AN AUTOMATED LEARNER-BASED READING ABILITY ESTIMATION STRATEGY USING CONCEPT INDEXING WITH INTEGRATED PART-OF-SPEECH N-GRAM FEATURES

by

ABIGAIL R. RAZON

A thesis submitted to the University of Birmingham
for the degree of DOCTOR OF PHILOSOPHY

School of Computer Science
University of Birmingham
March 2016

Abstract

This study is about the development of a retrainable reading ability estimation system based on concepts from the Text Readability Indexing (TRI) domain. This system aims to promote self-directed language learning and to serve as an educational reinforcement tool for English language learners. Student essays were used to calibrate the system which provided realistic approximations of their actual reading levels.

In this thesis, we compared the performance of two vector semantics-based algorithms, namely, Latent Semantic Indexing (LSI) and Concept Indexing (CI) for content analysis. Since these algorithms rely on the bag-of-words approach and inherently lack grammatical analysis, we augmented them using Part-of-Speech (POS) n-gram features to approximate the syntactic complexity of text documents.

Results show that directly combining the content- and grammar-based feature sets yielded lower classification accuracies than utilising each feature set alone. Using a sparsification strategy, we were able to optimise the combination process and, with the integration of POS bi-grams, we achieved our overall highest mean exact agreement accuracies (MEAA) of 0.924 and 0.952 for LSI and CI, respectively.

We have also conducted error analyses on our results where we examined overestimation and underestimation error types to uncover the probable causes for the systems' misclassifications.

Acknowledgements

First of all I would like to thank the Lord for His guidance and for blessing me with good health and peace of mind during the course of my research .

I would like to express my sincere gratitude to my supervisor, Prof. John A. Barnden, for his patience, motivation, and immense knowledge. His guidance and words of wisdom gave me the strength to carry on with the challenges in my PhD life.

I would like to acknowledge the financial support provided by the Department of Science and Technology of the Philippines through its Engineering Research and Development for Technology (ERDT) Program. I am also very grateful for the support given by the English departments of the University of the Philippines Integrated School (UPIS) and the Philippine Science High School (PSHS), and most especially for the dedication given by Dr. Ma. Lourdes J. Vargas of the UPIS, in the collection of the datasets used in this thesis.

To all my friends and colleagues who have always been there for me even through the worst days of my research, thank you so much.

Last but not the least, I would like to thank my family: my beloved parents, my brothers and sisters, and my husband and son, for supporting me and being patient with me throughout my PhD studies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAA Adjacent Agreement Accuracy.

ATOS Advantage-TASA Open Standard.

CD Concept Decomposition.

DC Dale-Chall Formula.

DET Detection Error Tradeo.

EAA Exact Agreement Accuracy.

EM Expectation Maximisation.

FK Flesch-Kincaid Grade Level.

FOG Fog Index.

FRE Flesch Reading Ease Formula.

GR Guided Reading.

GUM Georgetown University Multilayer.

IELTS International English Language Testing System.

IR Independent Reading.

ML Machine Learning.

MSE Mean Squared Error.

NCLB No Child Left Behind.

NLP Natural Language Processing.

O-type Overestimation error.

RANLP Recent Advances in Natural Language Processing.

RBF Radial Basis Function.

RMM Reading Maturity Metric.

RTF Raw Term Frequency.

SAT Scholastic Achievement Test.

SD Standard deviation.

SI Sparsity index.

SMOG Simple Measure of Gobbledygook.

SS Sparsication Strategy.

SVD Singular Value Decomposition.

SVM Support Vector Machines.

SVN Support Vector Networks.

TASA Touchstone Applied Science Associates.

TF-IDF Term Frequency-Inverse Document Frequency.

TRA Text Readability Analysis.

TTM Time to Maturity.

U-type Underestimation error.

UPIS University of the Philippines Integrated School.

WM Word Maturity.

Chapter 1

Introduction

This chapter provides an overview of the thesis. In Section 1.1, we discuss what *Reading* and *Reading Comprehension* entail. Then, we introduce the concepts of *Text Readability* and *Text Readability Analysis* (TRA) in Section 1.2. Discussion of the importance of TRA follows in Section 1.3. After which, we enumerate this study's main contributions and present the publication resulting from this research in Section 1.4 and Section 1.5, respectively. Finally, Section 1.6 concludes this chapter by providing the coverage of the succeeding chapters of this thesis.

1.1 Reading and Reading Comprehension

In Snow (2002), reading is defined as a process which involves simultaneous extraction and construction of meaning from written language and is composed of three basic elements, the reader, the text, and the activity. It is considered to be a problem solving *activity* in which the *reader* attempts to comprehend the ideas within the *texts* (Snow, 2002). In Biddulph (2002), it is defined as "an interactive process in which readers actively engage with texts, building their own understanding of the author's message". Braunger and Lewis (1997) define it "as an active, cognitive and affective process which leads to the construction of meaning from written texts". Summarising, we can say that reading is a process or an activity wherein a reader

converts texts into ideas or concepts based on his own understanding.

Reading comprehension is affected by two main factors, the reader's ability and the features of the texts. Lorge, the author of the Lorge's Formula which is one of the very first readability formulas, viewed reading comprehension as "the interaction between reading ability and text readability" (Lorge, 1944). On one hand, it is dependent on the reader's ability to understand what he is reading which inherently depends on his cognitive capacities, motivation, and knowledge (Snow, 2002). On the other hand, it is influenced by text features, such as content, grammar and vocabulary, which affect the readability of the texts.

Nowadays, reading ability is formally assessed using standardised tests such as the Scholastic Achievement Test (SAT) which is mostly taken by American middle school children. Another example is the International English Language Testing System (IELTS) given to aspiring immigrants of English-speaking countries. In these tests, reading ability is estimated by requiring the test takers to answer reading comprehension questions which come after each passage.

In this thesis, we aim to provide evidence that Machine Learning strategies can be used effectively to approximate the reading ability levels of the English language learners. Specifically, we would like to address the questions in Chapter 3 which can help in the development of a reading reinforcement tool for teachers and learners. This tool will help guide them in choosing appropriate materials to read and will support both *Independent Reading* (IR) and *Guided Reading* (GR) approaches to literacy education which will be discussed in the next chapter.

1.2 Text Readability and Text Readability Analysis

Text Readability has been defined by several authors in different ways. According to DuBay (2004), "it is what makes some texts easier to read than others." In the same literature, he cites the definitions of the term given by other researchers on text readability analysis. One of these authors is Klare (1963) who defined text readability as "the ease of understanding or comprehension due to the style of writing". McLaughlin (1969), the creator of the Simple Measure of Gobbledygook (SMOG) readability formula, stated that text readability is "the degree to which a given class of people find certain reading matter compelling and comprehensible". But, according to DuBay (2004) the most comprehensive definition of readability is the one given by Dale and Chall (1949) which states that readability is "The sum total (including all the interactions) of all those elements within a given piece of printed material that affects the success of a group of readers...The success is the extent to which they understand it, read it at an optimal speed, and find it interesting". In this thesis, we combine these definitions together and define Text Readability as a measure of the required reading level of the reader for him or her to understand the content, distinguish grammar structures and know the majority of the reading material's vocabulary.

Moreover, we will define TRA as the extraction and utilisation of valuable features (e.g. grammar, content, and vocabulary) from written documents to be able to decide on its level of readability. Past research on the TRA domain, such as Si and Callan (2001) and Heilman et al. (2007), rely greatly on syntactic features as indicators of text readability. Such features include sentence length, syllable count, character count per word, part-of-speech (POS), and word frequency. Although these features are important linguistic components, these have not been sufficient to model text difficulty levels. As a result, recent studies are geared towards using content

learning techniques from the Natural Language Processing (NLP) area using features based on word unigrams, lists of hard and/or easy words, and word frequencies. Such techniques include Latent Semantic Indexing (LSI) and Concept Indexing (CI) which have the ability to extract text content-related features from documents using frequency measures of the words present in the text samples. This thesis presents a comparative study on these two techniques and discusses how the integration of grammar-related features can affect them.

1.3 Importance of Text Readability Analysis

In education, providing suitable reading materials to students is crucial. As stated in Milone (2009), students learn more efficiently if their books are neither too hard nor too easy. On one hand, if a reading material is too hard, the student will not understand it and he may feel intimidated. DuBay (2004) also stated that it is likely for readers to stop when they cannot understand what they are reading. Once they stop reading, the learning process is also hindered. On the other hand, if a reading material is too easy, then the student is bound to feel less intellectually motivated by it. This may result in boredom and lose of interest in language learning.

Readability analysis has several benefits, not just in education, but also in healthcare, industry and government. In healthcare, it can be used in writing medical instructions, which need to be correctly understood by an average patient or person (Al-Khalifa and Al-Ajlan, 2010). It can be used to write textbooks and other reading resources which can be easily understood by students of healthcare-related courses. It is also one of the technologies being utilised to provide more effective clinical guidelines (i.e. specialised clinical documents describing appropriate treatment and care for patients with special conditions) for healthcare professionals. In industry, text

readability assessment of user manuals and instructions is also of top priority since confusion on these documents can result in product or property damage and even death. For example, traffic accidents, which led to deaths among children aged 1 to 14 in 1998, were suspected to be because of the gap between the average reading ability level of 80% of the adult readers in the U.S. (i.e. 7th grade level) and the average readability level of the child-safety seat installation instructions (i.e. 10th grade level) (DuBay, 2004). The industry also uses readability analysis to be more effective in promoting products by making more easily readable materials for their target customers. Government agencies can also benefit from text readability analysis since their official documents or forms are required to meet specific readability levels to make them suitable for every member of the society, including people with low educational levels and people with reading difficulties (Al-Khalifa and Al-Ajlan, 2010). It is also necessary for the government to deliver sensitive information and to do international transactions as clearly and as understandably as possible to avoid misunderstanding and even war.

1.4 Main Contributions of the Study

The main contributions of this study are as follows:

1. This study proposes a new approach to reading ability estimation using concepts in Text Readability Analysis, where 1.) the main features used are not explicitly based on text features (i.e. syntax- and vocabulary-based features), but rather based on content similarity features between instructional reading materials and actual essays written by primary and secondary school students, and 2.) the secondary features used are POS n-grams instead of word n-grams.

2. To the best of our knowledge, this is the very first study which augments CI-based content features with POS n-gram features on English text documents. Although, a similar set of main features has already been applied on Tagalog texts in Razon et al. (2011), the algorithm and feature set combinations we propose in this thesis are different from what was proposed in that study.

3. The experiments conducted in this study provide further evidence that combined grammar- and content-based features can yield better results in the text readability indexing domain, as also indicated in Schwarm and Ostendorf (2005) and Heilman et al. (2007). This study also provides empirical justification that the success in combining grammar- and content-based features involves more sophisticated feature analysis than just directly mixing feature sets together, which was commonly done in previous research. Based on the results, elimination of sparse POS n-gram feature vectors has proven to improve the performance of the combined CI and POS-based systems in general.

4. This study delivers a retrainable learner-focused approach to reading ability estimation using concepts and strategies in the Text Readability Analysis (TRA) domain. Since the associated system is calibrated using the learners' written essays, it has the intrinsic ability to provide learners with reading materials which are more closely fitting to their reading ability. This would also allow a more flexible self-directed learning of the English language. Moreover, should the system become outdated, it can also be retrained easily by feeding new text samples into it.

1.5 Publication Resulting from the Thesis Work

In the course of the thesis, we were able to present a paper in the *Recent Advances in Natural Language Processing* (RANLP) conference which was held in Hissar, Bulgaria in September 2015. The paper is entitled *A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features* (Razon and Barnden, 2015).

1.6 Organisation of the Thesis

This section provides an overview of the rest of the thesis. Chapter 2 provides some general information on reading, discusses prominent readability formulas and related research on TRA, and presents existing comparative studies involving LSI and CI. The issues and inadequacies of previous studies on TRA which we aim to address in this study are pointed out in Chapter 3. Chapter 4 provides the implementation details of the development of the combined CI and POS n-gram-based reading ability estimation system. Chapter 5 presents all the experiments conducted in this study, together with the discussion of the results and the analysis of the errors encountered by the system. Lastly, Chapter 6 concludes this work and enumerates some possible future studies related to it.

Chapter 2

Related Literature

In this chapter, we will provide some general information on literacy education related to reading ability assessment, including the relationship of reading and writing abilities, in Section 2.1. Then, we will focus our discussion on the several approaches applied on readability analysis to date. These approaches can be classified into two major categories which revolve around the use of 1.) Readability Formulas and 2.) Machine Learning (ML) strategies. We will present some of the well-known readability formulas in Section 2.2 and we will also discuss the more recent ML-based strategies in Section 2.3. Then, in Section 2.4, we will present existing comparative studies on LSI and CI. Lastly, we will summarise this chapter in Section 2.5.

2.1 Approaches to Literacy Education Related to Reading

Reading can be learned in different ways and it can be complemented with writing activities. In this section, we will discuss reading as an approach to literacy education.

2.1.1 Independent Reading vs. Guided Reading

In Independent Reading (IR), learners choose what they want to read. As stated in Cullinan (2000), "IR is done for information or for pleasure. No one assigns it; no one requires a report; no one checks on comprehension." Therefore, IR is purely

dependent on the reader's preference.

Unlike IR, Guided Reading (GR) requires a facilitator as discussed in Biddulph (2002). In this approach, a facilitator (e.g. teacher) selects a reading material for the learners to read. He introduces the material to them by discussing relevant experiences before the reading activity. Discussions of the material before, during and after reading are encouraged to boost the learning experience. Thus, facilitators play a very important role in this approach. Their choice of materials is crucial in the learning process.

2.1.2 Relationship of Reading and Writing Abilities

Reading and writing abilities grow together as a learner progresses in school. However, as stated in Graham and Hebert (2010), writing is often disregarded as a tool in improving reading when in fact it has a theoretical potential in doing so. Similarly, we can also argue that reading is a tool in improving writing. This connection between these two activities has been the focus of recent studies in literacy education.

Graham and Hebert (2010) provides three ways in which writing can improve reading. 1.) Being both functional activities, reading and writing can be combined to accomplish specific learning goals. When writing about a concept in a Science course, learners tap into the information they acquired by reading and this event "provides the reader with a means for recording, connecting, analysing, personalising, and manipulating key ideas from the text." Consequently, 2.) we can also infer that reading and writing activities draw upon a common source of knowledge. Thus, improving one will also improve the other. 3.) Reading and writing are both communication skills. Writers gain insights from what they read, and write about them. To be able to produce beautiful write-ups, they should develop better comprehension of texts produced by others.

Graham and Hebert (2010) also presented three recommendations on how to strengthen reading through writing. These recommendations are:

1. *HAVE STUDENTS WRITE ABOUT THE TEXTS THEY READ. Students' comprehension of Science, Social Studies, and Language Arts texts is improved when they write about what they read, specifically when they:*
   - *respond to a text in writing (Writing personal reactions, analysing and interpreting the text)*
   - *write summaries of a text*
   - *write notes about a text*
   - *answer questions about a text in writing, or create and answer written questions about a text*
2. *TEACH STUDENTS THE WRITING SKILLS AND PROCESSES THAT GO INTO CREATING TEXT. Students' reading skills and comprehension are improved by learning the skills and processes that go into creating text, specifically when teachers:*
   - *teach the process of writing, text structures for writing, paragraph or sentence construction skills (Improves Reading Comprehension)*
   - *teach spelling and sentence construction skills (Improves Reading Fluency)*
   - *teach spelling skills (Improves Word Reading Skills)*
3. *INCREASE HOW MUCH STUDENTS WRITE. Students' reading comprehension is improved by having them increase how often they produce their own texts.*

- Adapted from Graham and Hebert (2010)

2.1.3 Quantitative Research on Reading and Writing Connection

Quantitative analysis of the effects of different writing practices on reading was also conducted in Graham and Hebert (2010). Figure 2.1 shows the confidence intervals in which the "true" effect of each of these practices lies. An effect size value of zero means that the practice or activity does not have any effect on reading and values greater than zero mean that that particular activity can be used to enhance reading. As shown, none of the activities have an effect size value less than or equal to zero. Thus, it was concluded in Graham and Hebert (2010) that writing activities generally have a positive effect in reading.

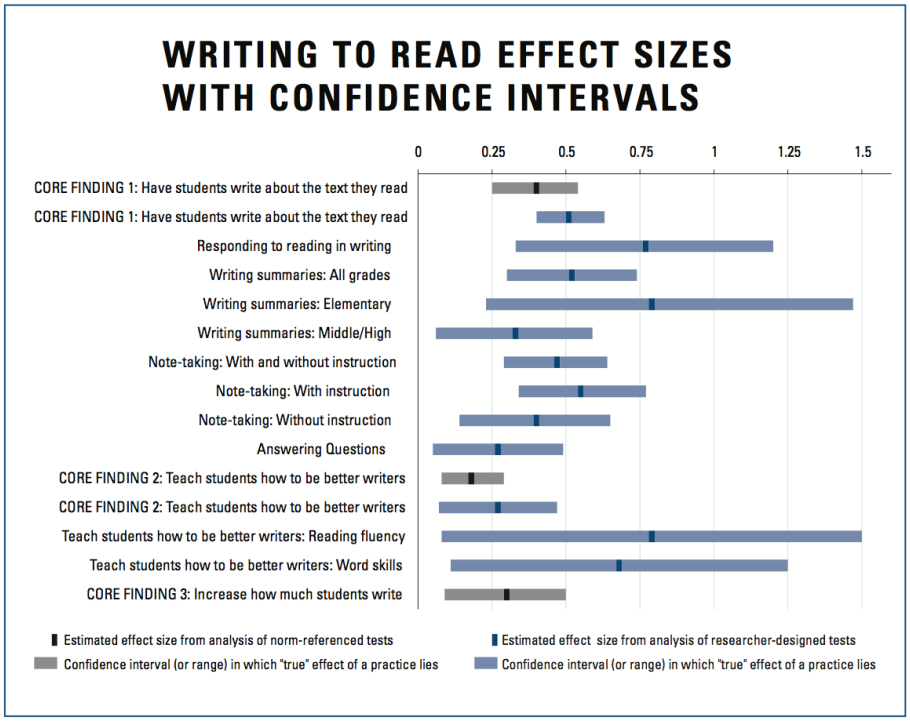


Figure 2.1: Result of the Graham and Hebert (2010) Study on the Effects of Different Writing Practices on Reading

Metametrics, the developer of the Lexile measure which is being used in 50 states in the US, also conducted a study on the connection between reading and writing (Smith III, 2009). With a vast amount of data they acquire each year, reaching as high as 28 million Lexile measures, they developed the Lexile Framework in which both reading and writing abilities can be estimated on the same developmental scale. Based on the results of one of their studies using 589 students across eight grade levels, reading ability is consistently lower than writing ability as shown in Figure 2.2. By looking at the line graphs for the reading and writing abilities, we can also see that there is a positive correlation between these abilities across different student grade levels. This is denoted by the relatively equally-spaced gap between the two lines.

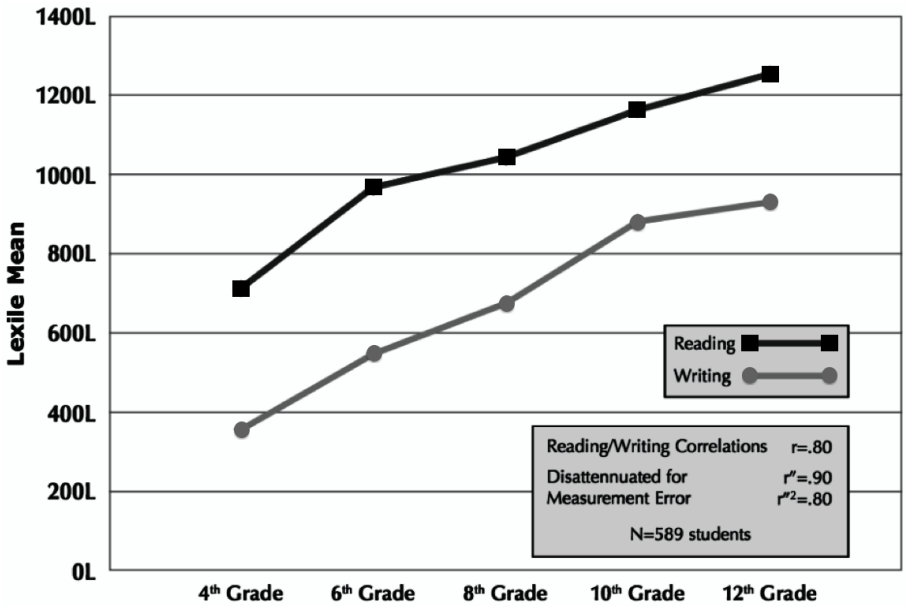


Figure 2.2: Cross-Sectional Reading and Writing Lexile Means (Smith III, 2009)

2.2 Prominent Readability Formulas

The study of readability analysis has been going on for over a century now. As stated in DuBay (2004), after realising that students learn reading in steps and that they learn more efficiently when the materials they use are suitable for their reading ability, educators promoted the use of prepared levelled reading materials as a standard instructional procedure. With the rise of this new system, several readability formulas were proposed. DuBay estimated that there were at least 200 readability formulas by the 1980s. This number kept increasing in the 1990s not just for the English language but also for other languages.

Below are some of the popular readability formulas for the English language:

1. 1939 Lorge's Formula, revised in 1948 (Larsson, 2006) — This formula is based on American standards. It has three factors, namely, 1.) average number of word tokens per sentence, 2.) number of difficult words which are not in the Dale-Chall list of 769 easy words (i.e. subset of the Dale-Chall 3000 word list[1]) (Dale and Chall, 1948) divided by total number of words, 3.) number of prepositional phrases divided by the total number of words. In this formula, the readability index (RI) is rounded off to the nearest half value as shown in the sample computation in Figure 2.3.

[1]This is a list of words which 80% of 4th grade American students can understand

FORMULA FOR ESTIMATING GRADE PLACEMENT OF READING MATERIAL

WORK SHEET

Title of article: Gettysburg Address — Edition: first revision
Name of author: Abraham Lincoln
Publisher: — Date of Publication: Nov. 19, 1863
Location of sample in text: Complete — R. I. = 6.5

BASIC DATA

1. The number of words in the sample 269
2. The number of sentences in the sample 10
3. The number of prepositional phrases in the sample 26
4. The number of hard words in the sample 43

COMPUTATION[28]

Item 6, average sentence length: Divide 1 by 2 = 26.90 x .07 = 1.8830
Item 8, ratio of propositional phrases: Divide 3 by 1 = .0967 x 13.01 = 1.2581
Item 9, ratio of hard words: Divide 4 by 1 = .1599 x 10.73 = 1.7151
Constant = 1.6126
Add 6, 8, 9, and C
Readability Index: 6.4694

NOTES

lives, n. called easy
Last sentence, although long, is broken up by adequate punctuation

Name of Analyst: I. D. L — Date of analysis: Nov 23, 1943
Name of Computer: I. D. L. — Date of computing: Nov. 23, 1943
Name of checker: J. C.

