking, taken,takes
WHT	question, all forms	which,what,who,whose,how,whoever
PP	Punctuation,	., !, ?

Table 6 - continued

	sentence ender	
PPC	Punctuation, comma	,
PPD	Punctuation, dollar sign	$
PPL	Punctuation, quotation mark left	``
PPR	Punctuation, quotation mark right	''
PPS	Punctuation, colon, semicolon, ellipsis	:, ..., -
LRB	Punctuation, left bracket	(, {, [
RRB	Punctuation, right bracket	), },]

4.2.2 Construction of Grade-level syntax models

Our training data contains 100 POS-skeletons (corresponding to 100 training texts), with each of the 5 Grades accounting for 20 of those POS-skeletons. We call

each of those 5 Grade-level POS-skeleton sets a Grade-level syntax corpus. From the Grade-level syntax corpora, Grade-level syntax models are built using the CMU-SLM. The process of building Grade-level syntax models from text corpora is described in the Figure 3.

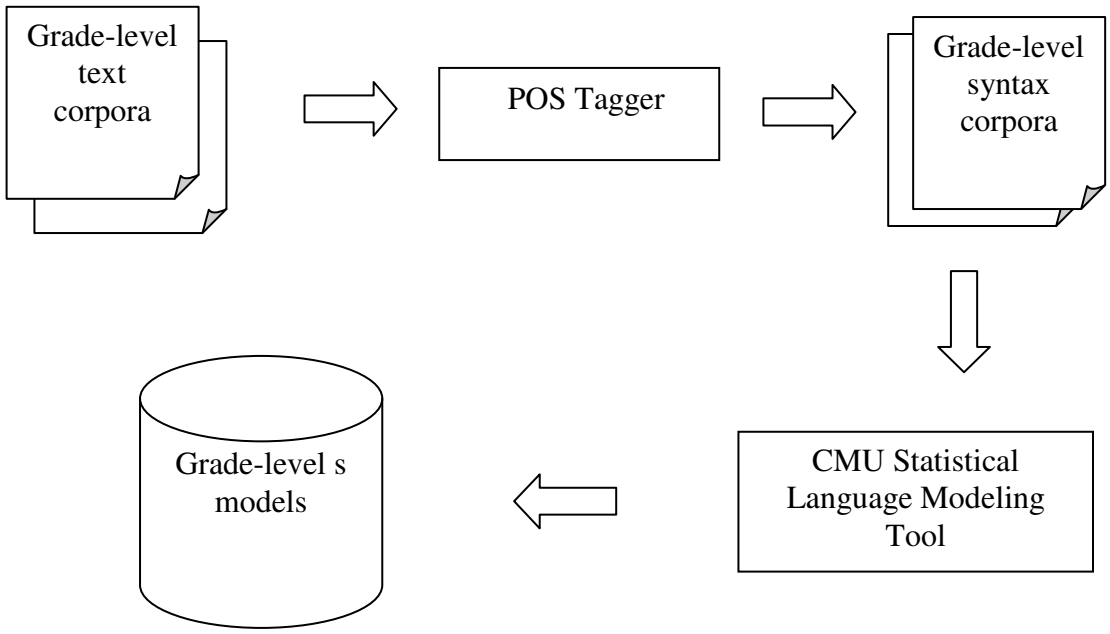

Figure 3: Construction of Grade-level syntax models

4.3 Computing Syntax Grade

We collected 33 pre-classified texts for testing purpose. Computing Syntax Grade involves extracting syntax (POS-skeleton) of each test-text and calculating the perplexities of POS-skeleton of the test-text with respect to each of the Grade-level syntax models. Syntax Grade of a test-text is the Grade associated with the Grade-level

syntax model that gives the lowest perplexity with POS-skeleton of that test-text. This process is illustrated in Figure 4.