Figure 2.3: Sample Actual Computation of the 1939 Lorge's Formula (DuBay, 2006)

2. 1948 Flesch Reading Ease (FRE) Formula (Larsson, 2006) — This formula maps readability scores between 0 and 100 to American reading grade levels (i.e. from 5th level to college graduate level) where a lower score indicates a more difficult text (refer to Table 2.1). It is dependent on two variables: 1.) average number of word tokens per sentence and 2.) average number of syllables per word.

Table 2.1: FRE Scores Interpretation (Badgett, 2010; DuBay, 2006)

FRE Score	Interpretation	Estimated Reading Grade	% US Adults
00-30	very difficult	College grad level	4.5
30-50	difficult	13th-16th level	33.0
50-60	fairly difficult	10th-12th level	54.0
60-70	standard	8th-9th level	83.0
70-80	fairly easy	7th level	88.0
80-90	easy	6th level	91.0
90-100	very easy	5th level	93.0

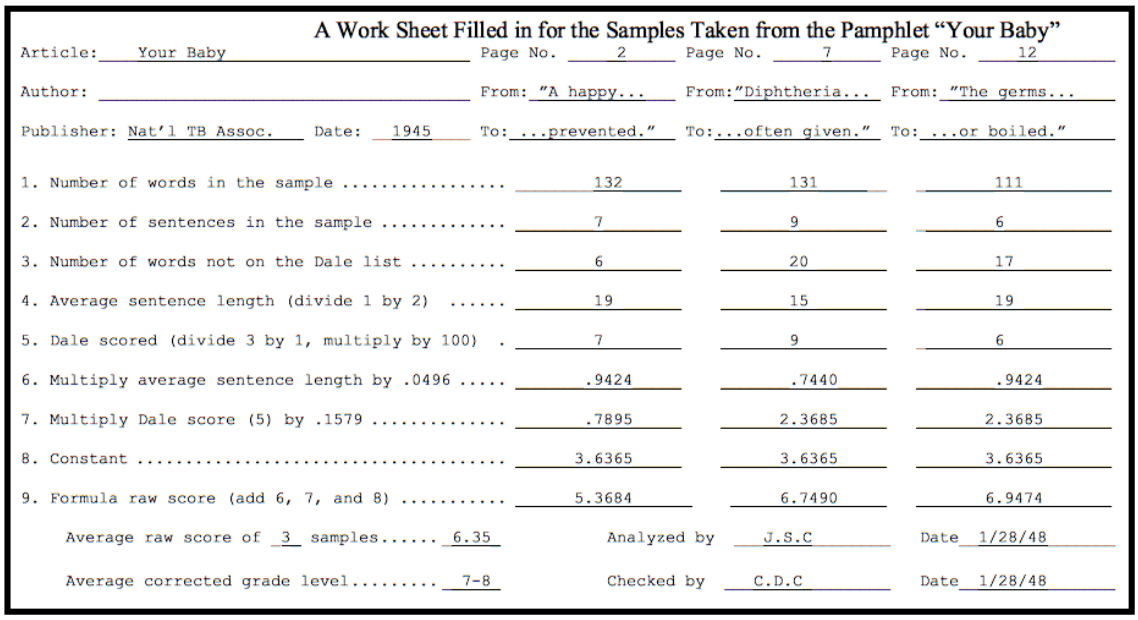
3. 1948 Dale-Chall (DC) Formula, revised in 1995 (Dale and Chall, 1995) — These formulas are based on American grade levels. The 1948 version of the formula only covered grade levels 4 to 16, whereas the 1995 version covered grade levels 1 to 16. Note that these formulas can only deliver a grade level range and not a specific grade level classification of a text document. They are dependent on two variables: 1.) the average number of word tokens per sentence, and 2.) percentage of the number of words not occurring in the Dale-Chall list of 3000 easy words. Tables 2.2 and 2.3 provide the mapping between the score ranges and grade levels for the 1948 and 1995 versions of the formula, respectively. To illustrate how this formula is used, a sample computation is shown in Figure 2.4.

Table 2.2: Mapping between the Grade Levels and the Dale-Chall's 1948 Score Ranges (Dale and Chall, 1995)

Score Range	Grade Levels
4.9 and below	4 and below
5.0-5.9	5-6
6.0-6.9	7-8
7.0-7.9	9-10
8.0-8.9	11-12
9.0-9.9	13-15 (college)
10 and above	16 and above (college graduate)

Table 2.3: Mapping between the Grade Levels and the Dale-Chall's 1995 Score Ranges (Dale and Chall, 1995)

Score Range	Grade Levels
58 and above	1
57-54	2
53-50	3
49-45	4
44-40	5-6
39-34	7-8
33-28	9-10
27-22	11-12
21-16	13-15
15 and below	16



A Work Sheet Filled in for the Samples Taken from the Pamphlet "Your Baby"

Article: Your Baby	Page No. 2	Page No. 7	Page No. 12
Author:	From: "A happy...	From: "Diphtheria...	From: "The germs...
Publisher: Nat'l TB Assoc. Date: 1945	To: ...prevented."	To: ...often given."	To: ...or boiled."
1. Number of words in the sample	132	131	111
2. Number of sentences in the sample	7	9	6
3. Number of words not on the Dale list	6	20	17
4. Average sentence length (divide 1 by 2)	19	15	19
5. Dale scored (divide 3 by 1, multiply by 100) .	7	9	6
6. Multiply average sentence length by .0496	.9424	.7440	.9424
7. Multiply Dale score (5) by .1579	.7895	2.3685	2.3685
8. Constant	3.6365	3.6365	3.6365
9. Formula raw score (add 6, 7, and 8)	5.3684	6.7490	6.9474
Average raw score of 3 samples..... 6.35	Analyzed by J.S.C		Date 1/28/48
Average corrected grade level..... 7-8	Checked by C.D.C		Date 1/28/48

Figure 2.4: Sample Actual Computation of the 1948 Dale-Chall Formula (DuBay, 2006)

4. 1952 Gunning Fog Index (FOG) (Badgett, 2010) — This formula has two variables: 1.) average number of word tokens per sentence and 2.) percentage of difficult words (i.e. words with more than two syllables) in passages for which students from grade levels 6, 8, 10, and 12 correctly answered 90% of the comprehension questions (DuBay, 2004). Outputs are rounded off to the nearest grade level.

5. 1969 SMOG Formula (McLaughlin, 1969) — This formula only has one variable, the polysyllable count or the number of words with 3 or more syllables. Outputs are rounded off to the nearest grade level.

6. 1975 Flesch-Kincaid (FK) Grade Level (Larsson, 2006) — This is a modification of the FRE formula mentioned above. It is also dependent on two variables: 1.) average number of word tokens per sentence and 2.) average number of syllables per word. Outputs are rounded off to the nearest grade level.

7. 1997 Lexile Measure (The Lexile Website, 2013; Stenner et al., 2007; Burdick, 2010) — The standard scale for the Lexile measure ranges from 0L to above 2000L, where L stands for *Lexile*. It is dependent on: 1.) log of the average number of word tokens per sentence and 2.) average of the log word token frequencies. Outputs are rounded off to the nearest Lexile score.

8. 2000 Advantage-TASA Open Standard (ATOS) Formula, where TASA stands for Touchstone Applied Science Associates (Milone, 2009) — This formula is based on American grade levels and is dependent on three variables: 1.) average number of word tokens per sentence, 2.) average number of characters per word, and 3.) average grade level for words found in their derived graded vocabulary list excluding the top 100 most frequent words. Outputs are rounded off to the

nearest grade level.

Although these readability formulas have been widely used to measure text difficulty levels, they are often criticised because of their strong dependency on surface linguistic features. Other features, such as semantics and grammar, are often not considered. Moreover, as presented in DuBay (2004), discrepancies in output grade levels among these readability formulas have also been an issue. To illustrate the discrepancies, we picked 3 reference documents per grade level from the 2014 Grades 7-9 dataset and used the online resource, Readability Formulas (2015)[2], to calculate each one's readability using 5 of the formulas discussed above. The titles of these documents are listed below and their full texts can be found in Appendix B.

1. Grade 7 documents
   (a) E1: *Story of Maykapal*
   (b) E2: *Reproductive health bill: Facts, fallacies*
   (c) E3: *Belief in Supreme God*
2. Grade 8 documents
   (a) E4: *THE TIGER*
   (b) E5: *"MY GOD! WHAT HAVE WE DONE?"*
   (c) E6: *Bound Feet*
3. Grade 9 documents
   (a) E7: *ANGLO-SAXON INVASION OF BRITAIN*
   (b) E8: *THE COMING OF GRENDEL*
   (c) E9: *The Grapes of Wrath*

Table 2.4 shows the output American grade levels of FRE, FOG, FK, SMOG and DC formulas on the 9 reference documents using the Readability Formulas (2015) online resource. As evident in the table, the outputs of the readability formulas could

[2] `http://www.readabilityformulas.com`

greatly differ for 1 document. For example, E4 got grade level ratings of 8 and 13 for SMOG and FOG formulas, respectively. A similar case is true for E6.

Table 2.4: American Grade Level Outputs of Readability Formulas

FORMULA NAME	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**
FRE	7	13-16	13-16	8-9	8-9	10-12	8-9	8-9	6
FOG	9	16	11	13	9	13	10	14	8
FK	6	13	10	11	8	11	9	11	7
SMOG	7	12	10	8	8	8	8	8	4
DC	7-8	11-12	11-12	9-10	9-10	9-10	11-12	9-10	7-8

2.3 Machine Learning Strategies on Readability Analysis

In this section, we will discuss some recent (year 2001 onwards) prominent research on readability analysis using *Machine Learning* (ML) strategies. We will cite different systems based on different algorithms for ML-based text readability analysis such as:

- Expectation Maximisation
- Support Vector Machines
- Multinomial Naive Bayes
- Decision Trees
- Latent Semantic Indexing

2.3.1 2001 Expectation Maximisation-based System by Si and Callan

The study in Si and Callan (2001) combined content-based and surface linguistic features to create a text readability level classifier. The authors used the *Expectation Maximisation* (EM) algorithm to automatically calculate the weight values for their proposed models, namely, the *unigram language model* (i.e. using words in text) and the *sentence length distribution model*. On one hand, the unigram language model is based on the assumption that the probability that a word would appear in text is independent of its context and is not influenced by other words in it. On the other hand, the sentence distribution model assumes that a normal or Gaussian distribution with a specific mean and variance can be used to model sentence length distribution of texts.

To combine the two models, the authors used the formula, $P_c(g|d_i) = \lambda * P_a(g|d_i) + (1 - \lambda) * P_b(g|d_i)$, where g and d_i represent a readability grade level and a specific document, respectively. $P_c(g|d_i)$ represents the probability of a readability grade level given a specific document. It is equal to the linear combination of the unigram language model, represented as $P_a(g|d_i)$ and the sentence length distribution language model, represented as $P_b(g|d_i)$, with λ as the weight value between these models.

The study conducted in Si and Callan (2001) revolved around three major hypotheses: 1) Readability measures should be sensitive to content as well as to surface linguistic features. 2) Statistical language models could capture the content information related to reading difficulty. 3) The normal distribution with a specific mean and variance can be used to model the sentence length distribution of each readability grade level.

Results of their experiments revealed that: 1.) sentence length is a useful feature for readability analysis on their dataset since its mean value increases as the readability level of texts increases and 2.) syllable count is not a useful feature since it does not exhibit the same behaviour. The authors also reported that the system based on the unigram language model was able to achieve a higher accuracy value of 70.5% than the system based on the sentence length distribution model which only achieved 42.6% accuracy. Moreover, by combining these two models, they were able to achieve their highest accuracy of 75.4%.

2.3.2 2005 Support Vector Machines-based System by Schwarm and Ostendorf

In Schwarm and Ostendorf (2005), binary *Support Vector Machines* (SVM) were utilised to approximate the syntactic and semantic complexities of texts. Several

text features including sentence length, syllable count, word instances (i.e. tokens), unique words (i.e. types), part-of-speech tags, parse tree height, average number of noun phrases, average number of verb phrases, and word uni-, bi-, and tri-gram features were used in training the classifiers to distinguish articles for grade levels 2 to 5. The corpus used in this study was created from several sources including the pre-graded 2004 Weekly Reader, Encyclopædia Britannica, CNN News Stories (full and abridged versions), and Brittanica Elementary Encyclopædia, covering topics in science, history, and current events.

Detection Error Tradeoff (DET) curves, *Precision*[3] and *Recall*[4] metrics were used for system evaluation. DET curves show the tradeoff between *misses* (i.e. positive examples misclassified as negative) and *false alarms* (i.e. negative examples misclassified as positive) on different classifier threshold values. In the experiments, Schwarm and Ostendorf observed the contribution of individual features to the overall performance of the SVM classifiers and found that: 1.) no feature stood out as the most important one, and 2.) system performance was degraded when any particular feature was removed. They also realised that trigram models were noticeably more accurate than bigrams and unigrams.

As shown in Table 2.5, their system can sometimes achieve precision of 75% and recall of 87%, with adjacent accuracy classification error (i.e. percentage of articles which are misclassified by more than one grade level) of 3.3%. Comparison of their proposed approach versus Lexile and Flesch-Kincaid, two of the popular readability measures, is presented in Table 2.6, wherein we can see that their system achieved the lowest adjacent accuracy classification error values for all grade levels.

[3] $Precision = \frac{true\ positive}{true\ positive + false\ positive}$; the fraction of retrieved instances that are relevant
[4] $Recall = \frac{true\ positive}{true\ positive + false\ negative}$; the fraction of relevant instances that are retrieved

Table 2.5: The *Precision* and *Recall* of Schwarm and Ostendorf's SVM-based Classifiers (Schwarm and Ostendorf, 2005).

Grade	Precision	Recall
2	38%	61%
3	38%	87%
4	70%	60%
5	75%	79%

Table 2.6: Schwarm and Ostendorf's Approach vs. the Lexile and the Flesch-Kincaid Formulas (Schwarm and Ostendorf, 2005).

Grade	Adjacent Classification Error		
	Flesch-Kincaid	Lexile	Schwarm-Ostendorf System
2	78%	33%	5.5%
3	67%	27%	3.3%
4	74%	26%	13.0%
5	59%	24%	21.0%

2.3.3 2006 Support Vector Machines-, Decision Trees-, and Naive Bayes-based Systems by Wang

The study conducted in Wang (2006) focused on indexing consumer health information web sites. The aim was to classify reading materials into two categories, easy and hard. Easy to read materials should be readable to people who have difficulty reading or understanding information, typically in the fourth to sixth grade reading level. Hard to read materials, including patient education materials should be readable to audiences with sixth to eighth grade reading level.

Three approaches were investigated in this study, namely, SVM, Decision Trees and Naive Bayes. However, the objective of the study was not to compare the machine learning methods themselves, but to compare the performance of feature sets using these machine learning methods. The features considered in the research were

categorised into three levels: word, document and domain-dependent levels. The *word level* category used the Dale-Chall *easy* word list (Dale and Chall, 1948) and words with syllable count of three or more as the difficult word list. For the *document level* category, Wang used number of words per sentence, average number of characters per word, and average number of syllables per word as features. She then used unigram features, which were obtained from getting words that occurred more than three times in each document and also occurred three or more times in the training data set, for the *domain-dependent level* category.

Results of Wang's experiments are shown in Table 2.7. Although SVM did not outperform the other two approaches on all feature sets, it is important to note that the accuracy values of each of the approaches (i.e. Decision Tree, Naive Bayes and SVM) exhibit the same increasing pattern as you go down the columns of this table. This consequently implied that accuracy is not solely dependent on the approach used, but also on the features considered.

Table 2.7: Classification accuracy of Wang experiments on the three feature sets (Wang, 2006).

Feature Set	Accuracy (%)		
	Decision Tree	Naive Bayes	SVM
(1) Word Level	66.81	66.34	62.72
(2) Document Level	67.18	66.68	64.67
(1) and (2)	73.41	75.55	76.82
(3) Domain-Dependent	78.68	75.26	80.71
(1), (2) and (3)	79.72	76.18	84.06

2.3.4 2004 Multinomial Naive Bayes-based System by Collins-Thompson and Callan

In Collins-Thompson and Callan (2004), the authors implemented a text readability indexing system based on the Multinomial Naive Bayes algorithm. They used uni-grams and their corresponding uni-gram probabilities to estimate the most probable grade level of a given passage among the 12 American grade levels. They introduced the concept of the *Smoothed Uni-gram* language model, where *smoothing* referred to adjusting probability estimates of *types*, which are unique tokens in the dataset, by shifting part of the model's probability mass from observed types to unseen and rare types. This concept was based on the hypothesis that: *Adjacent grade level models are in fact highly related, so that even if a type is unobserved in one grade level's training data, its probability of belonging in that grade level can be derived from the interpolation of nearby grade level models' probability estimates.* Moreover, Collins-Thompson and Callan also stated that: *There are enough distinctive changes in word usage patterns between grade levels which enable accurate predictions using simple language models, even when the subject domain of the documents is unrestricted*, as part of their working hypotheses in the study.

Their dataset was created from 550 English documents composed of fiction, non-fiction, history, science and other genres in which they observed that: 1.) more difficult words were introduced at later grade levels, and 2.) concrete words like *red* exhibit a steady decline in usage as grade level increases, while the probability of more abstract words like *determine* increases along with it, as shown in Figure 2.5. Noticing also that *stopwords*[5] are prevalent in lower grade levels, Collins-Thompson and Callan did not remove these words from their system. However, they removed low-frequency

[5]common words in a language which are filtered out in text processing

types which occurred less than 3 times in the entire dataset and those types which occurred in less than 3 grade level models (no matter how high their frequency), claiming these to be considered more as site-specific noise than as genuine vocabulary items.

The proposed system in this study was able to achieve consistent correlations[6] of 0.63-0.79 with pre-tagged data across different grade levels. Experiments also revealed that with minimal retraining, the system can be utilised for other languages, such as French.

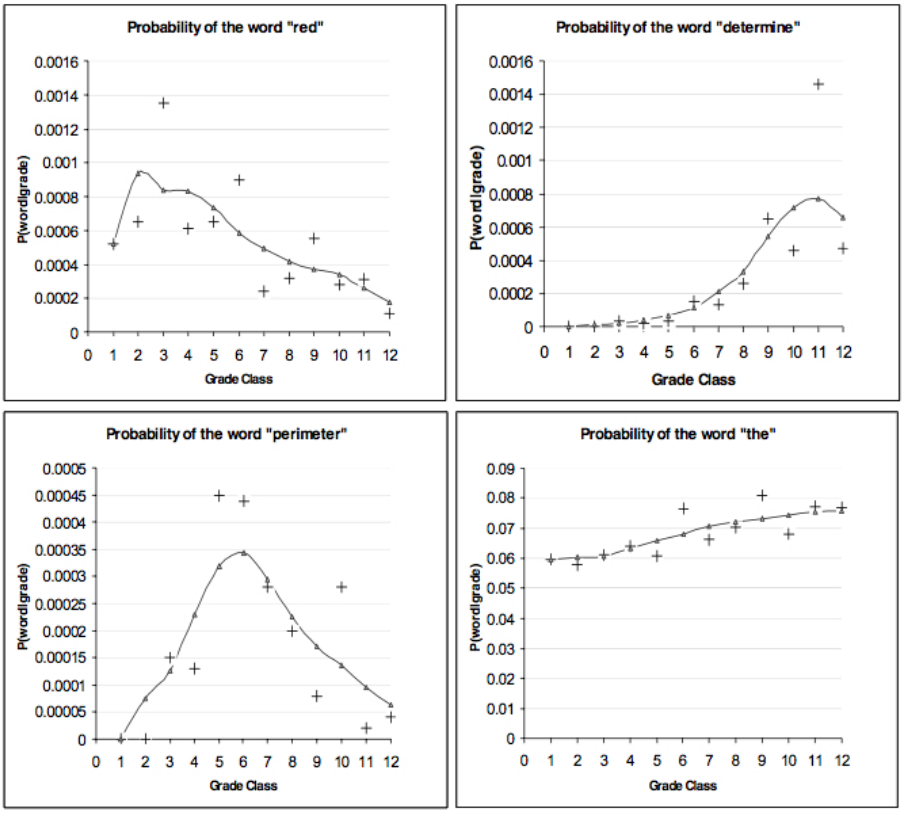


Figure 2.5: Examples of four different word usage trends across grades 1-12, as sampled from the authors' 400K-token corpus of Web documents (Collins-Thompson and Callan, 2004)

[6] correlation of x_i and $y_i = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$

2.3.5 2007 Multinomial Naive Bayes and k-Nearest Neighbour-based Systems by Heilman et al.

In Heilman et al. (2007), the authors had concluded from their interactions with instructors of second language learners of English that combining grammatical and lexical features as predictors of text readability could outperform those measures based solely on one of the two. They combined a vocabulary-based approach using *Multinomial Naive Bayes* classifier on unigrams, and a grammar-based approach using *k-Nearest Neighbour* algorithm on sentence parse trees, sentence length, verb forms, and POS tags features to evaluate text readability. Results of their study showed that the vocabulary-based approach alone is better than the grammar-based approach. However, the combined approach yielded the best performance, reducing the mean squared error value by as much as 22%.

2.3.6 2011-12 Pearson's Reading Maturity Metric

Kireyev, Way and Landauer introduced the Pearson's Reading Maturity Metric (RMM) (Kireyev and Landauer, 2011; Landauer, 2011; Landauer and Way, 2012). The core of the RMM algorithm is the *Word Maturity* (WM) concept which is an LSI-based computational model involving the development of individual word and paragraph meanings as learners become more exposed to the English language. The authors believe that words have different meanings for readers of different ages and reading experience. With the WM concept, their aim is to measure how knowledge of these meanings evolves toward that of literate adults. WM is obtained by cumulatively adding specific educational or naturally ordered samples of text paragraphs in quantities typical of student reading capacity. The order of these cumulative sets of

paragraphs has been selected from materials whose overall difficulty has been previously estimated using the LEXILE rating scale (The Lexile Website, 2013) or any other standard measure.