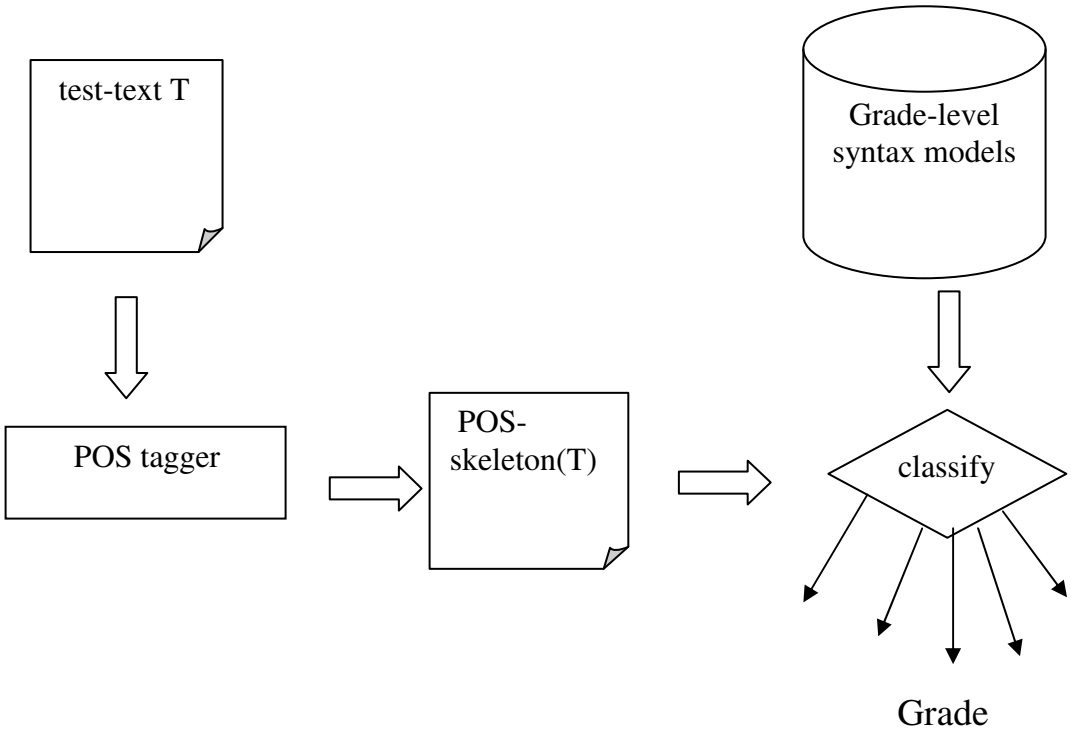

Figure 4: Computing Syntax Grade

4.4 Parameters of training and evaluation

4.4.1 Vocabulary

Our vocabulary is the set of tags from the Readability POS tagset, listed in section 4.2.3. Our model of vocabulary is "closed" which means that an Out-Of-Vocabulary word in training/evaluation causes an error.

4.4.2 Gram size

Our intention is to use n-grams that are long enough to cover any POS-sentence(POS-skeleton corresponding to a sentence in text) in our Grade-level syntax corpora and the in POS-skeletons corresponding to test-texts. We use a gram size of 139 as that is the length of the longest POS-sentence encountered in our training and testing.

4.4.3 Discounting

We used Witten-Bell scheme of discounting. Witten-Bell discounting is described in Section 2.4.1.4

4.4.4 Back-off

We discussed the concept of back-off in section 2.4.1.3. In the computation of perplexity, we execute an inclusive back-off at sentence boundaries (In POS-skeletons, beginning of a sentence is indicated by the tag SS). In other words, in predicting the probability of a word in a context, we disregard the context before the beginning of the current sentence.

4.4.5 Context cues

A context cue is defined in section 2.4.1.4. As the only context cue, we have SS, a special tag that indicates the start of a POS-sentence.

CHAPTER 5

RESULTS AND ANALYSIS

In this chapter, we show the results of our experimentation, deduce a new readability formula and analyze it comparatively with other readability formulas.

5.1 Results

For our 33 test-texts, we computed the following parameters:

- Syntax Grade
- Word Length in syllables
- Percentage of complex words (words greater than 3 syllables in length)
- Average sentence length
- Percentage of mono-syllable words
- Dale-Chall percentage (percentage of words not appearing on the Dale-Chall word list)
- Percentage of difficult sentences (sentences longer than 20 words)

The results of our experimentation are shown in Table 7.