As discussed in Landauer and Way (2012): The RMM scoring process has two components: the WM-dependent and the linguistics-based components. On one hand, it has the WM-dependent component which is incorporated into the RMM through the *Time to Maturity* (TTM) factor. This TTM factor is derived by getting the number of paragraphs that have to be read in order to achieve the WM threshold value. For example (refer to Figure 2.6 on page 30), the word *turkey* reaches the WM threshold value of 0.65 when about 40,000 paragraphs have been encountered, while the word *productivity* does not reach that threshold value until about 68,000 paragraphs have been read. This number of paragraphs which need to be read are then rescaled to unit length such that words that mature early (e.g. *dog*) are mapped to values closer to 0 and words that mature late (e.g. *productivity*) are mapped closer to 1. Lastly, to get the complexity rating of a text document based on the WM-dependent component of the RMM, these normalised TTM values are averaged except for those n words with the highest TTM values which are given additional weight to minimise the skewing effect caused by the few rarely occurring words in the text that do not follow the general distribution. On the other hand, average sentence length, average word length, n-gram probabilities, and LSI-based features are considered for the linguistic component of the RMM. The relative importance of these features was calibrated on levelled reading passages from *Pearson's Summary Street* product and a collection of publicly available state readings used as items in No Child Left Behind (NCLB) reading comprehension tests in 27 states and 2 national assessments[7].

[7]See Landauer and Way (2012) for details.

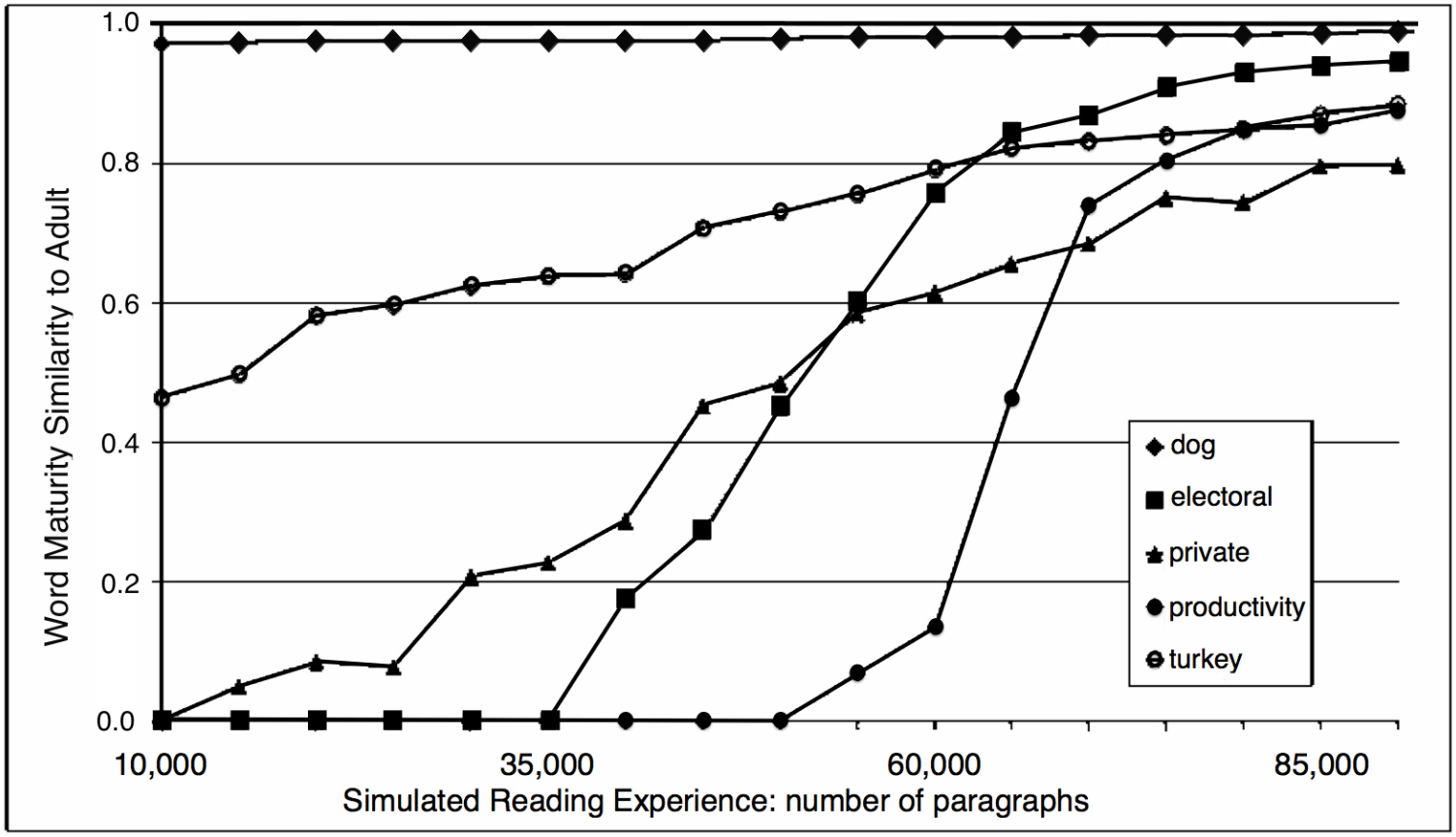


Figure 2.6: Examples of Word Maturity (WM) Trajectories for Five Words (Landauer, 2011).

Results of experiments on the RMM in Landauer (2011) show that it can yield high correlation values of up to 0.88 on pre-tagged corpora. Moreover, these also proved that the RMM can outperform the Flesch-Kincaid and Coleman-Liau measures in assigning readability levels to text documents.

2.4 Existing LSI vs. CI Studies

LSI has been a well-known information retrieval algorithm, patented in 1988 (Deerwester et al., 1989). CI, however, was proposed more recently by Karypis and Han (2000) as a faster alternative for LSI. Both algorithms are based on vector semantics using dimensionality reduction.

In this section, we present existing research comparing the performances of LSI and CI on text content and readability analyses. Then, in Chapter 4, we will provide a detailed discussion on the implementation of these algorithms.

2.4.1 English Essay Content Analysis

The study presented in Razon (2010) focused on comparing LSI and CI as applied on English essay scoring. Through several experiments, the study was able to prove that CI can outperform LSI in grading essays using content features alone. Table 2.8 shows the result of one of the experiments the authors conducted, where *accuracy* was calculated based on the exact agreement between the predicted and actual essay scores (i.e. predicted score by the system = actual essay score). As indicated on this table, CI outperformed LSI on all datasets reaching as high as 84.21% accuracy. It is also important to note that, as shown in the Grade8 dataset results, the difference between the accuracies of the two algorithms can reach as high as 18.75 in favour of CI.

Table 2.8: LSI vs. CI Accuracies (%)

Dataset	LSI Accuracy	CI Accuracy
Grade7	78.947	84.210
Grade8	62.500	81.250
Grade9 Set1	50.000	58.824
Grade9 Set2	64.102	69.231

2.4.2 Filipino Essay Content Analysis

The study in Ong (2011) was an attempt to implement a CI-based Filipino essay grader. Filipino language experts were consulted to validate the outputs. Experiments comparing CI and LSI showed that CI may perform better than LSI for some experts. The experimental results have demonstrated that upon measuring the agreement between the CI-based essay grader system and human raters (i.e. teachers), accuracies between 0.755 and 0.799 are obtained 95% of the time, i.e. within the 95% confidence interval. These are even slightly better than the agreement accuracies among human raters, themselves, which were calculated to be only between 0.706 and 0.709. This implies that the proposed essay grader system is as if behaving as another human rater.

As also stated in Ong (2011), CI, with a small number of vectors representing each pre-defined class or group in the dataset, can run faster than LSI. The time complexity for CI is $O(iekn)$ while LSI is $O(en^2)$, where i is the number of iterations until convergence is achieved, k is the number of vectors representing a set of documents, e is the number of word tokens, and n is the number of essays (Ong, 2011).

2.4.3 Tagalog Text Readability Indexing

A comparative study between LSI- and CI-based algorithms, as applied on readability analysis for Filipino text documents, was conducted in Razon et al. (2011). In the

experiments, cosine similarity of each training document, d, against all the model text documents is calculated. These similarity values for d form a "similarity-to-model" vector representation for document d. To create the training set's matrix representation, the set of these similarity vectors is created over all training documents. The test set's matrix representation was also constructed using the same process. Then, each test document vector representation is correlated against all the vectors in the training set. Grade levels were then assigned to each test document based on the grade level of the corresponding training document with the highest correlation to it. The authors' investigation also focused on the effects of the weighting schemes applied on the cosine similarity matrices by conducting experiments using Raw Term Frequency (RTF) and Term Frequency-Inverse Document Frequency (TF-IDF).

Table 2.9: Exact Agreement Accuracy (%) using Raw Term Frequency (RTF) and Term Frequency-Inverse Document Frequency (TF-IDF) Weighting Schemes

	RTF		**TF-IDF**	
Grade Level	LSI	CI	LSI	CI
2	61.67	80.00	76.67	66.67
3	40.00	52.00	62.00	52.00
4	16.67	36.67	23.33	33.33
6	65.00	47.50	32.50	20.00

As shown in Table 2.9, CI using the RTF weighting scheme outperformed LSI on all the datasets except Grade 6. However, for the TF-IDF weighting scheme, LSI outperformed CI on all the datasets except Grade 4.

With these results, it is inconclusive whether which algorithm is better. However, we can say that CI consistently performed better with the RTF weighting scheme than with the TF-IDF. As for LSI, the inverse is true since it was able to yield higher accuracy values in 3 out of 4 grade levels using TF-IDF.

2.5 Chapter Summary

In this chapter, we presented pertinent literature in reading and TRA. First, we discussed about the importance of reading in literacy education. With the combination of GR and IR approaches, learning the English language can be more effective and efficient. We also established that a learner's reading ability is closely related to his writing ability. Then, we enumerated some of the popular readability formulas developed to objectively match text difficulty levels to learners' reading ability. These formulas mostly rely on surface syntactic features, such as average number of words per sentence and average number of syllables per word. However, these features are subject to change as evident in the study conducted by L.A. Sherman which shows that average words per sentence decreased from 50 words per sentence in the Pre-Elizabethan times to 23 words per sentence during Sherman's time in the 1880s (DuBay, 2004). With this, we can say that readability formulas do not have the intrinsic ability to adapt to the language's evolution. Lastly, we presented some ML strategies applied to the TRA domain. Systems based on these strategies can be easily retrained which addresses the adaptability issue of using readability formulas. Moreover, these ML approaches account for, not just syntactic features, but also semantic text features which make them a more holistic approach to TRA. With the knowledge we gathered from the literature review provided in this chapter, we will formally state the research hypotheses and questions for this thesis in the next chapter.

Chapter 3

Problem Statement

As discussed in the previous chapter, instead of using readability formulas, ML approaches can be utilised for TRA. These approaches have two major advantages, namely, 1) they can yield retrainable systems and 2) they can account for both syntactic and semantic features of texts. With these, we formulated our first and second research questions.

Research Question 1
How can we create an easily retrainable reading ability estimation system using ML strategies?

On top of the second advantage of ML approaches stated earlier, past studies have proven that combining language models using different feature sets can yield better performing systems (Si and Callan, 2001; Heilman et al., 2007; Landauer and Way, 2012). However, there are still several feature set combinations which have not been explored yet. Thus, we would like to contribute to this body of knowledge by integrating feature sets for content- and grammar-based analyses which have not been investigated so far for TRA. Thus, our first hypothesis is as follows, together with our second research question.

Research Hypothesis 1

The combination of content- and grammar-based text features yields better performing systems.

Research Question 2

Which feature set or feature set combinations are most relevant and effective in modelling each school grade level in the datasets?

In most studies, feature sets are just directly combined without prior investigation on how to do the combination process efficiently. Authors of existing systems in this domain of study often do not consider the effect brought about by the new feature relationships established by joining several feature sets together. As stated and proven in Boulis and Ostendorf (2005), it could be the case that the feature is relevant by itself but irrelevant or redundant when considered jointly with other features. We speculate that if we optimise this feature set combination process, we can yield better performing systems. With these, we formulate our second hypothesis alongside our third research question.

Research Hypothesis 2

Optimisation of the feature set combination process yields better performing systems.

Research Question 3
How can we efficiently combine and/or augment the content-based features from CI or LSI with the grammar-based features represented by the POS n-grams?

Text data used to train existing ML-based systems mostly come from pre-classified English reading materials just like in Schwarm and Ostendorf (2005) and Collins-Thompson and Callan (2004). Although the classification of these materials is considered to be verified by experts, in real classroom environment, reading abilities of students vary within each grade level. Thus, there is no guarantee that all students in one grade level have the reading ability suitable for these materials. In effect, students with advanced reading ability may experience boredom and loss of interest in learning, while students with low reading ability may not be able to understand the reading materials and eventually feel intimidated. Moreover, classification categories in existing systems do not always directly correspond to individual school grade levels, as in Si and Callan (2001) which has ranges of grade levels (i.e. Grades K-2, 3-5, 6-8) as final output of the system. Hence, these systems do not have the required calibration to appropriately recommend reading materials to students belonging to a specific school grade level. On top of that, these systems do not also have the ability to account for the different reading abilities within each grade level. With these, we state our fourth research question as follows:

Research Question 4

How can we create a learner-focused reading ability estimation system to be able to recommend reading materials to students in each grade level and to promote self-directed learning?

This research also provides a comparative review of the CI algorithm against the well-known LSI. As shown in Section 2.4, CI has the potential to outperform LSI on text classification problems (Razon, 2010; Ong, 2011; Razon et al., 2011). In this study, we aim to further validate the effectiveness of CI and to provide evidence that CI can be used as an alternative to LSI. Thus, giving us our last research question:

Research Question 5

What performance metrics can we use to validate the effectiveness of the systems?

Chapter 4

Methodology

In this study, multi-class Support Vector Machine (SVM) models (Hsu and Lin, 2002; Chang and Lin, 2011), otherwise known as Support Vector Networks (SVN) (Cortes and Vapnik, 1995), are created using content-based features from LSI and CI, and grammar-related features represented by POS n-grams. These models are then used to classify student reading ability profiles per grade level.

In this chapter, we will present the methodological details of the study. First, the working assumptions are stated in Section 4.1. Then, Section 4.2 provides details of the datasets used in the development of the systems while Section 4.3 discusses the sampling procedure done on these datasets to be able to come up with unbiased training, test and reference sets. Section 4.4 presents the data pre-processing steps considered for the creation of the matrix representations of the text documents. After which, we discuss the content- and grammar-based algorithms followed in the actual implementation of the systems in Section 4.5 and 4.6, respectively. In Section 4.7, we discuss the details of the SVM classifier used in the study. The metrics we used to measure the performance of the systems are presented in Section 4.8. Finally, we summarise this chapter in Section 4.9.

4.1 Assumptions

We have two working assumptions in this study. The first one is that ***written essays by students can be used to approximate their lowest possible reading level***. This assumes that whatever the students can write, they can also read. As discussed in Section 2.1.2, it was empirically proven in Smith III (2009) that students' reading ability is consistently higher than their writing ability. The gap between the mean Lexile score for the two abilities per grade level is approximately 300 Lexile or 300L on the standardised Lexile scales for writing and reading metrics. This empirically-derived information serves as a basis for the aforementioned assumption. The second assumption is that ***statistical n-gram analysis of POS tags can yield useful information to approximate text readability levels***. This assumption is drawn out from the other studies discussed in Chapter 2, such as Schwarm and Ostendorf (2005) and Heilman et al. (2007).

4.2 Datasets

One of the challenges in this research domain is creating a suitable dataset to model and test readability levels of reading materials. The datasets used in this study are based on the Philippines' educational system. These were acquired through collaboration with the English Department coordinators of University of the Philippines Integrated School's (UPIS) primary (i.e. Grades 3 to 6) and secondary (i.e. Grades 7 to 10) school levels. UPIS is a public school in the Philippines and functions as a laboratory school for the University of the Philippines, College of Education. Future researchers who wish to use the datasets in this study should acquire written permission from the UPIS before using them.

There are two categories of data in this thesis. The first one is composed of English

essays written by high school students. Under this category, we have the *2010 Grades 7-9*, *2014 Grades 3-6*, *2014 Grades 7-9*, and the *2014 Grades 3-9* (i.e. Full Range Dataset) datasets. These are used to model student reading abilities per school grade level. Each of these datasets is divided into two, $\frac{2}{3}$ for **training** and $\frac{1}{3}$ for **test**. The second data category is the teacher-prepared instructional materials which we call the *Reference Reading Materials* (i.e. Ref. Reading Mats). These materials are selected by the schools' instructional materials experts and are classified from grade 3 to grade 9. In the experiments, these are used to create the **reference** set for both the training and testing processes which will be discussed in the later sections of this chapter. Summary of these datasets are shown in Table 4.1.

Table 4.1: Summary of Datasets Used

Dataset	**Grade3**	**Grade4**	**Grade5**	**Grade6**	**Grade7**	**Grade8**	**Grade9**	**Total**
2010 Gr 7-9	-	-	-	-	47	54	112	213
2014 Gr 7-9	-	-	-	-	67	62	64	193
2014 Gr 3-6	27	64	96	46	-	-	-	233
2014 Gr 3-9	27	64	96	46	67	62	64	426
Ref. Reading Mats	9	10	6	10	12	6	10	63

4.3 Sampling

Sampling is another very important factor considered in the implementation of the system. For both the 2010 Grades 7-9 and 2014 Grades 7-9 datasets, a stratified 3-fold cross-validation is implemented, such that, essays in each grade level (i.e. Grade7, Grade8, Grade9) are roughly divided into three equal static partitions. In each run, one part is set aside for testing and the other two for training. Note that since there are 3 grade levels with 3 partitions each, 27 test-training combinations are created to exhaust all possible partition combinations with 1:2 test-to-training partition ratio

for each grade level.

The 2014 Grades 3-6 and the 2014 Grades 3-9 datasets have a different sampling procedure since applying the previously discussed procedure will yield too many test-training combinations. For these datasets, R software's *caret* package is used to derive the random stratified 3-fold cross-validation sets. This implements a standard stratified 3-fold data splitting which also makes sure that each sample has a chance to be part of both the test and training sets. For each run of this splitter, three distinct test-train combinations are created. The splitter is then executed ten times to produce 30 test-train distinct combinations for the experiments.

4.4 Preliminary Processing

Stopwords (e.g. 'a', 'an', 'the') removal and stemming (i.e. process for reducing inflected (or derived) words to their word stem, base or root form (Manning, Raghavan and Schütze, 2009; Meyer, Hornik and Feinerer, 2008)) are not implemented in the systems. As mentioned in Schwarm and Ostendorf (2005), stopwords tend to be more prevalent in lower grade levels making it an essential characteristic which we need to consider. Furthermore, since word variations tend to be more extensive as grade level increases, this can also be used as an indicator of text complexity. Hence, stemming has been disabled in all experimental setups to capture that information. Therefore, the only preliminary processing step done in this study is the tokenisation of the text documents used for training and testing the systems. All tokens in the documents are taken as they are and all are considered as valid tokens.

4.5 Content-based Analysis

4.5.1 Matrix Representation

After creating the vocabulary list from text samples (i.e. documents), the three sets (i.e. training, test and reference) are converted to their term-by-document matrix representations, where a term corresponds to a word token. In this representation, each column is equivalent to one document vector, each row represents a word or term vector, and each entry in the matrix is the number of occurrences of each term in each document. This step yields three matrices corresponding to the training, test and reference sets.

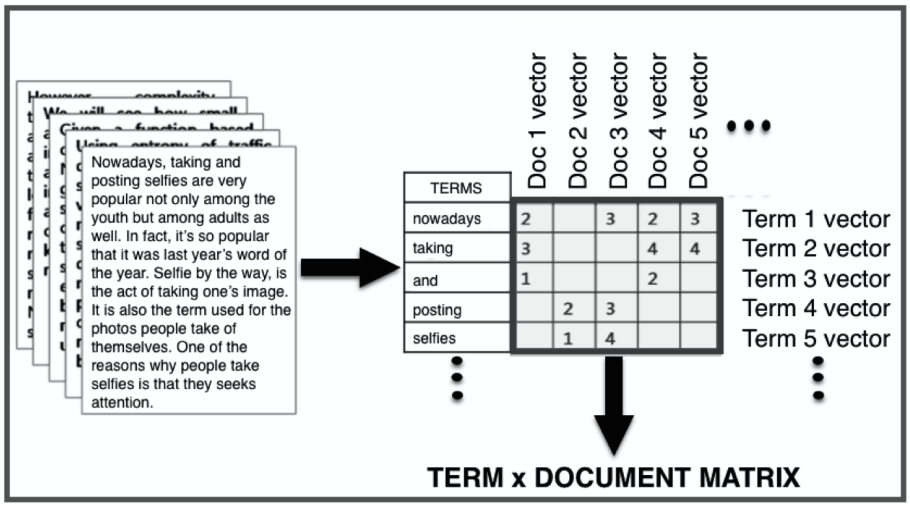


Figure 4.1: Term-by-Document Matrix

4.5.2 Dimensionality Reduction (Dobša and Dalbelo-Bašić, 2004; Garcia, 2006; Razon, 2010)

As discussed in Razon (2010), both LSI and CI dimensionality reduction strategies are implemented separately on the training sets. These are *Singular Value Decomposition*

(SVD) for LSI and *Concept Decomposition* (CD) for CI.

4.5.2.1 LSI's Singular Value Decomposition

SVD is defined as the decomposition of matrix X using:

$$X = UDV^T, where \tag{4.5-1}$$

$$U = eigenvectors\ of\ XX^T, \tag{4.5-2}$$

$$V = eigenvectors\ of\ X^TX, \tag{4.5-3}$$

V^T is the transpose of matrix V, and D is a matrix whose diagonals are the singular values of matrix X (i.e. square root of the eigenvalues of X).

Dimension reduction is accomplished by choosing only the top k biggest singular values of matrix D and setting the rest to zero, resulting in the new reduced matrices D_k, U_k and V_k^T. Optimisation of the dimensionality reduction process is done by empirically finding the optimal value for k.

4.5.2.2 CI's Concept Decomposition (Dobša and Dalbelo-Bašić, 2004; Razon, 2010)

CD is defined as the decomposition of matrix X using:

$$X = CZ^* \tag{4.5-4}$$

Matrix C is the reduced column vector representation of the training set. It is derived as:

1. Set the number of vectors, j, to represent each grade level. In this thesis, this number is also referred to as the number of *sub-clusters* (i.e. CI's *sub*

parameter) per grade level. A sub-cluster, p_j, corresponds to a set of documents belonging to the same grade level and a grade level can be represented using one or more sub-clusters, p_j. In our experiments, the number of sub-clusters per grade level (i.e. j) ranges from 1 to 5. This sub-clustering strategy enables our system to represent each school grade level using 1 to 5 reading ability levels, corresponding to the aforementioned sub-clusters. For the same reasons stated in Razon (2010) (i.e. simplicity and speed in implementation), *K-Means* clustering algorithm (Jin and Han, 2010) was used to determine these sub-clusters.