Table 7: Results of experimentation

Test text	Orig Grade	Syntx Grde	Avg word length	%Cmplx words	Avg Sentce Length	%Mono syllable	Dale Chall %	%Diff sntncs
Test-1	4	4	1.4	7.1	6.8	74.6	20.1	0
Test-2	4	8	1.4	9.4	25.7	72.2	9.6	46.7
Test-3	4	4	1.4	6.9	12	72.7	9.3	8.6
Test-4	4	4	1.4	6.4	14.4	78.9	6.9	19.2
Test-5	4	4	1.2	2.1	8.3	80.5	5.8	4.3
Test-6	4	4	1.4	2	8.7	70.3	6.1	0
Test-7	4	4	1.5	11.7	12.4	72.2	16.1	10.3
Test-8	6	4	1.2	5	13.7	81.4	5.3	24.1
Test-9	6	8	1.4	4.2	17.9	70.6	14	30
Test-10	6	8	1.6	17	14.7	65.3	19	20
Test-11	6	10	1.4	6.6	13.8	68.7	13.3	8.3
Test-12	6	10	1.6	9.4	8.7	65.5	22.3	0
Test-13	6	6	1.5	12.5	14.1	65.9	21.3	9.5
Test-14	8	8	1.6	17	21.5	58.9	30.7	58.8
Test-15	8	10	1.6	12.7	20.4	66.5	19.4	40.9
Test-16	8	6	1.6	11.5	16.9	57.8	24.5	25
Test-17	8	8	1.6	14.1	18.2	62.9	21.3	37.5
Test-18	8	4	1.4	6.1	12.1	72.4	8.6	12.7
Test-19	8	8	1.6	14	18.3	60.7	23.6	33.3
Test-20	8	6	1.6	13.1	18.8	61.8	24.1	42.3
Test-21	8	6	1.5	11.5	14.1	66.9	20.4	8.3
Test-22	10	8	1.4	6.8	16.5	72.2	15.9	34.4
Test-23	10	6	1.6	18.1	16.3	62.7	20	22.5
Test-24	10	8	1.6	14.8	21.2	62	24.7	51.4
Test-25	10	8	1.4	9.1	14.8	76.6	15.8	22.1
Test-26	10	8	1.6	14.3	18.8	62.1	18.1	36.7
Test-27	10	8	1.6	14.9	19.4	61.5	21.4	35.7
Test-28	12	12	1.7	18.6	25.6	65.1	23.1	75
Test-29	12	12	1.6	15.7	22.7	61.3	23	55.6
Test-30	12	12	1.9	22.2	23	44.9	38.6	66.7
Test-31	12	12	1.7	18.2	16.6	62.2	28.5	27.3
Test-32	12	12	1.7	28.6	21.4	51.2	38	64.7
Test-33	12	10	1.8	21.1	24.4	58.1	32.6	64.3

5.2 Analysis

In this section, we present the correlations of various metrics with the original grade as assigned by the U.S. state education boards. We also come up with a new readability formula, the Readability Index and compare its grade predictions with the grades predicted by some of the popular readability formulas. Table 8 shows correlations of various metrics with original Grade.

Table 8: Correlations of various measures with original Grade

Parameter	Correlation with Original Grade
Percentage of Complex words	0.74
Syntax Grade	0.74
Percentage of Difficult sentences	0.72
Average Word Length	0.72
Dale-Chall percentage	0.72
Percentage of Mono-syllable words	-0.68
Average Sentence Length	0.66

It can be seen that Syntax Grade and percentage of complex words are the best two predictors of readability.

In our hypothesis in section 3.2, we stated that our readability formula should include Syntax Grade as a measure of syntactic complexity and some measure of lexical complexity. The measures of lexical complexity we computed are average word length, percentage of complex words and Dale-Chall percentage. We want to choose one of them and combine it with Syntax Grade to generate our readability formula.

Table 9 shows the correlations of different lexical complexity measures with Syntax Grade. Among the measures of lexical complexity, we chose average word length because it is least correlated with Syntax Grade.

Table 9: Correlations of various lexical complexity measures with Syntax Grade

Measure of Lexical complexity	Correlation with Syntax Grade
Average Word Length	0.68
Dale Chall Percentage	0.69
Percentage of complex words	0.73

5.2.1 The Readability Index

We assume that our new readability formula is of the form:

Raw Score = x × Syntax Grade + y × Average Word Length + z (Equation 4.1) Where x,y,z are constants.

We define Readability Index [RI] to be the function that calculates Raw Score using Equation 4.1 and maps Raw Score to a Grade value using the mapping definition shown in Table 10.

Table 10: RI's scheme for mapping Raw score to Grade

Raw Score Range	RI output (Grade)
Upto 4.5	4
Between 4.5 and 6.5	6
Between 6.5 and 8.5	8
Between 8.5 and 10.5	10
Above 10.5	12

For finding the values of x, y, z we did a gradient descent over the different triplets <original Grade, Syntax Grade, average word length>.

This process yielded the following values for x, y, z:

x = 0.87

y = 5.2

z = -5.9

Equation 4.1 now becomes

Raw Score = 0.87 × Syntax Grade + 5.2 × Average Word Length - 5.9 (Equation 4.2)

5.2.2 Comparing RI with other readability formulas

In this section we compare RI with the Dale-Chall formula, the Flesch-Kincaid Grade Level formula, the SMOG formula and the Gunning Fog Index. For the sake of comparison with RI, we convert the outputs of Flesh-Kincaid Grade Level formula, SMOG formula and Gunning Fog Index to the Grade values according to the mapping defined in Table 10.

5.2.2.1 RI vs The Dale-Chall formula

The Dale-Chall formula is defined in section 2.1.4.3 It is based on the Dale-Chall percentage(the percentage of words not appearing on the Dale-Chall list) and the average sentence length.

Table 11 shows the Grade predictions by Dale-Chall formula and RI on our set of test texts.

Table 11 : RI vs The Dale-Chall formula

Test text	original Grade	Syntax Grade	Word Length	RI Grade	Dale-Chall %	Average Sentence Length	Dale Chall Grade
Test-1	4	4	1.4	4	20.1	6.8	6
Test-2	4	8	1.4	8	9.6	25.7	10
Test-3	4	4	1.4	4	9.3	12	4
Test-4	4	4	1.4	4	6.9	14.4	4
Test-5	4	4	1.2	4	5.8	8.3	4
Test-6	4	4	1.4	4	6.1	8.7	4
Test-7	4	4	1.5	4	16.1	12.4	6
Test-8	6	4	1.2	4	5.3	13.7	4
Test-9	6	8	1.4	8	14	17.9	8
Test-10	6	8	1.6	10	19	14.7	8
Test-11	6	10	1.4	10	13.3	13.8	6
Test-12	6	10	1.6	10	22.3	8.7	8
Test-13	6	6	1.5	8	21.3	14.1	8
Test-14	8	8	1.6	10	30.7	21.5	12
Test-15	8	10	1.6	10	19.4	20.4	10
Test-16	8	6	1.6	8	24.5	16.9	10
Test-17	8	8	1.6	10	21.3	18.2	10
Test-18	8	4	1.4	4	8.6	12.1	4
Test-19	8	8	1.6	10	23.6	18.3	10
Test-20	8	6	1.6	8	24.1	18.8	10
Test-21	8	6	1.5	8	20.4	14.1	8
Test-22	10	8	1.4	8	15.9	16.5	8
Test-23	10	6	1.6	8	20	16.3	8
Test-24	10	8	1.6	10	24.7	21.2	12
Test-25	10	8	1.4	8	15.8	14.8	8
Test-26	10	8	1.6	10	18.1	18.8	8
Test-27	10	8	1.6	10	21.4	19.4	10
Test-28	12	12	1.7	12	23.1	25.6	12
Test-29	12	12	1.6	12	23	22.7	12
Test-30	12	12	1.9	12	38.6	23	12
Test-31	12	12	1.7	12	28.5	16.6	12
Test-32	12	12	1.7	12	38	21.4	12
Test-33	12	10	1.8	12	32.6	24.4	12

Grades predicted by Dale-Chall has a correlation of 0.76 while grades predicted by RI correlate at 0.78 with the original Grades. The predictions of RI deviated from the actual Grade by 0.6 on an average, where the deviations of predictions by Dale-Chall from the actual Grade averaged at 0.72 Grade.