2. For each sub-cluster, derive the mean concept vector, m_r, using:

$$m_r = \frac{1}{n} \sum_{x \in p_j} x_n \tag{4.5-5}$$

where x_1, x_2, x_3, ..., x_n are the term frequency values of matrix X in one sub-cluster, and n is the total number of documents in that sub-cluster. This step will produce r number vectors, where $r = j * g$, in which j is the number of sub-clusters per grade level and g is the number of grade levels.

3. Normalise each of the mean concept vectors, m_r, and get its corresponding c_r using:

$$c_r = \frac{m_r}{\|m_r\|} \tag{4.5-6}$$

4. Construct matrix C by putting together all the c_rs as its column vectors. Consequently, C will be a term-by-r matrix.

$$C = [c_1 \;\; c_2 \;\; c_3 \; ... \; c_r] \tag{4.5-7}$$

The next task is to find Z^* that minimises the distance between semantic space spanned by CZ^* and X, which leads us to using the equation:

$$Z^* = \arg_Z \min \|X - CZ^*\|_F^2 \tag{4.5-8}$$

where F depicts the Frobenius norm.

As proven in *Linear Algebra*, this minimum distance is equivalent to the projection of X onto the column space of C. Therefore, we would like to find a matrix CZ^* which is the projection of matrix X onto the column space of matrix C. This problem is otherwise known as the *Least Squares Problem* with an approximate solution called the *least squares approximation* given by the equation:

$$Z^* = (C^T C)^{-1} C^T X \tag{4.5-9}$$

After the matrix operation above, Z^* will be an r by N matrix (r=total number of sub-clusters, N=total number of documents in the training or test set) whose columns are the projections of the document vectors onto the reduced semantic space (i.e. column space of C) (Razon, 2010).

4.5.3 Folding-In

Folding-in refers to the projection of the original training, test and reference document vectors onto the reduced semantic space derived in the previous step. For LSI, as discussed in Garcia (2006), this process involves using the equation:

$$q_{reduced} = q_i^T U_k D_k^{-1} \tag{4.5-11}$$

for all document vectors, q_i, of the original training, reference and test sets, where U_k and D_k are the respective low rank approximations of the matrices U and D, derived by keeping only the top k largest singular values in D.

For CI, as discussed in Dobša and Dalbelo-Bašić (2004), we use the equation:

$$q_{reduced} = (C^T C)^{-1} C^T q_i \tag{4.5-12}$$

for all document vectors, q_i, of the original training, reference and test sets.

4.5.4 Similarity Measurement

For each training document vector A, we find its cosine similarity with each reference document vector R as shown in Equation 4.5-13. These similarity values for A form a "similarity-to-reference" vector for A, as shown in Figure 4.2. Then, we take the set of such similarity vectors over all the documents of the training set and put them together in one matrix. After which, we do the same procedure for all test set documents. Consequently, this step yields two matrices composed of similarity-to-reference vectors, one for the training set and the other for the test set.

$$sim(A, R) = \frac{A \cdot R}{|A||R|} \tag{4.5-13}$$

4.6 POS-based Grammar Analysis

Grammar-based features can provide useful information in text analysis. As part of our working assumptions discussed in Section 4.1, POS n-grams can be used to provide a rough approximation of the texts' syntactic information at the least. For example, POS unigrams can provide information regarding which of the POS tags

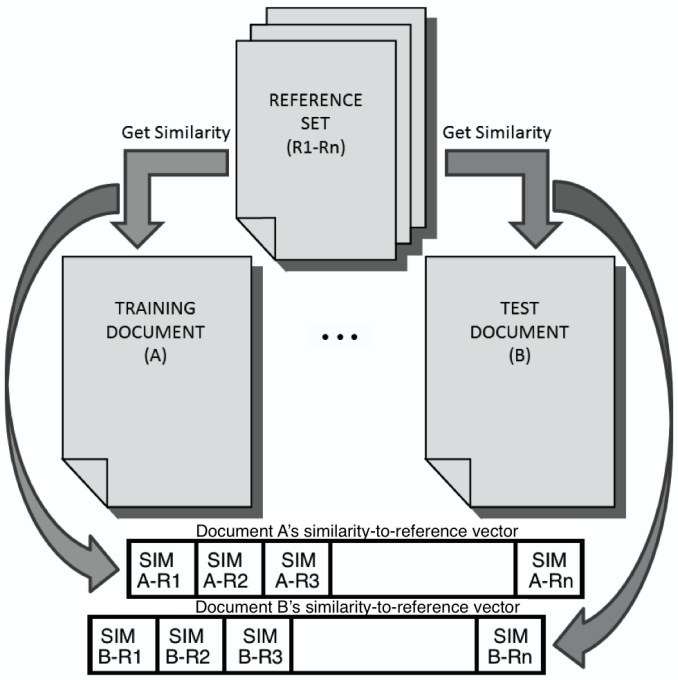


Figure 4.2: Similarity Vector Diagram

are prevalent for each grade level and which are not. Moreover, POS bi- and tri-grams can capture grammar-related information which can serve as a basis for syntax complexity.

Apache OpenNLP *Maximum Entropy POS Tagger* (i.e. `Maxent_POS_Tag_Annotator`), together with its Maximum Entropy Sentence and Word Annotators (i.e. `Maxent_Sent_Token_Annotator` and `Maxent_Word_Token_Annotator`), is used to tag all documents in this study (Apache OpenNLP, 2015; Hornik, 2016). The complete list of tags for this tagger is given in Appendix C.

After getting uni-, bi- and tri-gram POS features, we construct the term-by-document matrix for each of these, where the POS n-grams are treated as the terms of the matrices. Thus, we call this *POS n-gram-by-document matrix*. *Sparsification Strategy* (SS) is then conducted on these matrices.

SS is the removal of sparse term vectors (i.e. the exclusion of n-gram row term vectors which have mostly zero values). This procedure aims to reduce the dimensionality of the POS n-gram-by-document matrix without sacrificing the loss of significant information inherent in the matrix. In this study, the term *sparsity* refers to the maximum sparse percentage, called the *sparsity index* (SI), to consider in the experiment. For example, SI value of 0.7 means that all term vectors which are 70% sparse (i.e. 70% of elements in the vector are zero) and below will be considered. Therefore, higher sparsity values allow more POS n-gram vectors to be included in the analysis.

4.7 The SVM Classifier

SVMs have been successfully implemented in numerous text classification tasks which usually involve separating data into training and test sets, as mentioned in Hsu, Chang, and Lin (2003). As also stated in the same literature, the main goal of SVM is to create a model which has the ability to predict the classification of unseen test data based on several attributes, i.e. features, taken from the training set. We chose SVM for two main reasons, namely, 1.) SVMs are proven to perform well on text classification tasks as manifested in Schwarm and Ostendorf (2005) and Wang (2006), and 2.) There are several references and resources readily available for the implementation of SVMs.

In this study, Radial Basis Function (RBF) or Gaussian is used as the kernel function for all SVM classifiers in all the experiments. The kernel's flexibility in handling linear and nonlinear relationships between output class labels and features makes it a practical choice for this study. As stated in Hsu, Chang, and Lin (2003), a linear kernel can be viewed as a special case of RBF since the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some

parameters C and γ. It can also map samples into higher dimensional space as explained in the same literature, thus handling the nonlinear case.

Our SVM with RBF kernel function has the following parameters:

1. γ: kernel parameter which controls the width of the Gaussian (i.e. width of the Gaussian is inversely proportional to γ)

2. C: misclassification cost or penalty constant which is used as the regularisation parameter

3. k: number of folds in training cross-validation (i.e. k-fold cross-validation constant)

In classification tasks, the γ parameter defines how far the influence of a single training example reaches, with low values having high reach and high values having low reach (Hsu, Chang, and Lin, 2003). The C parameter, however, can be seen as the regularisation constant. It controls the trade off between misclassification of training examples and simplicity of the decision surface, with low values making the decision surface smooth and high values making it considerably more wiggly but with more correctly classified training examples (Hsu, Chang, and Lin, 2003).

To determine suitable values for the γ and C parameters, we conducted exploratory experiments using the built-in SVM parameter grid search function in R software's `e1071` package, the *tune.svm()*. The function takes in a set of γ (e.g. $\gamma = 2^{(-10\ to\ 1)}$) and C (e.g. $C = [0\ to\ 100]$) values and creates a grid search space for these values. Then, it outputs the best paired values of these parameters which yield the lowest classification error in its built-in 10-fold cross-validation of the training set.

To account for all the basic system setups (i.e. CI-, LSI-, POS-based setups) in this research, we identified candidate C and γ paired values using the function discussed

above. From these candidates, we selected the most frequently occurring C value for each dataset, i.e. 10 and identified the range of values for the γ parameter which falls between 0.001 and 1.0. Keeping C=10, we conducted experiments per dataset for the γ parameter using values 0.001, 0.01, 0.1 and 1.0 to represent the aforementioned range of values. Based on the results, γ=0.1 yielded the highest accuracy values for the LSI- and CI-based systems. Doing a similar procedure for the POS-based systems, we achieved the best result using γ=0.001. Details of the experiments on C and γ parameters are presented in Appendix A.

4.8 Performance Metrics

The primary metric used to measure the performance of the system is the *Mean Exact Agreement Accuracy* or MEAA. For each test-training combination mentioned in Section 4.3, the fraction of documents in the test set having the same actual and predicted grade level classification is calculated. This fraction is the *Exact Agreement Accuracy* (EAA).

$$EAA = \frac{No.\ of\ test\ docs\ with\ the\ same\ actual\ and\ predicted\ grade\ level}{Total\ number\ of\ test\ docs} \quad (4.8\text{-}1)$$

To get the MEAA, the sum of EAA for all the test-training combinations is divided by the total number of combinations.

$$MEAA = \frac{Sum\ of\ all\ EAA\ for\ all\ test-training\ combos}{Total\ number\ of\ test-training\ combos} \quad (4.8\text{-}2)$$

We also measured the standard deviation (SD) across the EAA values. SD quantifies the extent of variation or dispersion of data values. A low SD value indicates that the EAA data points are very close to the MEAA, thus making the latter a good

representation of the system's overall performance.

Lastly, we also mention the *Adjacent Agreement Accuracy* (AAA) and Mean AAA metrics in Phase 3-B of the experiments discussed in the next chapter. AAA will be defined as the fraction of essays in the test set which are wrongly classified by the system by only one (1) grade level (e.g. a grade 7 essay classified as a grade 8).

$$AAA = \frac{No.\ of\ wrongly\ classified\ test\ docs\ by\ one\ grade\ level}{Total\ number\ of\ test\ docs} \tag{4.8-3}$$

Correspondingly, to get the MAAA, the sum of AAA for all the test-training combinations is divided by the total number of combinations.

$$MAAA = \frac{Sum\ of\ all\ AAA\ for\ all\ test-training\ combos}{Total\ number\ of\ test-training\ combo} \tag{4.8-4}$$

4.9 Chapter Summary

In this chapter we presented the details on the research methods we used in this study. We started by establishing our research assumptions. Then, we discussed the datasets which are composed of the 4 essay datasets (i.e. 2010 Grades 7-9, 2014 Grades 7-9, 2014 Grades 3-6, 2014 Grades 3-9) and the Reference Reading Materials dataset. Next, we discussed the sampling procedures implemented on the essay datasets. Details on the LSI, CI, POS and SVM algorithms were also presented in this chapter. Finally, we provided the metrics we are going to use to measure the performance of the systems in our experiments. These experiments are going to be discussed in the following chapter of this thesis.

Chapter 5

Experiments and Results

In this chapter, we give the full details on the experiments we conducted. Section 5.1 presents the list of feature sets we used and the experimental phases we undertook to develop our proposed system. Section 5.2 provides a comprehensive discussion on the results per experimental phase. Lastly, we wrap up this chapter in Section 5.3 by giving a summary of all the experimental results with their corresponding analyses and implications.

5.1 Feature Sets and Phases of Experiments

Five (5) feature sets are investigated in this study. These are:

1. POS: POS n-gram features only
2. LSI: LSI-based features only
3. CI: CI-based features only
4. LSI+POS: Combined LSI-based and POS n-gram features
5. CI+POS: Combined CI-based and POS n-gram features

With these feature sets, the following experimental phases are implemented using the similarity-to-reference matrices discussed in Section 4.5.4.

1. Phase 1: Experiments on Isolated Feature Sets — Baseline Experiments
    (a) POS
    (b) LSI
    (c) CI
2. Phase 2: Experiments on Combined Grammar and Content Features with SI = 1.0
    (a) LSI+POS with SI=1.0
    (b) CI+POS with SI=1.0
3. Phase 3: Experiments on POS n-gram Sparsification
    (a) POS with SI from 0.1 to 0.9
    (b) LSI+POS with SI from 0.1 to 0.9
    (c) CI+POS with SI from 0.1 to 0.9
4. Phase 4: Error Analysis
    (a) Overestimation Error
    (b) Underestimation Error

In Phase 1, we used feature sets 1, 2, and 3 which are the isolated feature sets LSI, CI and POS, respectively. This phase will serve as the baseline experiments of the study.

Phase 2 involved the integration of the POS feature set into the LSI and CI feature sets, separately, resulting in the LSI+POS and CI+POS feature sets. In this phase, the SS was not yet implemented which equivalently makes the SI value equal to 1.0.

An investigation into the effect of the sparsity index (SI) applied on the POS n-gram features is performed in Phase 3 to optimise the LSI+POS and CI+POS combination processes. Finally, in Phase 4, we take another step forward to analyse the errors in the optimal system setup derived from the previous phase.

5.2 Results of Experiments

In this section, we will present the results of the 3 phases of our experiments. As explained in Section 4.3, the following MEAA values are computed using the results of the 27 random sets for the 2010 and 2014 Grades 7-9 datasets, and the 30 random sets for the 2014 Grades 3-6 and Grades 3-9 datasets. These values are shown in Figures 5.1 to 5.14, along with their corresponding SDs which are represented as error bars.

To test for statistical significance of the results, we used the Wilcoxon Matched Pairs Signed-Rank Test (Hollander, Wolfe and Chicken, 2013) with a significance threshold of p-value=0.05. In the following discussions, a p-value lower than 0.05 means that there is strong evidence that the difference between the outputs is significant, otherwise, this difference can just be attributed to chance.

5.2.1 Phase 1: Baseline Experiments

Baseline experiments are those experiments done using isolated feature sets (i.e. feature sets 1, 2, and 3 as mentioned in Section 5.1). For both the 2010 and the 2014 Grades 7-9 datasets, CI with sub=2 achieved the highest MEAA values of 0.897 and 0.934, respectively. Significance tests between CI and LSI and between CI and POS yielded p-values of $1.49e^{-08}$ and $5.04e^{-04}$, respectively, for the 2010 Grades 7-9 dataset. In this case, we can say that CI outperformed both LSI and POS. For the

2014 Grades 7-9 dataset, the p-values are found to be 0.252 and $2.086e^{-06}$ for CI versus LSI and CI versus POS, respectively. With these p-values, we can say that CI outperformed POS, however, there is not enough evidence that it was also able to outperform LSI for this dataset.

The highest MEAA values of 0.897 and 0.830 for the 2014 Grades 3-6 and 2014 Grades 3-9 datasets (i.e. datasets involving primary school levels 3 to 6), respectively, were achieved using POS features with p-values=$1.863e^{-09}$ against LSI and CI outputs. Therefore, we can infer that the inclusion of essays written by students in grades 3 to 6 made the POS-based features the more informative and the more discriminant ones over content-based features.

Table 5.1: Phase 1: Baseline Experiment Summary

Feature Set	**Primary Param.**	**2010 Gr7-9**		**2014 Gr7-9**		**2014 Gr3-6**		**2014 Gr3-9**	
		MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD
POS n-gram	n=1, uni	0.749	0.064	0.786	0.096	0.747	0.042	0.602	0.038
	n=2, bi	**0.854**	0.027	**0.874**	0.041	0.881	0.029	**0.830**	0.032
	n=3, tri	0.853	0.035	0.845	0.044	**0.897**	0.030	0.822	0.026
CI	sub=1	0.891	0.052	0.933	0.039	**0.815**	0.031	**0.691**	0.026
	sub=2	**0.897**	0.051	**0.934**	0.041	0.809	0.039	0.689	0.038
	sub=3	0.884	0.071	0.931	0.042	0.806	0.037	0.687	0.036
	sub=4	0.873	0.045	0.927	0.042	0.805	0.031	0.685	0.033
	sub=5	0.882	0.053	0.929	0.043	0.806	0.040	0.677	0.030
LSI	dim=0.1	0.647	0.033	0.770	0.046	**0.726**	0.033	0.419	0.035
	dim=0.2	**0.751**	0.062	**0.925**	0.029	0.648	0.047	**0.612**	0.043
	dim=0.3	0.741	0.059	0.874	0.031	0.627	0.039	0.572	0.034
	dim=0.4	0.696	0.050	0.757	0.046	0.622	0.037	0.537	0.040
	dim=0.5	0.683	0.054	0.781	0.056	0.636	0.045	0.529	0.034
	dim=0.6	0.660	0.055	0.783	0.050	0.629	0.036	0.540	0.032
	dim=0.7	0.666	0.040	0.798	0.048	0.616	0.043	0.534	0.039
	dim=0.8	0.659	0.044	0.789	0.053	0.622	0.049	0.522	0.042
	dim=0.9	0.655	0.039	0.785	0.060	0.613	0.043	0.512	0.035

Referring to Table 5.1, we can also generally say that POS bi-gram and tri-gram features create more stable performing models across different datasets with varying

number of grade levels than POS uni-grams since the MEAA values achieved on these models are all above 0.800. Although these values were not very high, this stability in performance proves that POS bi-gram and tri-gram features are more reliable than uni-grams.

Another important observation is that CI was able to perform well on datasets with only 3 or 4 adjacent grade levels (i.e. 2010 Grades 7-9, 2014 Grades 7-9 and 2014 Grades 3-6 datasets). Its performance dramatically decreased on the 2014 Grades 3-9 dataset which is composed of seven (7) grade levels. The same behaviour can also be observed with the LSI models which also have their lowest MEAA values on the 2014 Grades 3-9 dataset.

Detailed discussions of the individual baseline experimental results for each dataset will be provided in the next subsections. Figures 5.1, 5.2, 5.3 and 5.4 give graphical representations of the system's MEAAs across different values of dim, sub and n for the LSI, CI and POS feature sets, respectively. The topmost graphs of each figure represent the LSI outputs. CI outputs are presented as the middle graphs. Lastly, POS outputs are shown on the bottommost graphs of these figures.

5.2.1.1 2010 Gr7-9 Dataset

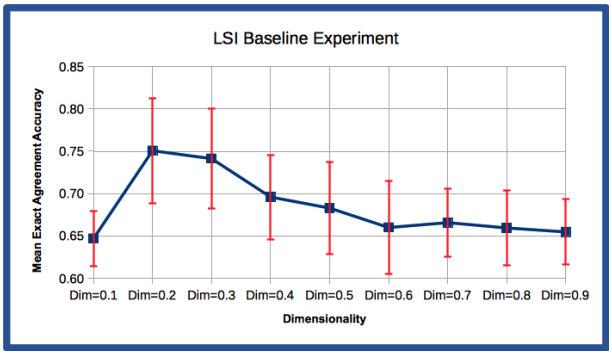
As shown in Figure 5.1a, LSI's lowest point corresponding to its lowest MEAA value of 0.647 is at *dim*=0.1. Increasing *dim* by 0.1, it achieved its highest MEAA value of 0.751 at *dim*=0.2. After which, a decline in MEAA values can be observed as *dim* approaches 0.9. Significance test between its highest and lowest points yielded a p-value of $1.49e^{-08}$.

In Figure 5.1b, CI achieved its highest MEAA of 0.897 at *sub*=2 and its lowest MEAA of 0.873 at *sub*=4, with a p-value=0.045 between these two *sub* values. This 0.024 difference in its MEAA values consequently means that CI with *sub*=2 has only around 2 more correctly classified documents than CI with *sub*=4.

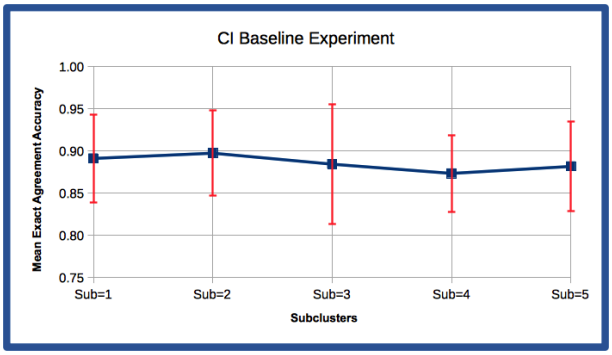
In Figure 5.1c, POS achieved its highest and lowest MEAA at n=2 (i.e. bi-gram) with a value of 0.854 and n=1 (i.e. uni-gram) with a value of 0.749, respectively, with a p-value of $1.49e^{-08}$ between these two n values. These values differ by 0.105 which is equivalent to around 7 out of 71 total documents in this dataset.

Significance test between the results of the POS experiments using n=2 and n=3 (i.e. tri-gram) yielded a p-value of 0.7017. However, it is important to note here that, although the 0.001 difference in the system's MEAA for n=2 and n=3 is negligible and that the p-value between them is higher than 0.05, our statistical significance threshold, the dimensionality of the matrix for n=3 (i.e. 6441-by-213) is almost eight (8) times bigger as that of n=2 (i.e. 842-by-213), making the latter a more practical choice over the other.

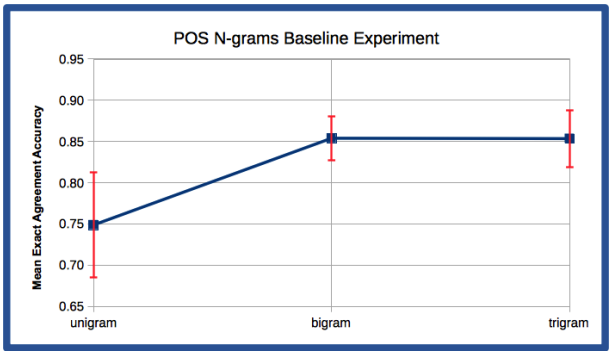
For this dataset, CI has outperformed both LSI and POS by 0.214 and 0.043, respectively. Upon evaluation of the statistical significance of the difference in their performance, corresponding p-values of $1.49e^{-08}$ and $5.039e^{-04}$ were obtained.