5.2.2.2 RI vs The Gunning Fog Index

The Gunning Fog Index is defined in section 2.1.4.4. It is based on the percentage of complex words(words longer than 2 syllables) and average sentence length. Table 12 shows the Grade predictions by the Gunning Fog Index and RI on our test texts.

Table 12 : RI vs The Gunning Fog Index

Test - text	Orig Grade	Syntx Grade	Avg Word Length	RI Grade	%mono syllables	% cmplx words	Average sentence Length	GF Indx
Test-1	4	4	1.4	4	74.6	7.1	6.8	6
Test-2	4	8	1.4	8	72.2	9.4	25.7	12
Test-3	4	4	1.4	4	72.7	6.9	12	8
Test-4	4	4	1.4	4	78.9	6.4	14.4	8
Test-5	4	4	1.2	4	80.5	2.1	8.3	4
Test-6	4	4	1.4	4	70.3	2	8.7	4
Test-7	4	4	1.5	4	72.2	11.7	12.4	10
Test-8	6	4	1.2	4	81.4	5	13.7	8
Test-9	6	8	1.4	8	70.6	4.2	17.9	10
Test-10	6	8	1.6	10	65.3	17	14.7	12
Test-11	6	10	1.4	10	68.7	6.6	13.8	8
Test-12	6	10	1.6	10	65.5	9.4	8.7	8
Test-13	6	6	1.5	8	65.9	12.5	14.1	12
Test-14	8	8	1.6	10	58.9	17	21.5	12
Test-15	8	10	1.6	10	66.5	12.7	20.4	12
Test-16	8	6	1.6	8	57.8	11.5	16.9	12

Table 12 - continued

Test-17	8	8	1.6	10	62.9	14.1	18.2	12
Test-18	8	4	1.4	4	72.4	6.1	12.1	8
Test-19	8	8	1.6	10	60.7	14	18.3	12
Test-20	8	6	1.6	8	61.8	13.1	18.8	12
Test-21	8	6	1.5	8	66.9	11.5	14.1	10
Test-22	10	8	1.4	8	72.2	6.8	16.5	10
Test-23	10	6	1.6	8	62.7	18.1	16.3	12
Test-24	10	8	1.6	10	62	14.8	21.2	12
Test-25	10	8	1.4	8	76.6	9.1	14.8	10
Test-26	10	8	1.6	10	62.1	14.3	18.8	12
Test-27	10	8	1.6	10	61.5	14.9	19.4	12
Test-28	12	12	1.7	12	65.1	18.6	25.6	12
Test-29	12	12	1.6	12	61.3	15.7	22.7	12
Test-30	12	12	1.9	12	44.9	22.2	23	12
Test-31	12	12	1.7	12	62.2	18.2	16.6	12
Test-32	12	12	1.7	12	51.2	28.6	21.4	12
Test-33	12	10	1.8	12	58.1	21.1	24.4	12

With the actual grades, Readability Index has a correlation of 0.78 where as the Gunning Fog Index has a correlation of 0.65. The Gunning Fog Index seems to overestimate the Grade. A good illustration of this tendency of the Gunning Fog Index is the Grade prediction it makes for Test-7. Test-7 is a text describing “guacamole” for 4^{th} grade. However, since it contains repeated occurrences of polysyllabic (word more than 2 syllables long) words like guacamole and avocado, the Gunning Fog Index predicts a higher Grade for it. But Test-7 also contains a huge percentage of monosyllabic words. Unlike the Gunning Fog Index, RI does not overestimate the Grade of a text with high percentage of monosyllabic and a high percentage of polysyllabic words because it is based on average word length.

5.2.2.3 RI vs The Flesch-Kincaid Grade Level formula

The Flesch Kincaid Grade Level formula is defined in section 2.1.4.2. It is based on the average word length in syllables and average sentence length. Table 13 shows the Grade predictions by the Flesch-Kincaid Grade Level and RI on our test texts.