(a) LSI



(b) CI



(c) POS

Figure 5.1: Baseline Experimental Results on 2010 Grades 7-9 Dataset

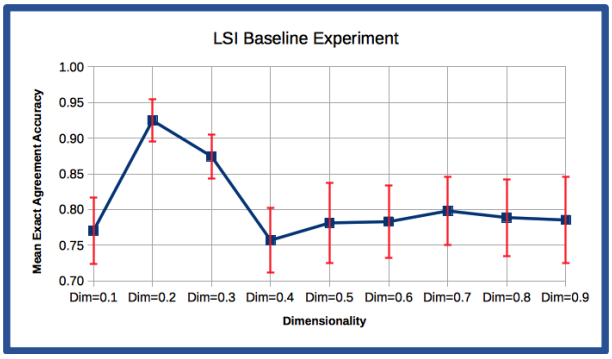
5.2.1.2 2014 Gr7-9 Dataset

As shown in Figure 5.2a, LSI achieved its highest MEAA at dim=0.2 with a value of 0.925 and its lowest MEAA at dim=0.4 with a value of 0.757. This 0.168 difference between its highest and lowest MEAA with a p-value of $1.49e^{-08}$ demonstrates the system's responsiveness to its dim parameter as also exhibited in the 2010 Grades 7-9 LSI results.

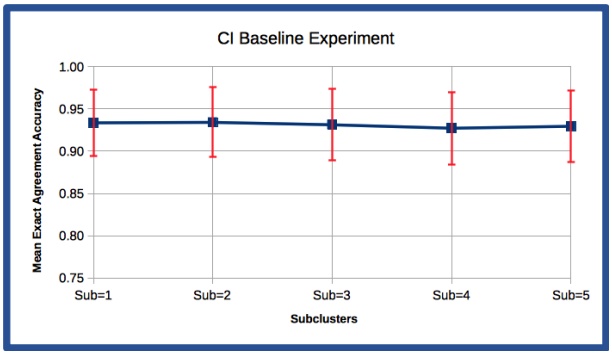
CI achieved its highest MEAA value of 0.934 at sub=2 as can be seen in Figure 5.2b. This is the overall highest MEAA value that the system was able to achieve among all the baseline experiments on all datasets. Additionally, CI's lowest MEAA occurred at sub=4 with a value of 0.927, which is still relatively higher than those achieved in other datasets. Statistical significance test between the results of these two sub values yielded a p-value of 0.1431, which implies that there is not enough evidence that CI with sub=2 is better than CI with sub=4.

As shown in Figure 5.2c, the POS-based feature set achieved its highest MEAA value of 0.874 at n=2 (i.e. bi-grams) and its lowest MEAA value of 0.786 at n=1 (i.e. uni-grams). Like LSI, it exhibited a relatively larger gap of 0.088 between its highest and lowest MEAA with a p-value of $9.266e^{-05}$, which in effect means that the system is also sensitive to the n parameter. This finding is also consistent with that on the 2010 Grades 7-9 dataset discussed earlier.

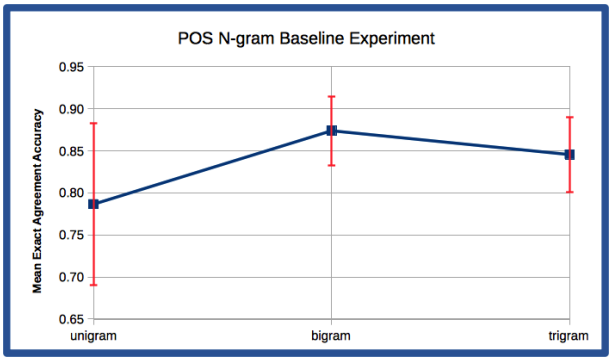
For this dataset, CI has outperformed POS by 0.060 with a p-value=$2.086e^{-06}$. It has successfully classified approximately 60 out of 64 test documents using 129 training documents. However, we can not claim that CI has also outperformed LSI since the significance test between them yielded a p-value of 0.252.



(a) LSI



(b) CI



(c) POS

Figure 5.2: Baseline Experimental Results on 2014 Grades 7-9 Dataset

5.2.1.3 2014 Gr3-6 Dataset

Figure 5.3 presents the outputs for the 2014 Grades 3-6 (i.e. primary school level) dataset. From top to bottom, it shows the graphs of the MEAA values against parameters *dim*, *sub* and n of the LSI, CI and POS feature sets, respectively.

As shown, the highest MEAA value of 0.726 was achieved by LSI at *dim*=0.1, while its lowest MEAA value of 0.613 was achieved at *dim*=0.9. Using these two *dim* values, 0.1 and 0.9, we found that there is significant difference between their corresponding results (i.e. p-value=$1.863e^{-09}$).

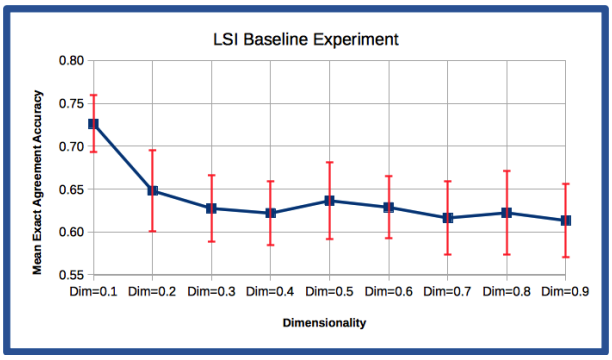
CI exhibited a smooth decreasing pattern in its MEAA values from *sub*=1 to *sub*=4 which slightly increased at *sub*=5. Its highest MEAA value of 0.815 was achieved at *sub*=1, while its lowest MEAA value of 0.805 was achieved at *sub*=4. Note that the 0.010 difference between these two MEAA values only accounts for approximately 1 out of 78 test documents in this dataset. Significance test between the results of CI with *sub*=1 and CI with *sub*=4 yielded a p-value of 0.039, which still implies that there is significant difference between them.

POS achieved its highest MEAA for this particular dataset among all the other datasets. For n=2 (i.e. bi-grams) and n=3 (i.e. tri-grams), it was able to achieve 0.896 and 0.897, respectively, with a p-value of 0.7317 which means that the difference between the results can just be attributed to chance. However, because of the very huge difference between the dimensionality of the n=2 and the n=3 matrices, which are 689-by-78 and 4597-by-78, respectively, we can say that using n=2 is a more practical choice. Additionally, lowest MEAA for the POS was achieved at n=1 (i.e. uni-grams) with a value of 0.701 and a p-value of $1.863e^{-09}$ against n=2.

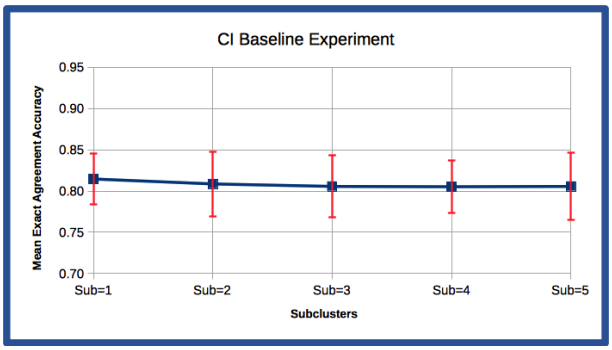
Note that the small difference between CI's highest and lowest MEAA values, along with the relatively larger gap for LSI and POS between these two values, is

consistent with the observation presented earlier in the discussion of the 2010 Grades 7-9 and 2014 Grades 7-9 datasets. Therefore, the outputs of this experiment further validate that the system is sensitive to the *dim* parameter of LSI and the n parameter of POS, but not to the *sub* parameter of CI.

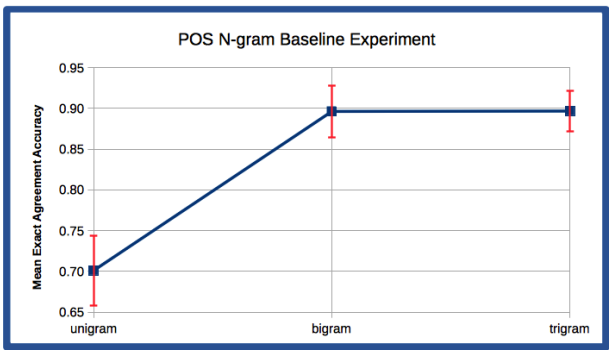
For this dataset, POS has significantly outperformed both LSI and CI by 0.261 and 0.082, respectively, with both p-values=$1.863e^{-09}$. It has successfully classified around 70 out of 78 test documents using 155 training documents.



(a) LSI



(b) CI



(c) POS

Figure 5.3: Baseline Experimental Results on 2014 Grades 3-6 Dataset

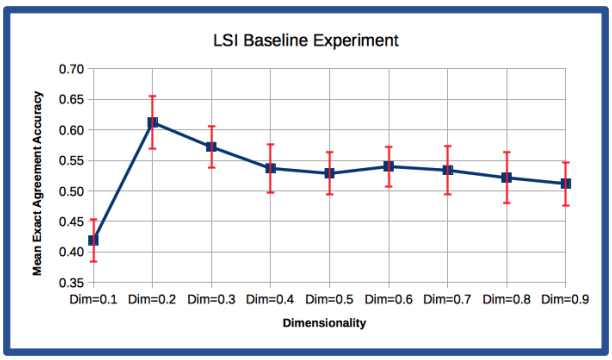
5.2.1.4 2014 Gr3-9 Dataset

The 2014 Grades 3-9 dataset is the biggest one in our research. It is composed of the combined 2014 Grades 3-6 and 2014 Grades 7-9 datasets. Figure 5.4 presents its MEAA output graphs for the LSI (i.e. topmost graph), CI (i.e. middle graph) and POS (i.e. bottommost graph) baseline experiments.

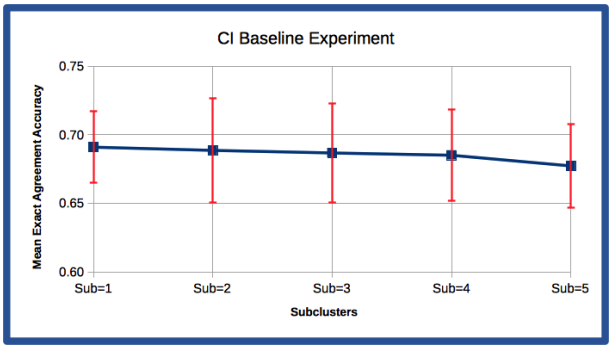
As shown in Figures 5.4a, LSI's output reached its peak MEAA value of 0.612 at dim=0.2 and achieved its lowest MEAA value of 0.419 at dim=0.1. CI's output graph (i.e. Figures 5.4b), however, started out with its highest MEAA value of 0.691 at sub=1 and then gradually decreased until it reached its lowest MEAA value of 0.677 at sub=5. Lastly, POS achieved its highest MEAA of 0.830 at n=2, while its lowest MEAA is at n=1 with a value of 0.602, as can be seen in Figures 5.4c.

The outputs of the experiments on this dataset exhibit similar behaviour to the other datasets. For example, CI also achieved small swings between its maximum and minimum MEAAs across different sub values. Additionally, LSI and POS demonstrated large gaps of 0.193 and 0.228, respectively, between their highest and lowest points.

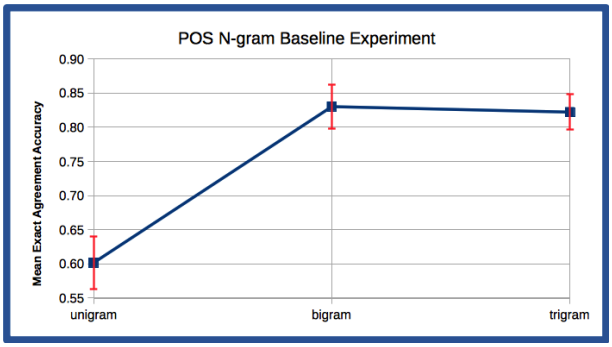
In summary, POS has outperformed both LSI and CI on this dataset by 0.290 and 0.139, respectively, with both p-values=$1.863e^{-09}$. It has successfully classified around 118 out of 142 test documents using 284 training documents.



(a) LSI



(b) CI



(c) POS

Figure 5.4: Baseline Experimental Results on 2014 Grades 3-9 Dataset

5.2.2 Phase 2: Experiments with Combined Features

Phase 2 experiments involve the direct combination of content- (i.e. from LSI and CI) and grammar-based (i.e. POS n-gram) feature sets with SI equal to 1.0 which means that the SS is not implemented. Figures 5.6 and 5.5 present the system outputs for LSI+POS and CI+POS experiments conducted on all the datasets, respectively. The goal of this phase is to verify if merely combining feature sets would yield better performing systems.

Table 5.2 summarises the highest MEAAs achieved per dataset in this phase with significant difference from the results of the previous phase. For comparison purposes, we have also included in the table the highest MEAAs achieved per dataset for the baseline experiments.

Based on the results, LSI's performance has been enhance for all datasets except for the 2014 Grades 7-9 dataset. Its MEAA values increased by 0.104, 0.150, and 0.173 for the 2010 Grades 7-9, 2014 Grades 3-6 and 2014 Grades 3-9 datasets, respectively. However, its MEAA dropped by 0.046 for the 2014 Grades 7-9 dataset. MEAA values for the LSI+POS n-gram systems across different values of the *dim* parameter are presented in Figures 5.5a–5.5d.

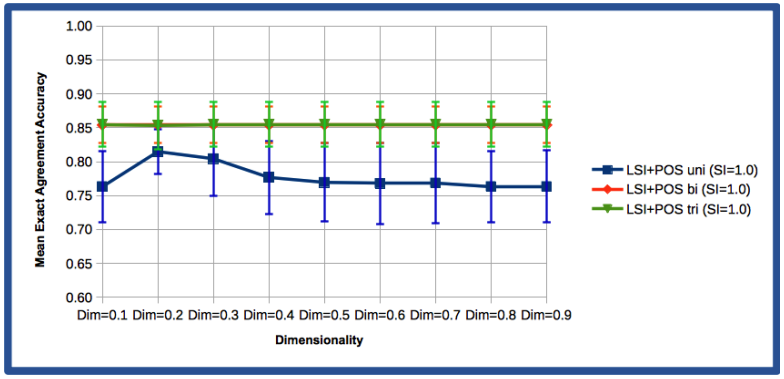
As also shown in Table 5.2, the performance of CI has degraded by 0.043 and 0.060 for the 2010 Grades 7-9 and 2014 Grades 7-9 datasets, respectively. However, the inverse is true for those datasets which include primary school levels (i.e. 2014 Grades 3-6 and Grades 3-9 datasets). For the 2014 Grades 3-6 dataset, CI's MEAA increased by 0.082. Similarly, its MEAA increased by 0.131 for the 2014 Grades 3-9 dataset. With these, we can say that the integration of POS-based features into the CI-based system has only been advantageous on those datasets involving essays from the primary school levels, but not on those datasets involving only the secondary

school levels. MEAA values for the CI+POS n-gram systems across different values of the *sub* parameter are shown in Figures 5.6a–5.6d.

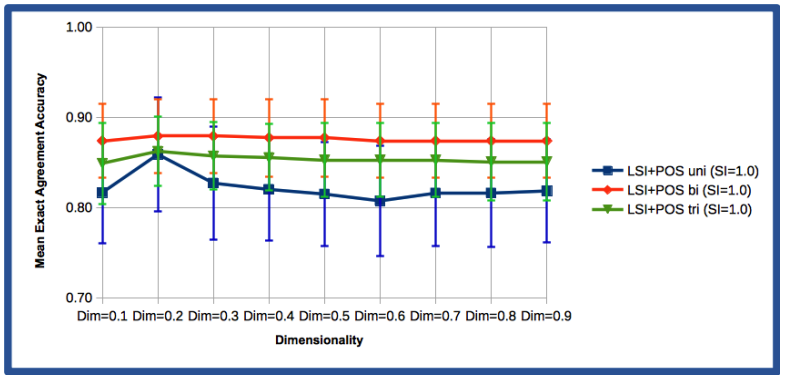
In this experiment, we have presented evidence that merely adding feature sets together can either improve or degrade the performance of the system. In the next phase, further feature analysis will be done in order to improve the combination process of the content- and grammar-based feature sets.

Table 5.2: Summary of the Highest MEAAs Achieved per Dataset in Phase 2 with Significant Difference from Phase 1 Results (p-value<0.05)

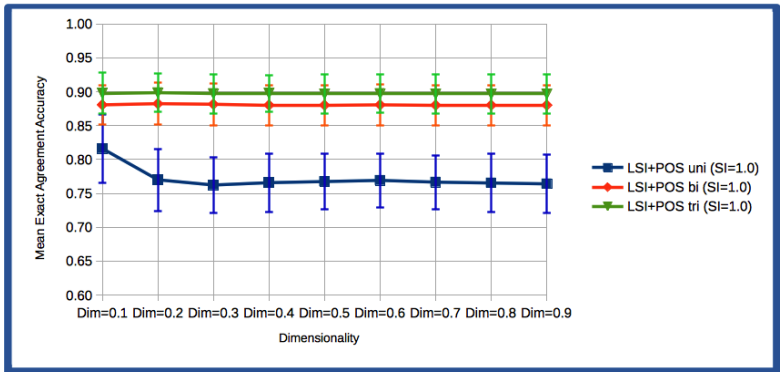
Feature Set	**2010 Gr7-9**		**2014 Gr7-9**		**2014 Gr3-6**		**2014 Gr3-9**	
	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD
Phase1: CI only (Baseline)	*0.897*	*0.051*	*0.934*	*0.041*	*0.815*	*0.031*	*0.691*	*0.026*
Phase2: CI+POS uni (SI=1.0)	0.849	0.067	0.856	0.071	0.791	0.042	0.749	0.044
Phase2: CI+POS bi (SI=1.0)	**0.854**	0.027	**0.874**	0.041	0.881	0.029	0.744	0.028
Phase2: CI+POS tri (SI=1.0)	0.853	0.035	0.845	0.044	**0.897**	0.030	**0.822**	0.026
Phase1: LSI only (Baseline)	*0.751*	*0.062*	*0.925*	*0.029*	*0.726*	*0.033*	*0.612*	*0.043*
Phase2: LSI+POS uni (SI=1.0)	0.815	0.033	0.859	0.063	0.816	0.051	0.716	0.037
Phase2: LSI+POS bi (SI=1.0)	0.854	0.027	**0.879**	0.041	0.882	0.031	0.749	0.028
Phase2: LSI+POS tri (SI=1.0)	**0.855**	0.033	0.862	0.039	**0.899**	0.028	**0.762**	0.032



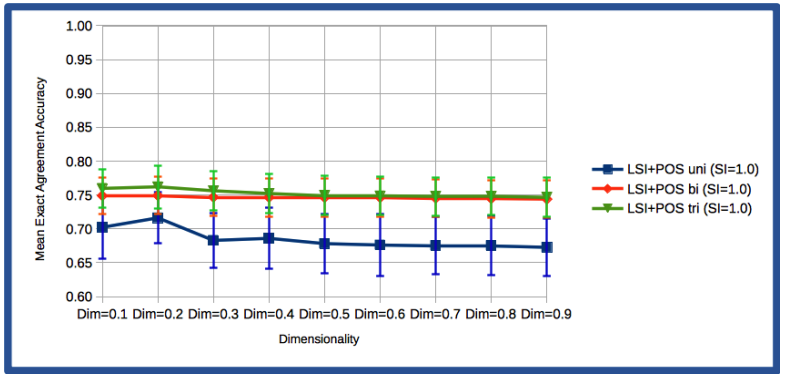
(a) 2010 Grades 7-9 Dataset



(b) 2014 Grades 7-9 Dataset

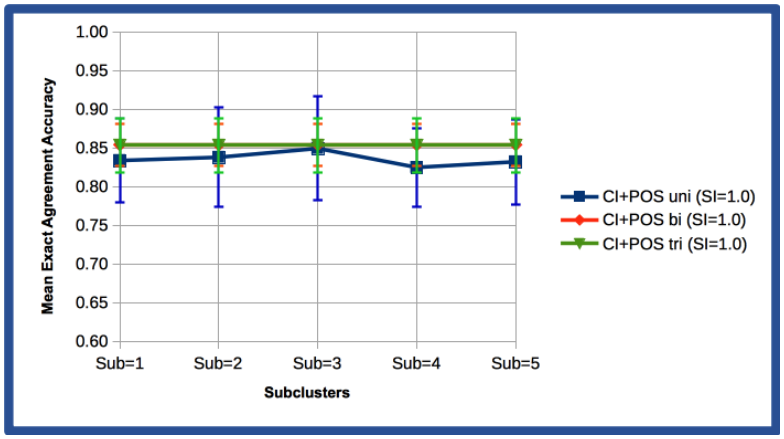


(c) 2014 Grades 3-6 Dataset

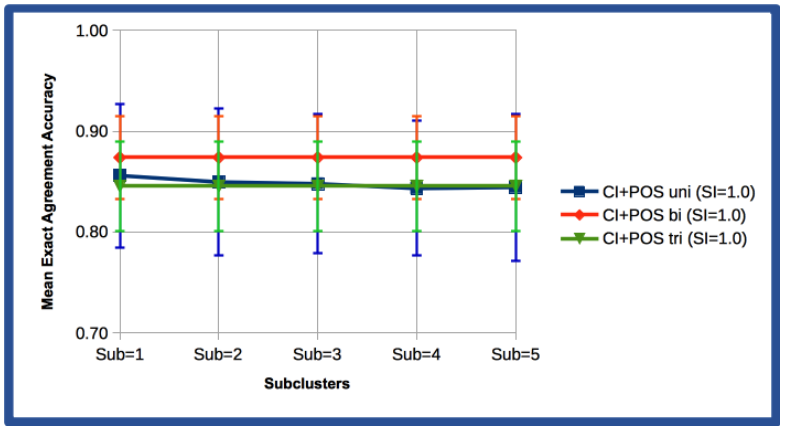


(d) 2014 Grades 3-9 Dataset

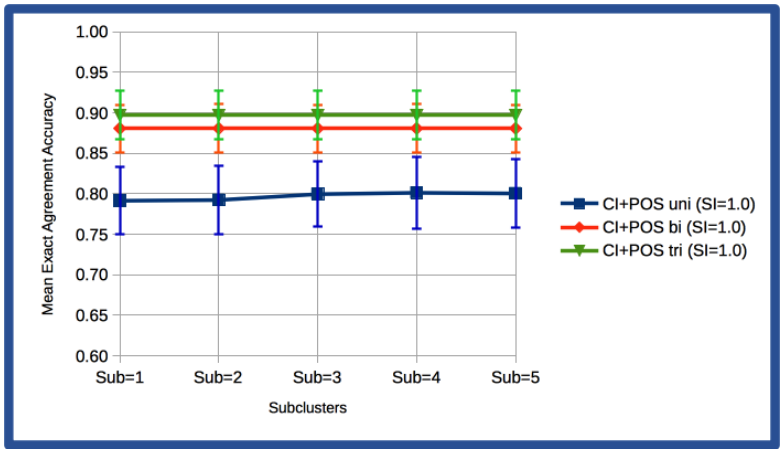
Figure 5.5: LSI+POS with SI=1.0 Experimental Results



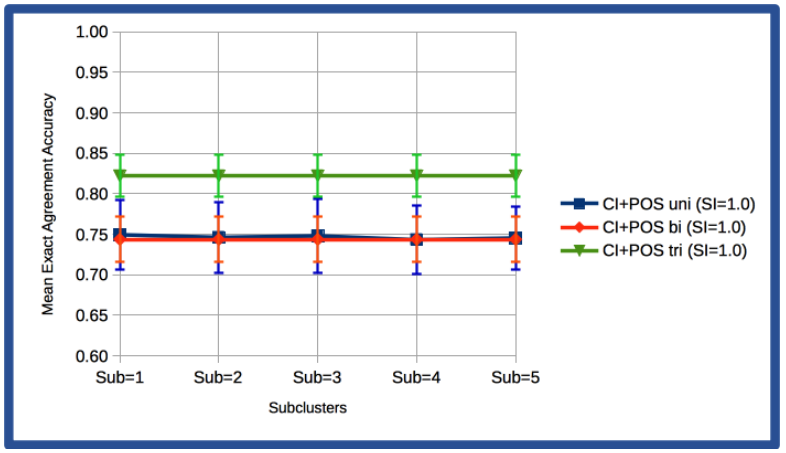
(a) 2010 Grades 7-9 Dataset



(b) 2014 Grades 7-9 Dataset



(c) 2014 Grades 3-6 Dataset



(d) 2014 Grades 3-9 Dataset

Figure 5.6: CI+POS with SI=1.0 Experimental Results

5.2.3 Phase 3: POS n-gram Sparsification

Phase 3 experiments evolved around the implementation of the SS discussed in Section 4.6. SIs from 0.5 to 0.9 were used to observe its effects on the integration of POS n-grams into the LSI- and CI-based systems. Figures 5.7 to 5.14 show the MEAA outputs of LSI+POS and CI+POS systems across different values of LSI's *dim* and CI's *sub* parameters, and the SI parameter of POS. The goal of this phase is to verify if further feature analysis using the SS would improve the performance of the systems.

Tables 5.3 and 5.4 present the summary of the highest MEAAs across LSI's *dim* and CI's *sub* parameters, respectively. These values are found to have statistically significant differences to the values achieved by LSI and CI in Phase 1. As shown, the optimal MEAA for each dataset is never achieved on SI=1.0 (i.e. without SS). This validates our claim that elimination of some features after the combination process is necessary to improve the overall performance of the systems.

The following are the general observations which can be derived from the results.

1. Although systems with integrated POS tri-grams may give out the highest MEAAs for SI=1.0 (i.e. without SS), their performance degrades as SI decreases to 0.5. It can be concluded that such systems are the most affected by the SS.

2. For all datasets, systems using POS bi-grams are the best performing ones across different values of LSI's *dim* and CI's *sub* parameters, and the SI parameter of POS, with p-values$<$0.05 against POS uni- and tri-gram outputs. These systems are also the ones which benefit most from the SS, reaching their highest MEAAs at either SI=0.8 or SI=0.9.

3. Systems with POS uni-grams mostly achieve the lowest rank in performance.

However, it is important to note that these systems, maintaining relatively similar MEAA output patterns across all values of SI, are the least sensitive to the SS. This behaviour can be attributed to the fact that POS uni-gram matrices are mostly dense.

Table 5.3: Summary of the Highest MEAAs Achieved by the LSI-based System on Varying SI Values per Dataset in Phase 3 with Significant Difference from Phase 1 Results (p-value<0.05)

		SI=1.0(Ph.2)		**SI=0.9**		**SI=0.8**		**SI=0.7**		**SI=0.6**		**SI=0.5**	
DATASET	FEATURE SET	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD
2010 Gr 7-9	LSI+POS uni	0.815	0.033	0.814	0.040	0.788	0.033	0.796	0.034	0.798	0.033	0.798	0.033
	LSI+POS bi	0.854	0.027	0.876	0.039	0.889	0.045	**0.899**	0.047	0.888	0.036	0.854	0.053
	LSI+POS tri	0.855	0.033	0.876	0.034	0.867	0.034	0.835	0.033	0.844	0.056	0.787	0.046
2014 Gr 7-9	LSI+POS uni	0.859	0.063	**0.907**	0.033	0.847	0.052	0.844	0.059	0.845	0.053	0.845	0.049
	LSI+POS bi	0.879	0.041	-	-	**0.907**	0.023	0.901	0.028	0.906	0.040	0.891	0.037
	LSI+POS tri	0.862	0.039	0.895	0.035	0.898	0.030	0.893	0.031	0.898	0.034	0.905	0.027
2014 Gr 3-6	LSI+POS uni	0.816	0.051	0.844	0.042	0.844	0.043	0.850	0.038	0.848	0.040	0.848	0.040
	LSI+POS bi	0.882	0.031	**0.924**	0.032	0.912	0.033	0.902	0.038	0.876	0.043	0.870	0.038
	LSI+POS tri	0.899	0.028	0.913	0.033	0.863	0.035	0.825	0.044	0.797	0.046	0.785	0.047
2014 Gr 3-9	LSI+POS uni	0.716	0.037	0.721	0.036	0.720	0.038	0.720	0.038	0.714	0.040	0.709	0.039
	LSI+POS bi	0.749	0.028	0.811	0.027	**0.864**	0.026	0.860	0.029	0.843	0.034	0.805	0.034
	LSI+POS tri	0.762	0.032	0.701	0.037	0.804	0.031	0.757	0.035	0.705	0.038	0.680	0.044

Table 5.4: Summary of the Highest MEAAs Achieved by the CI-based System on Varying SI Values per Dataset in Phase 3 with Significant Difference from Phase 1 Results (p-value<0.05)

DATASET	FEATURE SET	**SI=1.0(Ph.2)**		**SI=0.9**		**SI=0.8**		**SI=0.7**		**SI=0.6**		**SI=0.5**	
		MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD	MEAA	SD
2010 Gr 7-9	CI+POS uni	0.849	0.067	0.857	0.065	0.836	0.071	0.838	0.071	0.838	0.075	0.838	0.075
	CI+POS bi	0.854	0.027	-	-	-	-	0.854	0.048	**0.872**	0.030	0.863	0.051
	CI+POS tri	0.853	0.035	0.887	0.031	0.889	0.041	0.855	0.044	0.860	0.042	0.852	0.044
2014 Gr 7-9	CI+POS uni	0.856	0.071	0.884	0.063	0.884	0.061	0.885	0.063	0.886	0.061	0.886	0.064
	CI+POS bi	0.874	0.041	-	-	**0.952**	0.022	-	-	-	-	-	-
	CI+POS tri	0.845	0.044	0.903	0.026	0.899	0.030	0.878	0.033	0.885	0.041	0.850	0.044
2014 Gr 3-6	CI+POS uni	0.792	0.042	0.815	0.046	0.815	0.042	0.818	0.048	0.799	0.051	0.799	0.051
	CI+POS bi	0.881	0.029	**0.917**	0.030	0.903	0.030	0.895	0.036	0.871	0.033	0.888	0.039
	CI+POS tri	0.897	0.030	0.912	0.027	0.849	0.032	0.822	0.046	0.763	0.058	0.742	0.066
2014 Gr 3-9	CI+POS uni	0.749	0.044	0.756	0.040	0.755	0.041	0.755	0.041	0.755	0.041	0.746	0.044
	CI+POS bi	0.744	0.028	0.818	0.026	**0.859**	0.029	0.857	0.030	0.846	0.037	0.824	0.035
	CI+POS tri	0.822	0.026	0.855	0.028	0.824	0.036	0.778	0.033	0.735	0.040	0.692	0.046

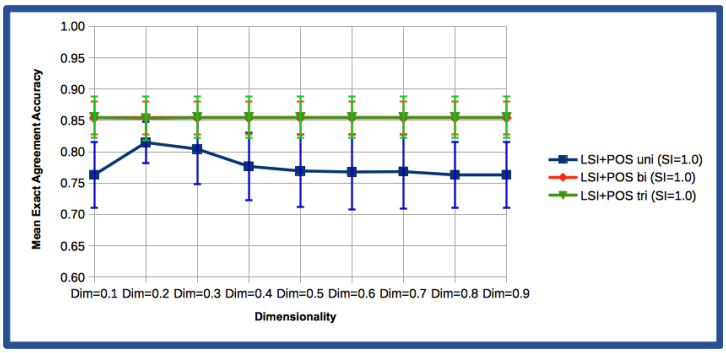
5.2.3.1 2010 Grades 7-9 Dataset

Outputs of the LSI+POS experiments on this dataset are shown in Figure 5.7. On one hand, LSI+POS uni-grams always yielded the lowest MEAA outputs across different SI values. On the other hand, LSI+POS bi-grams exhibited the best performance for SI values from 0.9 down to 0.5. For this dataset, the highest MEAA value achieved by LSI+POS that has significant difference from the LSI outputs of Phase 1 (i.e. p-value=$2.980e^{-08}$) is 0.899 at dim=0.2 using POS bi-grams with SI=0.7.

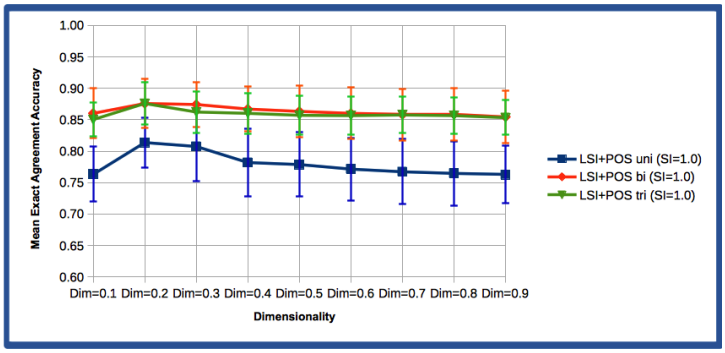
Similarly, the CI+POS outputs are presented in Figure 5.8. It is interesting to note that there are prominent similarities between the outputs and output patterns of LSI+POS and CI+POS. For one, systems using POS uni-grams always yielded the lowest MEAA outputs across different SI values. Additionally, the integration of POS bi-gram features into CI also yielded the highest MEAA for this dataset, reaching a value of 0.872 at sub=1 with a p-value of $5.63e^{-03}$ against the CI outputs of Phase 1.

Moreover, the MEAA values for CI+POS bi-grams and CI+POS tri-grams are very close for this dataset. It can be argued, however, that the former feature set (i.e. CI+POS bi-grams) is more preferable since it has 8 times smaller dimensionality than the other as mentioned in Section 5.2.1. In practical applications, larger dimensionality entails longer delays and higher computational costs. Hence, with almost the same performance, a system with lower dimensionality is more desirable.

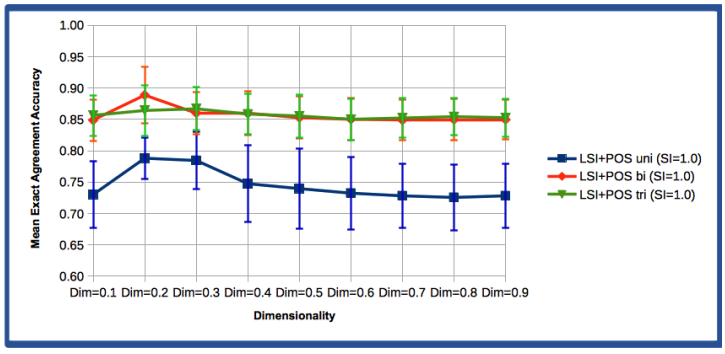
Overall, LSI+POS bi-grams outperformed CI+POS bi-grams by only 0.027 in terms of MEAA and a p-value of 0.02342 was obtained upon evaluation of the statistical significance of the difference in their outputs. It also achieved the overall highest MEAA for this dataset, surpassing the highest value achieved in Phase 2 by 0.044 with a p-value of $4.554e^{-05}$.



(a) $SI = 1.0$

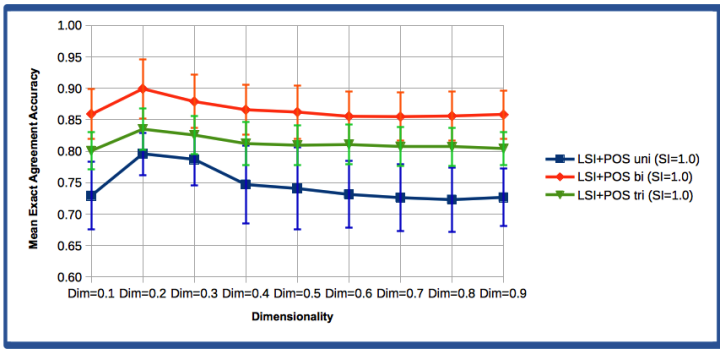


(b) $SI = 0.9$

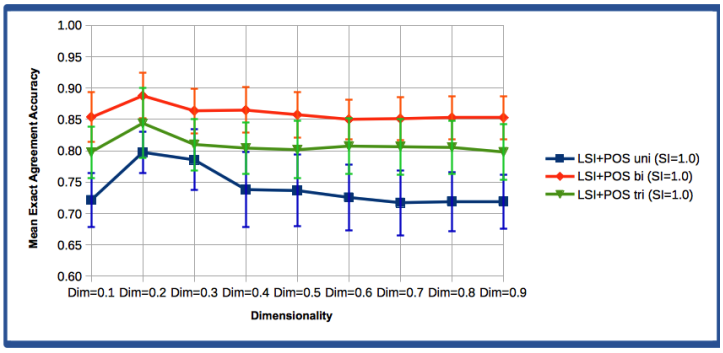


(c) $SI = 0.8$

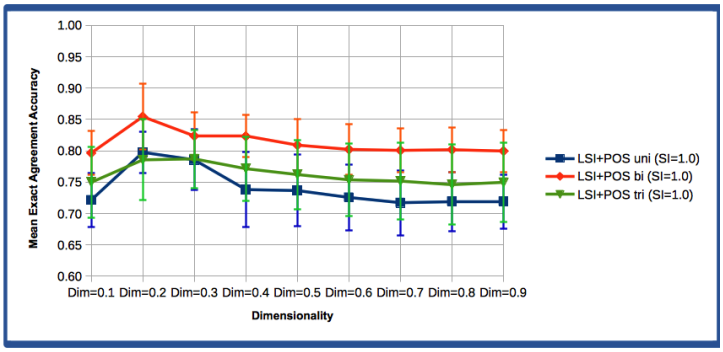
Figure 5.7: LSI+POS with Varying SI Values on the 2010 Grades 7-9 Dataset



(d) $SI = 0.7$

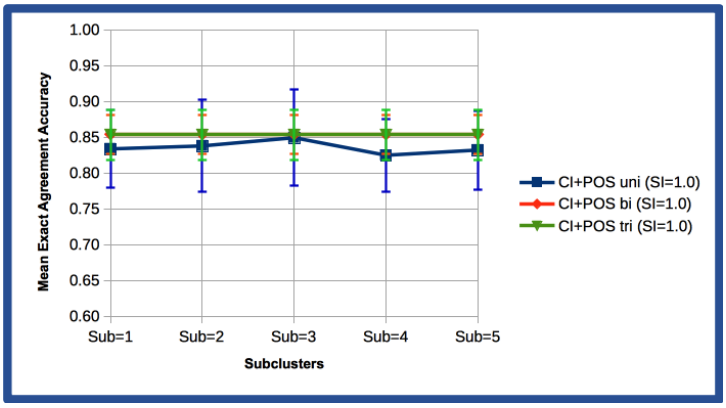


(e) $SI = 0.6$

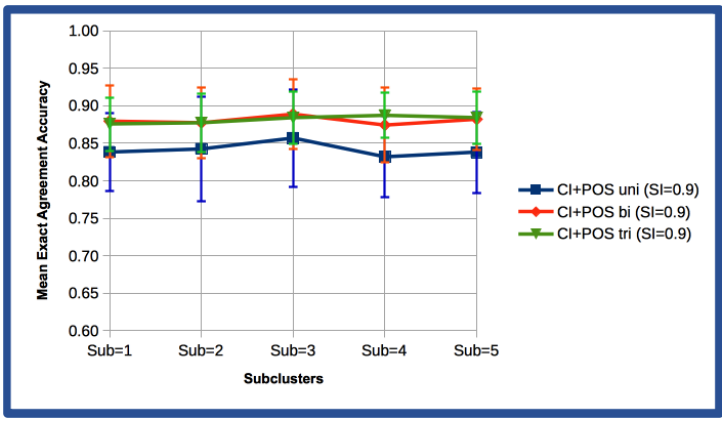


(f) $SI = 0.5$

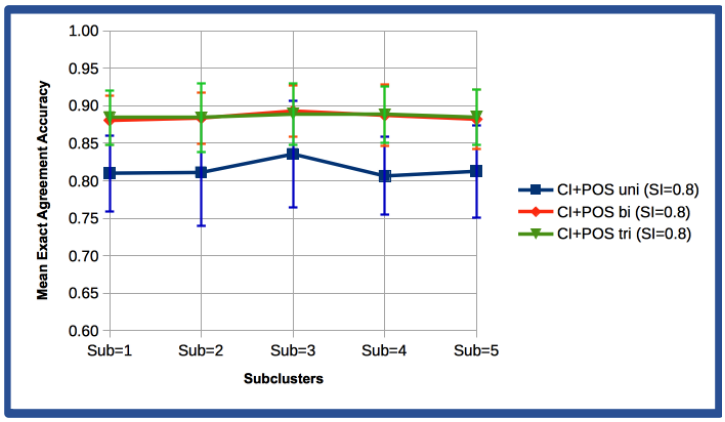
Figure 5.7: *Continuation of* LSI+POS with Varying SI Values on the 2010 Grades 7-9 Dataset



(a) $SI = 1.0$

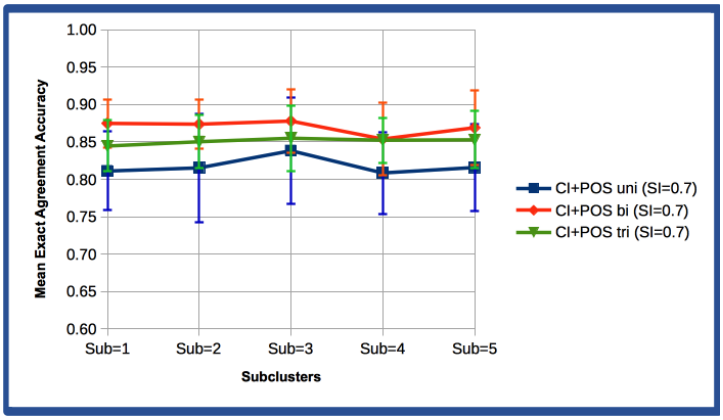


(b) $SI = 0.9$

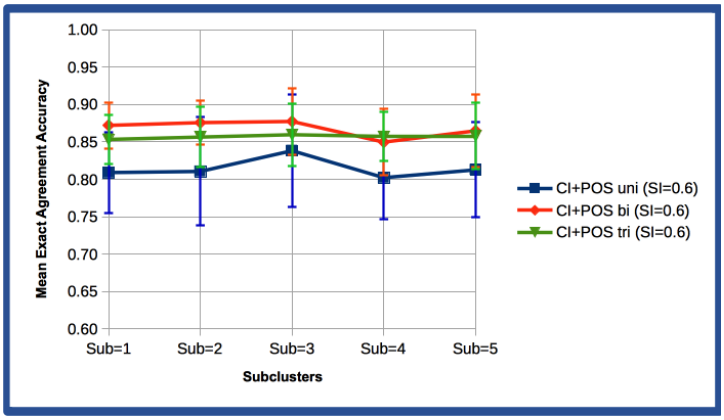


(c) $SI = 0.8$

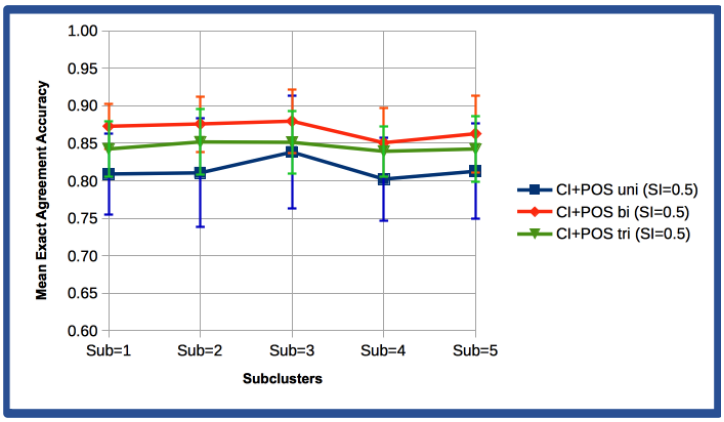
Figure 5.8: CI+POS with Varying SI Values on the 2010 Grades 7-9 Dataset



(d) $SI = 0.7$



(e) $SI = 0.6$



(f) $SI = 0.5$

Figure 5.8: *Continuation of* CI+POS with Varying SI Values on the 2010 Grades 7-9 Dataset

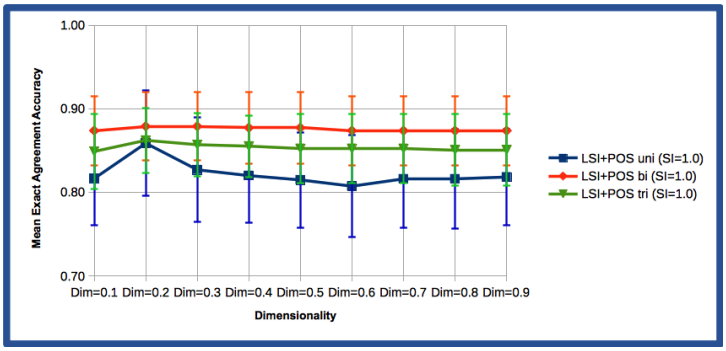
5.2.3.2 2014 Grades 7-9 Dataset

Based on Figures 5.9 and 5.10, the integration of POS uni-grams into LSI and CI yielded the lowest MEAAs for SI values from 0.6 to 1.0. However, systems with POS tri-grams degraded notably at SI=0.5, achieving even lower MEAAs than those systems with POS uni-grams. Additionally, the integration of POS bi-grams mostly yielded the best-performing systems for all SI values as evident in the figures.