Table 13: RI vs The Flesch-Kincaid Grade Level formula

Test text	Original Grade	Syntax Grade	Avg Word Length	RI Grade	Average sentence length	Flesch-Kincaid Grade
Test-1	4	4	1.4	4	6.8	4
Test-2	4	8	1.4	8	25.7	12
Test-3	4	4	1.4	4	12	6
Test-4	4	4	1.4	4	14.4	6
Test-5	4	4	1.2	4	8.3	4
Test-6	4	4	1.4	4	8.7	4
Test-7	4	4	1.5	4	12.4	6
Test-8	6	4	1.2	4	13.7	4
Test-9	6	8	1.4	8	17.9	8
Test-10	6	8	1.6	10	14.7	10
Test-11	6	10	1.4	10	13.8	6
Test-12	6	10	1.6	10	8.7	6
Test-13	6	6	1.5	8	14.1	8
Test-14	8	8	1.6	10	21.5	12
Test-15	8	10	1.6	10	20.4	10
Test-16	8	6	1.6	8	16.9	10
Test-17	8	8	1.6	10	18.2	10
Test-18	8	4	1.4	4	12.1	6
Test-19	8	8	1.6	10	18.3	10
Test-20	8	6	1.6	8	18.8	10
Test-21	8	6	1.5	8	14.1	8
Test-22	10	8	1.4	8	16.5	8
Test-23	10	6	1.6	8	16.3	10
Test-24	10	8	1.6	10	21.2	12
Test-25	10	8	1.4	8	14.8	6
Test-26	10	8	1.6	10	18.8	10

Table 13 - continued

Test-27	10	8	1.6	10	19.4	12
Test-28	12	12	1.7	12	25.6	12
Test-29	12	12	1.6	12	22.7	12
Test-30	12	12	1.9	12	23	12
Test-31	12	12	1.7	12	16.6	10
Test-32	12	12	1.7	12	21.4	12
Test-33	12	10	1.8	12	24.4	12

RI has a correlation of 0.78 with the original Grades where as Flesch-Kincaid Grade Level formula has a correlation of 0.70. Predictions of Flesch Kincaid Grade Level correlate with original Grade just as well as those of RI on most of our test texts. However, discrepancies occur when a text has a high average sentence length and low average word length. High average sentence length is not necessarily associated with high Syntax Grade (there is a correlation of 0.66 between average sentence length and Syntax Grade, which is decent but not very strong). So a text with high average sentence length and low average word length can result in varying predictions by RI and Flesch Kincaid Grade Level as evidenced by Test-2. It is interesting to note that correlations of RI and Flesh Kincaid Grade Level with original Grade would have been the same (0.78) had Flesh Kincaid Grade Level, like RI, predicted a Grade of 8 for Test-2.

5.2.2.4 RI vs The SMOG formula

The SMOG formula is defined in section 2.1.4.6. It is based on the number of complex words (words longer than 2 syllables) per sentence. Table 14 shows the Grade predictions by the SMOG formula and RI on our test texts.

Table 14: RI vs The SMOG formula

Test text	Original Grade	Syntax Grade	Average Word Length	RI Grade	#Complex words/sentence	Smog Grade
Test-1	4	4	1.4	4	0.48	8
Test-2	4	8	1.4	8	2.42	12
Test-3	4	4	1.4	4	0.83	8
Test-4	4	4	1.4	4	0.92	8
Test-5	4	4	1.2	4	0.17	6
Test-6	4	4	1.4	4	0.17	6
Test-7	4	4	1.5	4	1.45	10
Test-8	6	4	1.2	4	0.69	8
Test-9	6	8	1.4	8	0.75	8
Test-10	6	8	1.6	10	2.50	12
Test-11	6	10	1.4	10	0.91	8
Test-12	6	10	1.6	10	0.82	8
Test-13	6	6	1.5	8	1.76	12
Test-14	8	8	1.6	10	3.66	12
Test-15	8	10	1.6	10	2.59	12
Test-16	8	6	1.6	8	1.94	12
Test-17	8	8	1.6	10	2.57	12
Test-18	8	4	1.4	4	0.74	8
Test-19	8	8	1.6	10	2.56	12
Test-20	8	6	1.6	8	2.46	12
Test-21	8	6	1.5	8	1.62	10
Test-22	10	8	1.4	8	1.12	10
Test-23	10	6	1.6	8	2.95	12
Test-24	10	8	1.6	10	3.14	12
Test-25	10	8	1.4	8	1.35	10
Test-26	10	8	1.6	10	2.69	12
Test-27	10	8	1.6	10	2.89	12
Test-28	12	12	1.7	12	4.76	12
Test-29	12	12	1.6	12	3.56	12
Test-30	12	12	1.9	12	5.11	12
Test-31	12	12	1.7	12	3.02	12
Test-32	12	12	1.7	12	6.12	12
Test-33	12	10	1.8	12	5.15	12