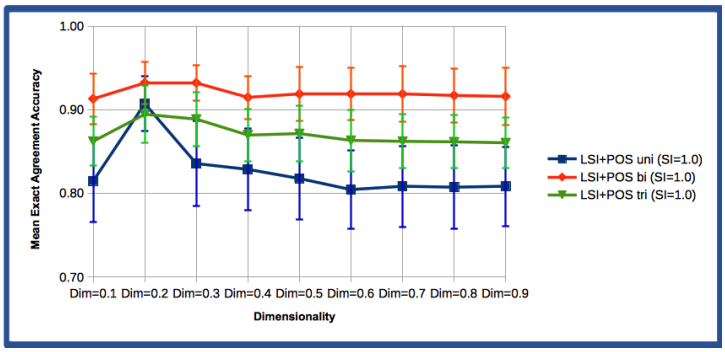
Looking at the line graphs of Figure 5.9, LSI+POS consistently exhibited a notable increase in MEAA at dim=0.2, where it achieved its highest value of 0.934 using POS bi-grams with SI=0.8. However, we found that this output is not statistically significant (i.e. p-value=0.349) with respect to the output of the LSI baseline experiment (i.e. LSI Phase 1 experiment). Upon further investigation, we identified the highest significant MEAA value that LSI+POS achieved in this phase to be 0.907 at dim=0.2 and at dim=0.9 using POS uni-grams with SI=0.9 and POS bi-grams with SI=0.8, respectively. The corresponding p-values for these two points against the LSI outputs in Phase 1 are found to be 0.0151 and 0.0010.

CI+POS, however, achieved its highest MEAA of 0.952 at sub=1 using POS bi-grams with SI=0.8 as can be seen in Figure 5.10c. It significantly surpassed the highest values achieved by LSI+POS uni-grams and LSI+POS bi-grams stated above by 0.045 with corresponding p-values of $4.172e^{-07}$ and $5.96e^{-08}$.

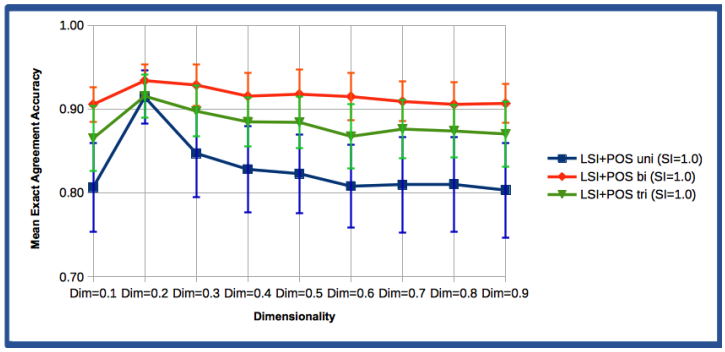
In summary, CI+POS bi-grams yielded the overall highest MEAA for this dataset. Its output also exceeded the highest MEAAs achieved in Phase 1 and Phase 2 by 0.018 and 0.073, respectively, with corresponding p-values of $1.8626e^{-09}$ and $1.49e^{-08}$.



(a) $SI = 1.0$

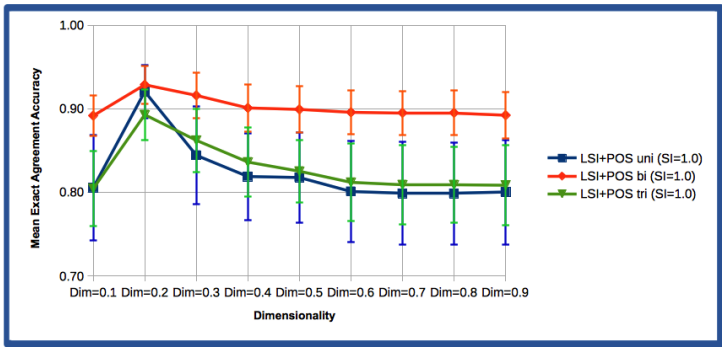


(b) $SI = 0.9$

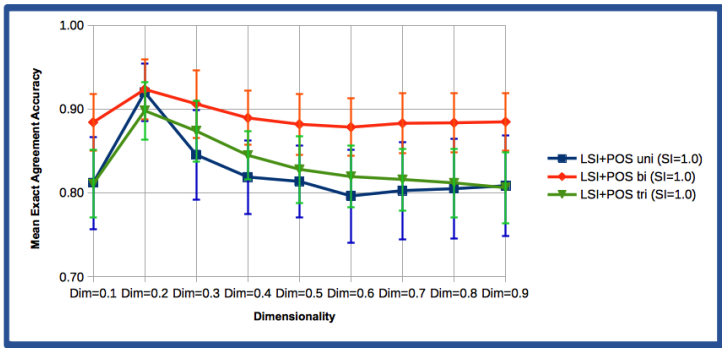


(c) $SI = 0.8$

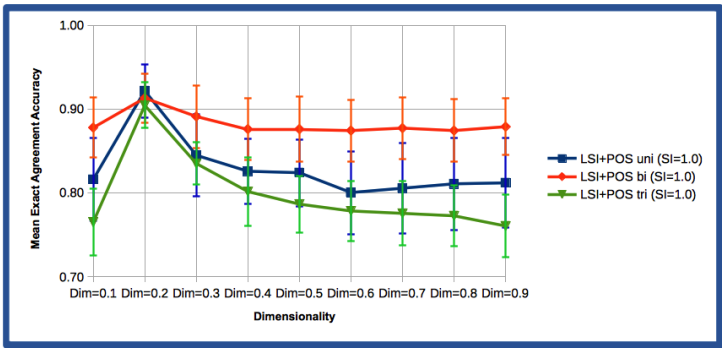
Figure 5.9: LSI+POS with Varying SI Values on the 2014 Grades 7-9 Dataset



(d) $SI = 0.7$

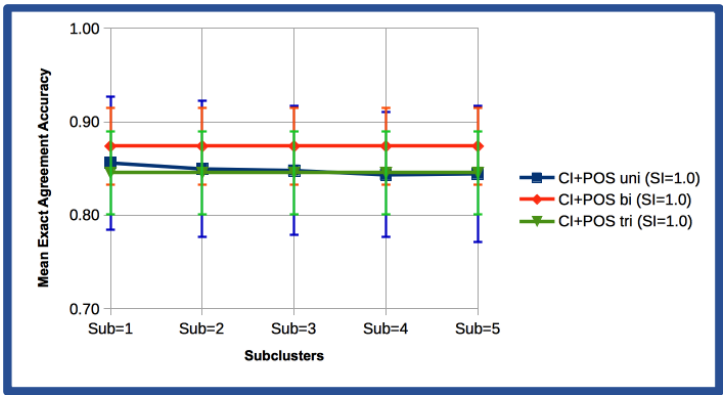


(e) $SI = 0.6$

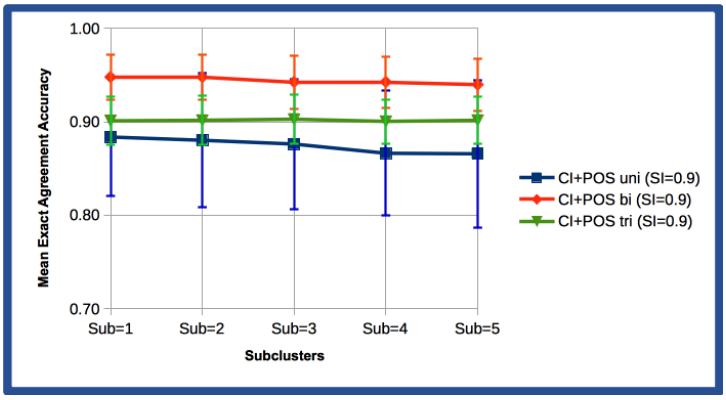


(f) $SI = 0.5$

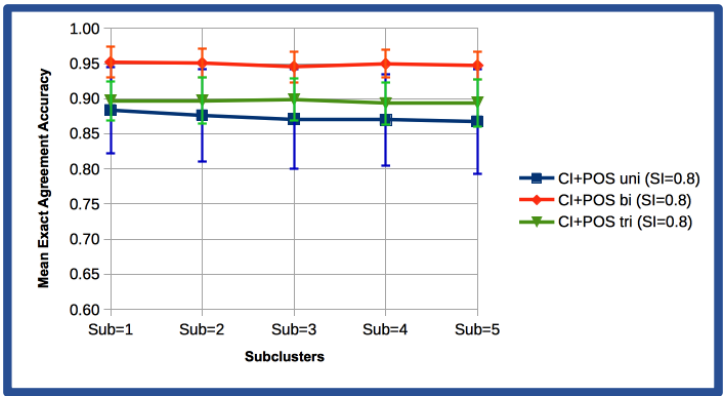
Figure 5.9: *Continuation of* LSI+POS with Varying SI Values on the 2014 Grades 7-9 Dataset



(a) $SI = 1.0$

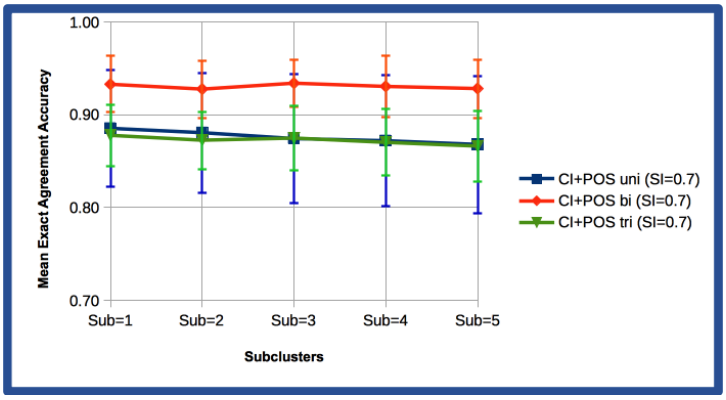


(b) $SI = 0.9$

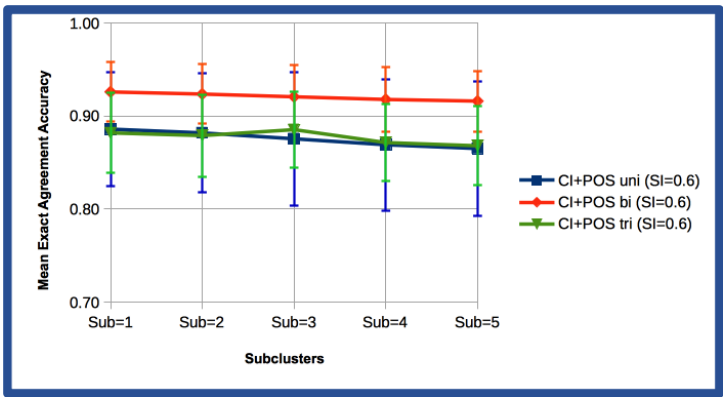


(c) $SI = 0.8$

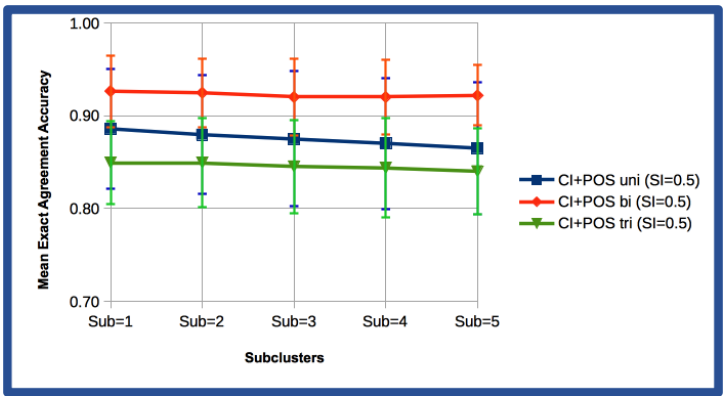
Figure 5.10: CI+POS with Varying SI Values on the 2014 Grades 7-9 Dataset



(d) $SI = 0.7$



(e) $SI = 0.6$



(f) $SI = 0.5$

Figure 5.10: *Continuation of* CI+POS with Varying SI Values on the 2014 Grades 7-9 Dataset

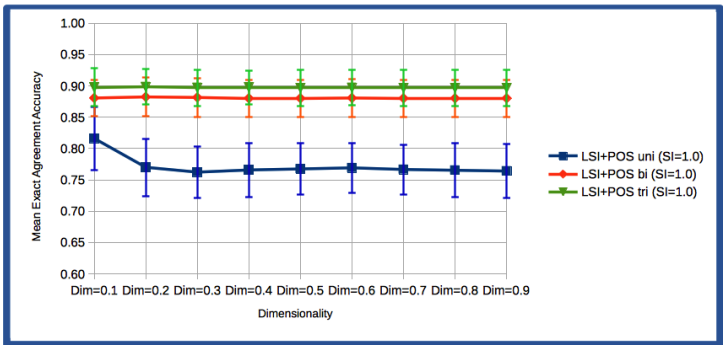
5.2.3.3 2014 Grades 3-6 Dataset

As depicted in Figures 5.11 and 5.12, systems with integrated POS uni-grams yielded the lowest MEAAs on this dataset for SI values ranging from 0.8 to 1.0, regardless of LSI's *dim* and CI's *sub* parameters. However, as SI approaches 0.5, the performance of the systems with POS tri-grams degrades, achieving the lowest MEAAs for all values of LSI's *dim* and CI's *sub* parameters.

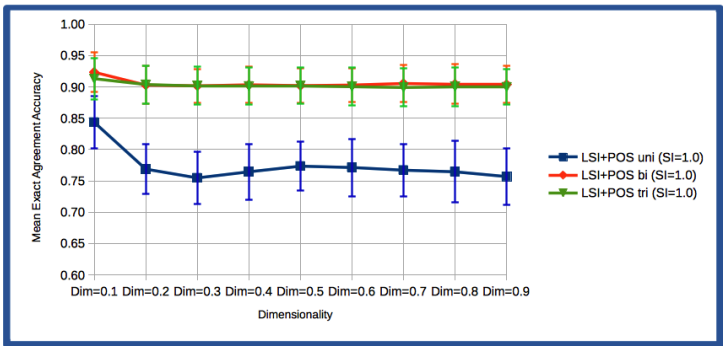
On one hand, LSI+POS-based systems generally achieved their highest MEAA values at dim=0.1 as evident in the line graphs of Figure 5.11. Specifically, the highest MEAA it was able to achieve is 0.924 at this *dim* value using POS bi-grams with SI=0.9. This output has p-values of $1.8626e^{-09}$ and $1.55e^{-05}$ against the highest values achieved in the LSI experiments of Phase 1 and Phase 2, respectively.

On the other hand, CI+POS, achieved its highest MEAA of 0.917 at sub=3 using POS bi-grams with SI=0.9. Evaluation of the statistical significance of this value against CI's highest output achieved in Phase 1 yielded a p-value of $1.8626e^{-09}$, while its statistical significance against CI's highest output in Phase 2 yielded a p-value of 0.003967.

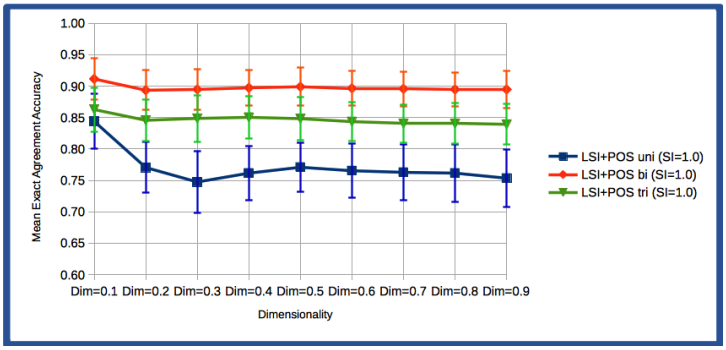
For this dataset, it is inconclusive as to whether LSI+POS bi-grams outperformed CI+POS bi-grams since the p-value between their highest achieved outputs is 0.2035. Additionally, the difference between their MEAA values is only 0.007 which only accounts for less than 1 correctly classified essay in favour of LSI+POS bi-grams. However, we can say that the integration of POS bi-grams features, along with the SS, has been the most beneficial for both algorithms, LSI and CI. This observation is consistent with the results in the previous datasets.



(a) $SI = 1.0$

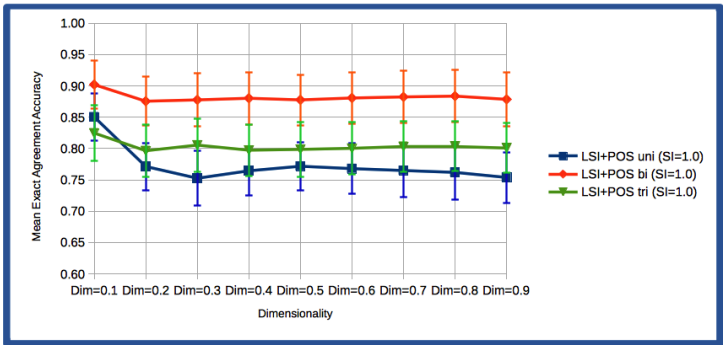


(b) $SI = 0.9$

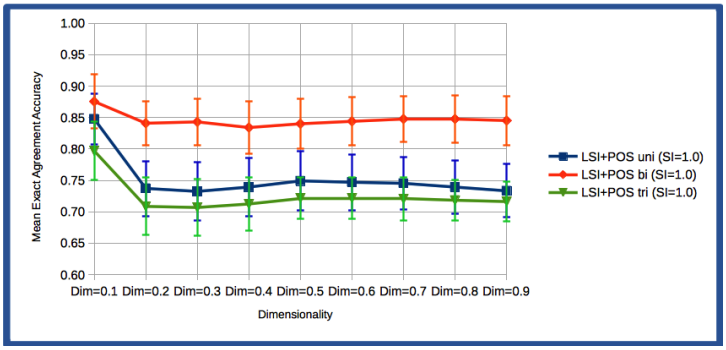


(c) $SI = 0.8$

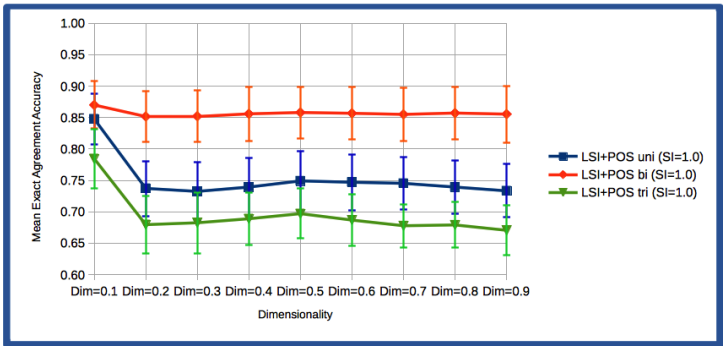
Figure 5.11: LSI+POS with Varying SI Values on the 2014 Grades 3-6 Dataset



(d) $SI = 0.7$

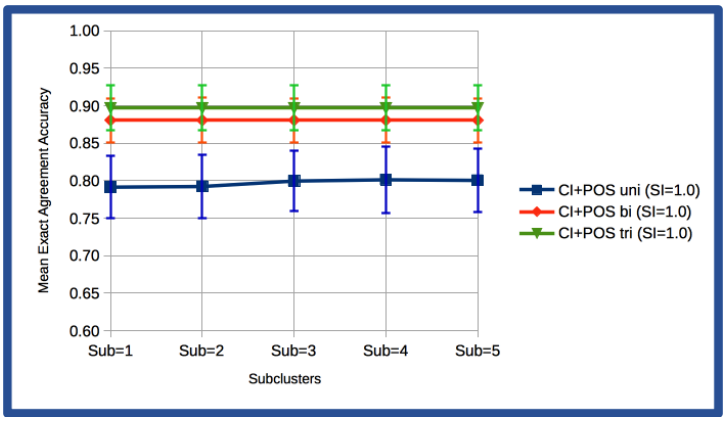


(e) $SI = 0.6$

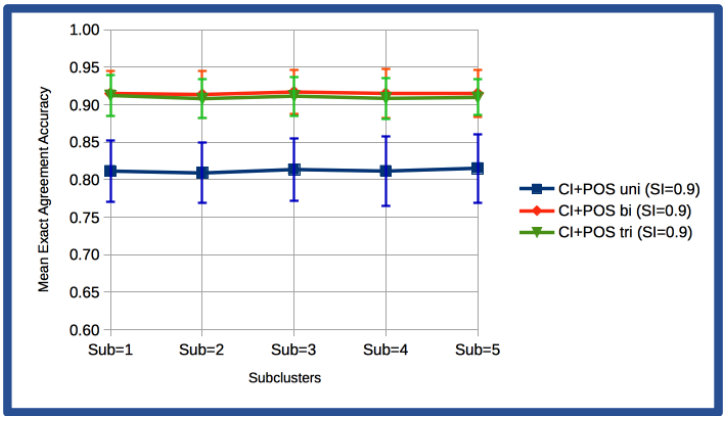


(f) $SI = 0.5$

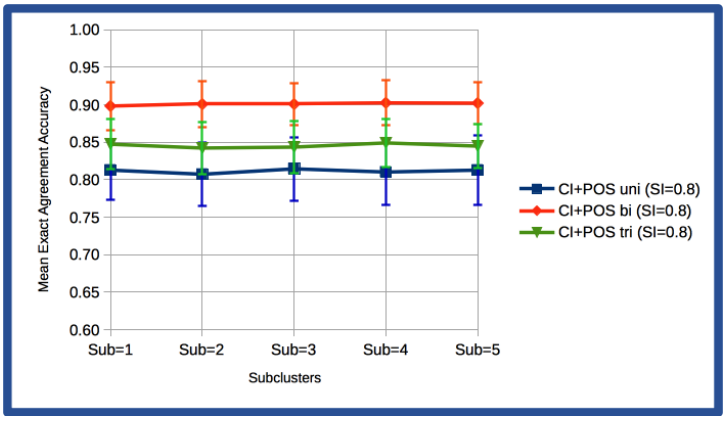
Figure 5.11: *Continuation of* LSI+POS with Varying SI Values on the 2014 Grades 3-6 Dataset