Results show that SMOG's Grade predictions have a correlation of 0.66 with the original Grades against RI's 0.78. It can easily be observed that SMOG overestimates Grade. SMOG is based on the average number of complex words per sentence. With the SMOG formula, texts with at least 1.01 complex words per sentence get a Grade of at least 10. Among our 33 test-texts, 23 (70%) have value greater than 1.01 for the parameter of complex words per sentence. So 70% of our test-texts are given a Grade of either 10 or 12 by the SMOG formula.

CHAPTER 6

CONCLUSIONS

We draw the following conclusions from analyzing the results of our experimentation:

- Average word length is as accurate a stand-alone predictor of readability as Dale-Chall percentage. There is a correlation of 0.9 between average word length and Dale-Chall percentage and both of them correlate at 0.72 with the original grades.
- Syntax Grade and average sentence length have a modest correlation of 0.66 between them. This clearly shows they do not measure the same thing. Also, the superiority of RI over the formulas based on average sentence length shows that Syntax Grade is a better approximation of structure of text than is average sentence length
- Our hypothesis that a readability formula based on the Syntax Grade will have better prediction accuracy than those based on average sentence length is vindicated by the results of our experimentation

CHAPTER 7

FUTURE WORK

We believe that the concepts developed as part of our work can be applied to the following tasks:

- Authorship attribution

We believe that authors use characteristic sentence structure patterns and that those can be captured in n-gram models. These n-gram structure models can be used to attribute authorship of a text to an author in the same way we associate Syntax Grade to texts.

- Genre identification

We hypothesize that sentence structure characterizes a literary genre (like Information Content, Poetry, Fiction etc). Using Language models representing the sentence structure characteristic of each genre, we can identify the genre of a given text.

REFERENCES

[1] Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad(1993). National Adult Literacy Study, National Center for Educational Statistics.

[2] Klare, G (1975). Assessing Readability. Reading Research Quarterly, Volume 10, Number 1, 62-102 .

[3] Sticht, T ., Caylor, J ., Kern, R ., and Fox, L (1972). Project REALISTIC: Determination of Adult Functional Literacy Skill Levels . Reading Research Quarterly, Volume7, Number 3, 424-465 .

[4] McLaughlin, G. Harry. (1969). SMOG grading: A new readability formula. Journal of Reading, 12(8), 639-646.

[5] William S. Gray and Bernice Leary (1935), What Makes a Book Readable, Chicago University Press

[6] Williams, A ., Siegel, A . and Burkett, J(1974) . Readability of Textual Material - A Survey of the Literature, AFHRLTR-74-29, Air Force Human Resources Labortory, Brooks A .F .B .,TX

[7]Rankin, Earl F (1965). Cloze Procedure—A Survey of Research, ES . Thurston and L .E. Hafner (eds .), The Philosophical and Sociological Bases of Reading, Fourteenth Yearbook of the National Reading Conference, 133-150.

[8] Daniel Jurafsky and James H Martin Speech(2000). Speech and Language Processing- An introduction to computational linguistics,Pearson Education, 217-253

[9] Mitchell Marcus, et al. (1993). Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics, vol. 19.

[10] R Rosenfeld(1995). The CMU statistical Language Modeling Toolkit, and its use in the 1994 ARPA SCR Evaluation, ARPA Spoken Language Technology workshop

BIOGRAPHICAL INFORMATION

Mr. Jagadeesh Kondru received his Bachelors degree in Computer Science and Engineering from Acharya Nagarjuna University, India in May, 2004 and his Masters in Computer Science from the University of Texas at Arlington in December, 2006. His research interests are in natural language processing and data mining. He is a recipient of the Dean's Graduate Fellowship at the University of Texas at Arlington.