(a) $SI = 1.0$

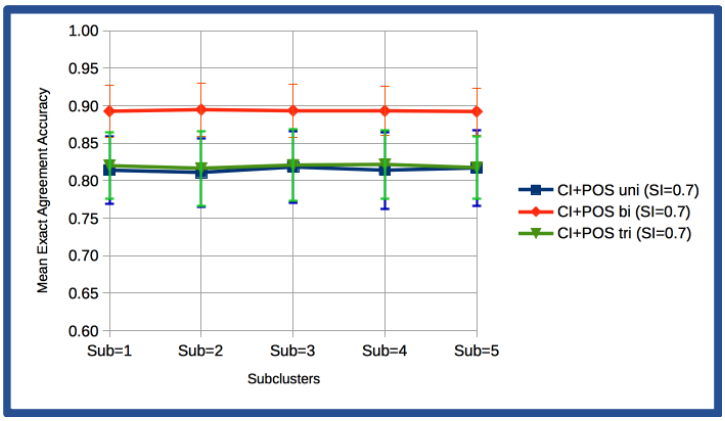


(b) $SI = 0.9$

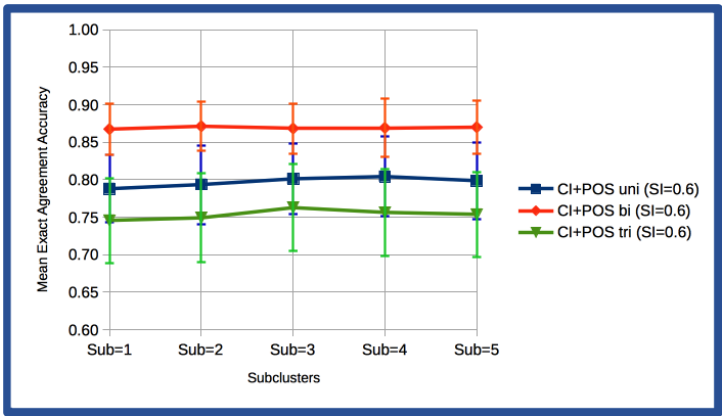


(c) $SI = 0.8$

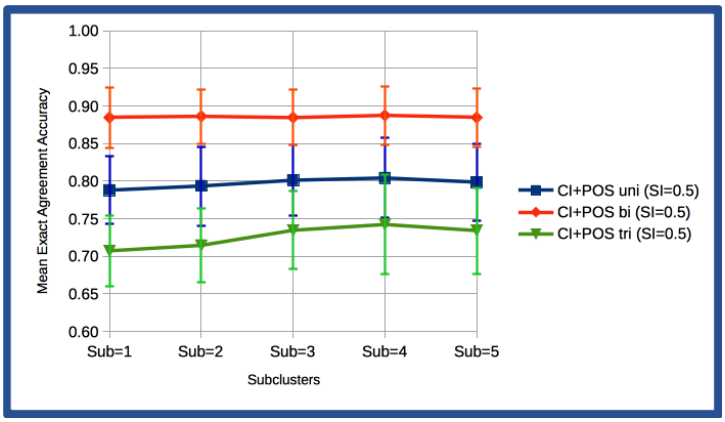
Figure 5.12: CI+POS with Varying SI Values on the 2014 Grades 3-6 Dataset



(d) $SI = 0.7$



(e) $SI = 0.6$



(f) $SI = 0.5$

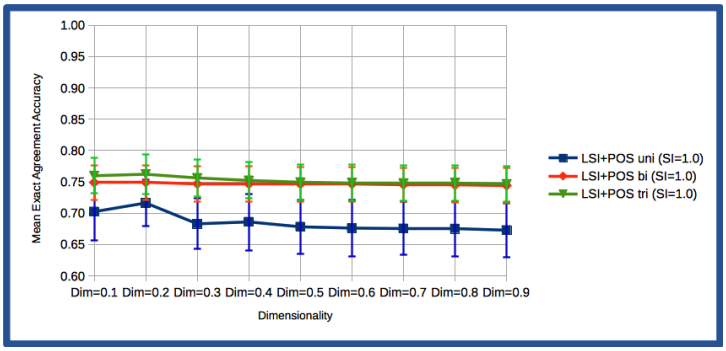
Figure 5.12: *Continuation of* CI+POS with Varying SI Values on the 2014 Grades 3-6 Dataset

5.2.3.4 2014 Grades 3-9 Dataset

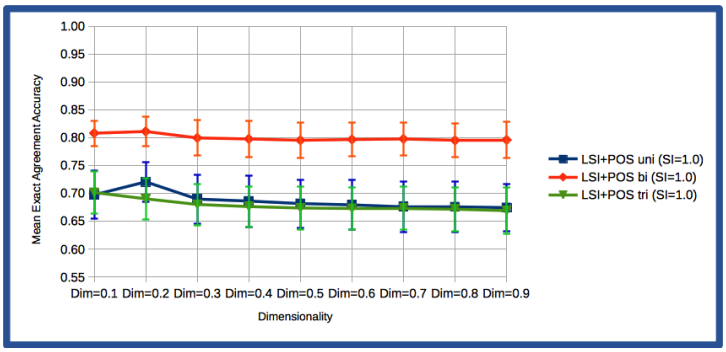
For this dataset, the integration of POS bi-grams with SI=0.8 into the LSI- and CI-based systems yielded the best performance. LSI+POS bi-grams reached its peak MEAA of 0.864 at dim=0.2, while CI+POS bi-grams achieved its highest MEAA of 0.859 at sub=3. It can also be observed that, although systems with POS tri-grams started out with the highest MEAAs at SI=1.0 for both LSI and CI, their MEAAs dropped as the SI value reaches 0.5 as shown in Figures 5.13 and 5.14. In these figures, we can also notice that the graphs for the POS uni-grams are almost the same across different values of SI.

Looking back at the outputs of the previous phases of the experiment, we found that the MEAA achieved by LSI+POS bi-grams in this phase is 0.252 and 0.102 higher than LSI's highest outputs in Phase 1 and 2, respectively, with both p-values=$1.8626e^{-09}$. Similarly, the highest MEAA achieved by CI+POS bi-grams in this phase is found to be 0.168 higher than what CI alone achieved in Phase 1 and 0.037 higher than what CI+POS tri-grams achieved in Phase 2. Corresponding p-values of these comparisons are found to be $1.8626e^{-09}$ and $2.459e^{-06}$.

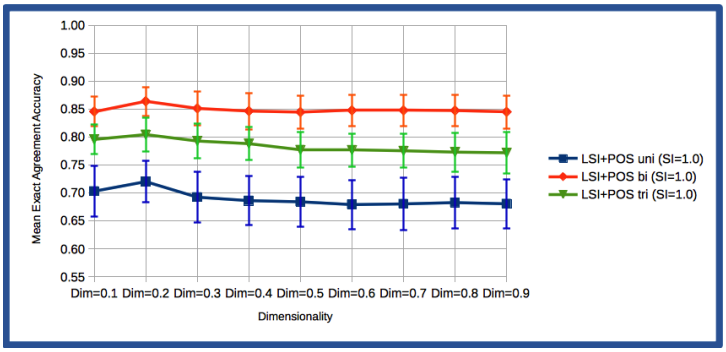
As in the 2014 Grades 3-6 dataset, it is also inconclusive as to whether LSI+POS bi-grams outperformed CI+POS bi-grams for this dataset since the p-value between their highest outputs is 0.2186. Moreover, the difference between their MEAA values is only 0.005 which also only accounts for less than 1 correctly classified essay in favour of LSI+POS bi-grams. What we can claim, however, is that in this phase the integration of POS bi-grams features, along with the implementation of the SS, has significantly enhanced the performance of LSI and CI, allowing them to yield higher MEAA values than what they achieved in Phase 1 and 2.



(a) $SI = 1.0$

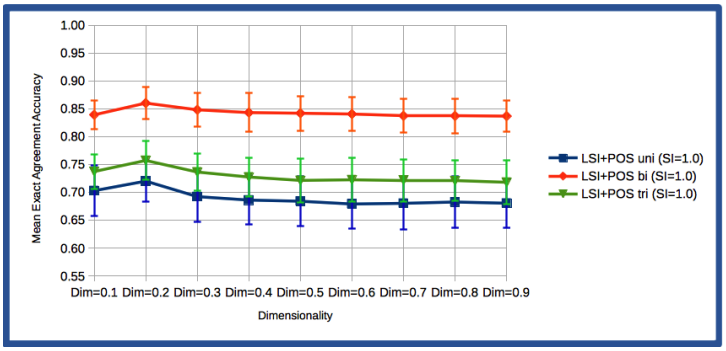


(b) $SI = 0.9$

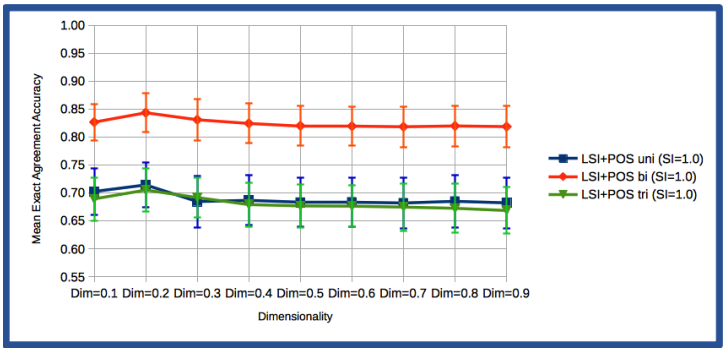


(c) $SI = 0.8$

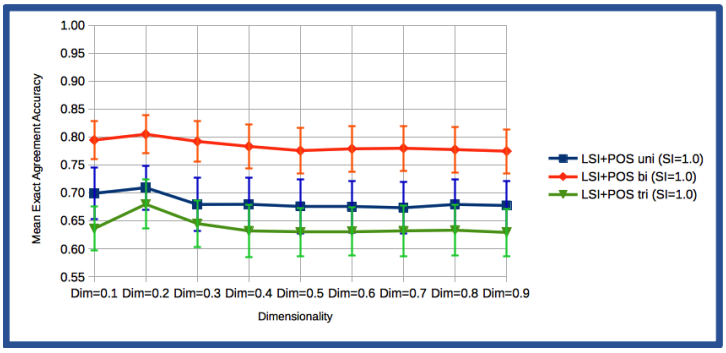
Figure 5.13: LSI+POS with Varying SI Values on the 2014 Grades 3-9 Dataset



(d) $SI = 0.7$

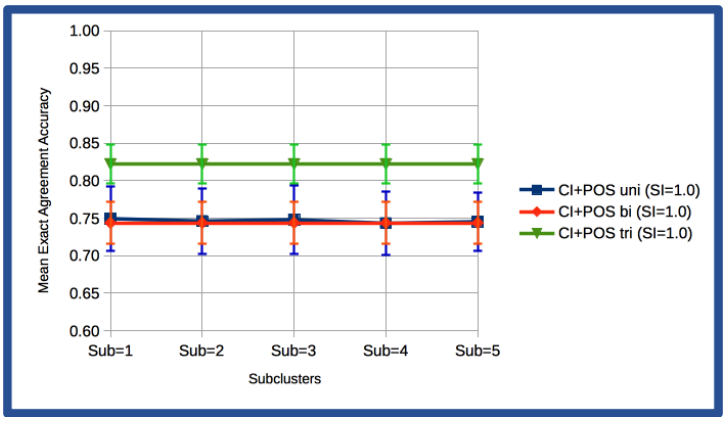


(e) $SI = 0.6$

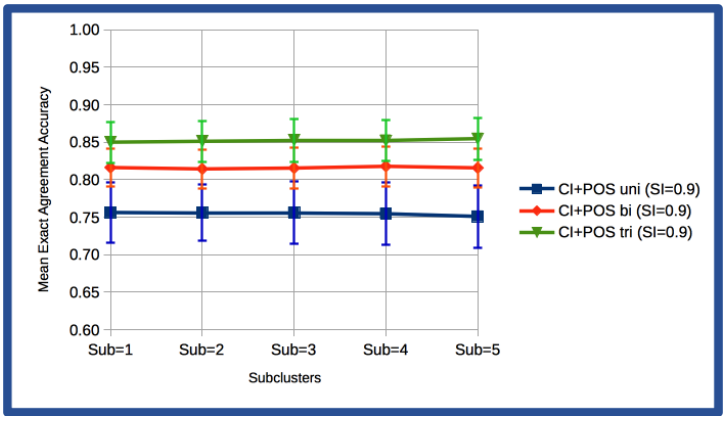


(f) $SI = 0.5$

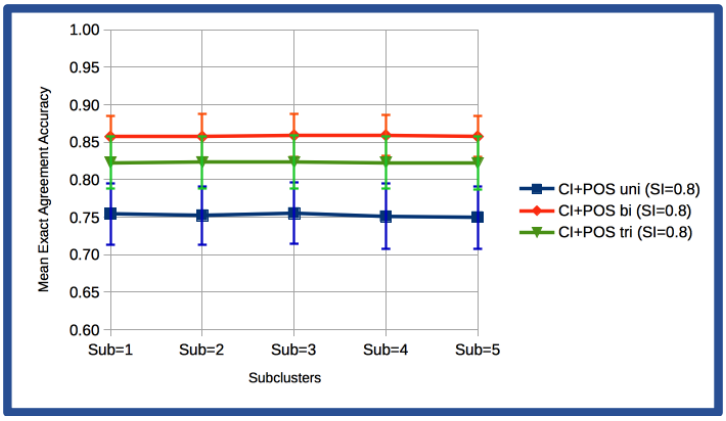
Figure 5.13: *Continuation of* LSI+POS with Varying SI Values on the 2014 Grades 3-9 Dataset



(a) $SI = 1.0$

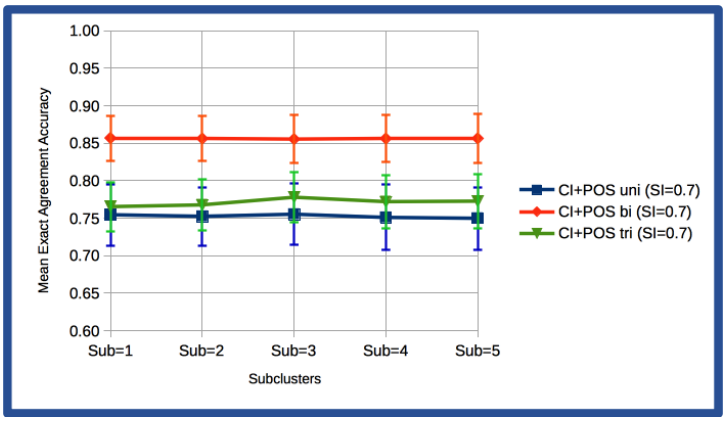


(b) $SI = 0.9$

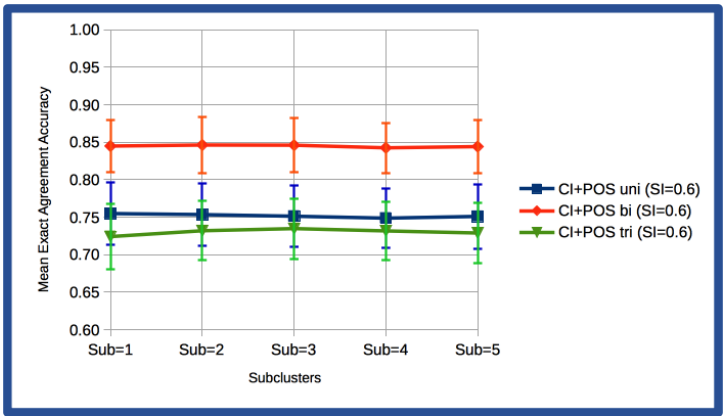


(c) $SI = 0.8$

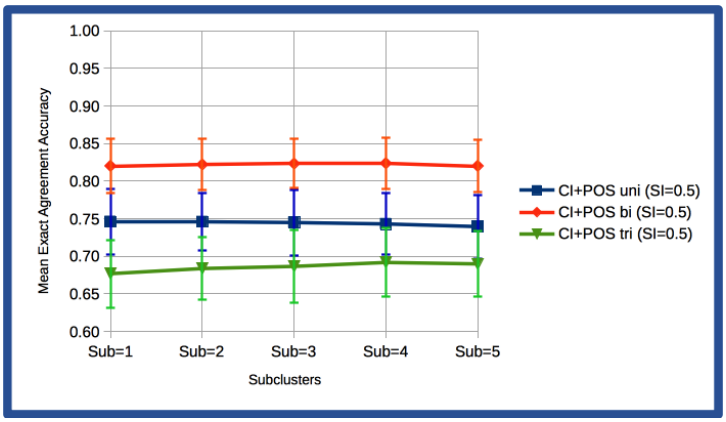
Figure 5.14: CI+POS with Varying SI Values on the 2014 Grades 3-9 Dataset



(d) $SI = 0.7$



(e) $SI = 0.6$



(f) $SI = 0.5$

Figure 5.14: *Continuation of* CI+POS with Varying SI Values on the 2014 Grades 3-9 Dataset

5.2.4 Phase 4: Error Analysis

Errors occur when the predicted grade level of the system differs from the actual grade level of the document. Adjacent grade levels often share common features which in effect makes them harder to set apart. However, adjacency errors can be considered as non-critical because reading materials for students in adjacent grade levels tend to have no sizeable difference. In fact, authors of some text classification studies even consider the AAA as one of the metrics to measure the performance of their systems. Thus, in this phase, our focus would be more on those errors resulting from more than 1 difference between actual and predicted grade levels.

One possible cause for these errors is the POS tagger. As indicated in the study of Horsmann et al. (2015), the openNLP POS tagger only achieved 92.8% accuracy on written documents composed of news articles, travel reports and how-to guides taken from the British National Corpus[1], the Brown Corpus[2] and the Georgetown University Multilayer (GUM) Corpus[3]. However, the more than satisfactory results that we obtained when using POS-based features indicate that any potential errors generated by the POS tagger have minimal impact on the overall performance of the system. The models demonstrated robustness to such errors in these features, most likely due to the consistent manner in which the automatic POS tagger generated them.

Reflecting on the probable effects of these errors on the practical applications of the algorithm, the errors are classified into two categories, an *overestimation error* (O-type) and an *underestimation error* (U-type).

On one hand, O-type errors occur when the following two conditions are met:

[1] `http://www.natcorp.ox.ac.uk/`
[2] `http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/`
[3] `https://corpling.uis.georgetown.edu/gum/`

1.) when the difference between the predicted and actual grade level is more than 1, AND 2.) when the actual grade level is lower than the predicted grade level (i.e. overestimated grade level). In a real life scenario, when a grade 7 essay is wrongly predicted as two or more levels higher, there is a high chance that the student who wrote this essay may not be able to understand the reading materials which will be assigned to him/her and this could consequently bring a negative learning experience for this student. As mentioned in DuBay (2004): "When texts exceed the reading ability of readers, they usually stop reading." Thus, minimisation of this type of error should be prioritised.

On the other hand, U-type errors occur when the following conditions hold true: 1.) when the difference between the predicted and actual grade level is more than 1, AND 2.) when the actual grade level is higher than the predicted grade level (i.e. underestimated grade level). This error type can be considered as less critical than O-type errors since erroneously assigning reading materials with lower readability level to students with higher reading ability does not result in a high-impact negative learning behaviour. They may not find it as challenging or stimulating, but they will surely understand what they are reading which could even promote a positive reading experience.

Summary of these O- and U-type errors for the 2010 Grades 7-9, 2014 Grades 7-9, and 2014 Grades 3-6 datasets can be found in Tables 5.5–5.7, respectively.

In the succeeding subsection, we will present the case analyses for O-type errors such as Essay3 of the 2010 Grades 7-9 dataset, Essay14 of the 2014 Grades 7-9 dataset, and Essay1 of the 2014 Grades 3-6 dataset, and U-type errors such as Essay167 of the 2010 Grades 7-9 dataset, Essay148 of the 2014 Grades 7-9 dataset, and Essay163 of the 2014 Grades 3-6 dataset. The goal of these analyses is to provide insights on the probable reasons for the two aforementioned types of grade level misclassification.

Table 5.5: Sample Errors for Three Random Sets of the 2010 Grades 7-9 Dataset

Set	Error Essay	Predicted Level	Actual Level	Error Type
1	3	9	7	O
2	3	9	7	O
3	3	9	7	O
3	167	7	9	U
4	167	7	9	U
5	167	7	9	U
6	167	7	9	U
7	167	7	9	U
8	167	7	9	U
9	167	7	9	U

Table 5.6: Sample Errors for Five Random Sets of the 2014 Grades 7-9 Dataset

Set	Error Essay	Predicted Level	Actual Level	Error Type
1	14	9	7	O
1	130	7	9	U
1	131	7	9	U
1	146	7	9	U
1	148	7	9	U
2	14	9	7	O
2	160	7	9	U
3	14	9	7	O
4	14	9	7	O
4	131	7	9	U
4	146	7	9	U
4	148	7	9	U
5	14	9	7	O
5	160	7	9	U
6	14	9	7	O
6	130	7	9	U
6	131	7	9	U
6	146	7	9	U
6	148	7	9	U
7	35	9	7	O
7	131	7	9	U
7	146	7	9	U
7	148	7	9	U
8	35	9	7	O
8	131	7	9	U
8	148	7	9	U
9	35	9	7	O
9	131	7	9	U
9	148	7	9	U
10	131	7	9	U
10	146	7	9	U
10	148	7	9	U
11	63	9	7	O
11	131	7	9	U
11	146	7	9	U
11	148	7	9	U
12	63	9	7	O
12	131	7	9	U
12	146	7	9	U
12	148	7	9	U

Table 5.7: Sample Errors for Four Random Sets of the 2014 Grades 3-6 Dataset

Set	Error Essay	Predicted Level	Actual Level	Error Type
1	1	5	3	O
1	142	3	5	U
1	163	3	5	U
1	185	3	5	U
1	195	4	6	U
2	1	5	3	O
3	1	5	3	O
3	12	5	3	O
3	163	3	5	U
4	1	5	3	O
4	21	5	3	O
4	24	5	3	O
4	59	6	4	O
5	40	6	4	O
5	59	6	4	O
5	163	3	5	U
5	185	3	5	U
6	59	6	4	O
6	163	3	5	U
7	163	3	5	U
7	185	3	5	U

5.2.4.1 O-type Error Investigation

In Tables 5.5–5.7, Essay3 of the 2010 Grades 7-9 dataset, Essay14 of the 2014 Grades 7-9 dataset, and Essay1 of the 2014 Grades 3-6 dataset have been consistently tagged as O-type errors by their corresponding best-performing systems using CI+POS bigrams. Essay3 and Essay14 are essays written by grade 7 students which are wrongly classified as essays belonging to the grade 9 level, while Essay1 is an essay written by a grade 3 student which is wrongly tagged as a grade 5 essay by the system.

Quoted below are the entirety of Essay3, Essay14 and Essay1:

> **Have faith in your dreams and someday it will come true!**
>
> *What is faith? I think that faith is a simple word yet hard to explain and define. For me, faith means something that I believe in...it is something that you should have in times of need.*
>
> *Filipino people are great followers of god. It is evident in our way of living. We show faith by praying in everything we do. We pray when we wake up, thanking that supreme being that we come to see a new day! We pray before we sleep to thank him that we are still alive. Truly, praying is a...uh, staple in our daily lives!*
>
> *But, faith isn't only for a supreme being...It is also for a certain person that you believe in. My example is ex President Cory Aquino...she believed that being the first woman president wouldn't stop her to be a great leader to our nation. She had faith in herself that she can free her people from the dictatorial hands of Marcos, and apparently she did. Thanks to the people who had supported her and had faith in her. If not for them, she wouldn't have enough strength to do what she did.*
>
> *We, Filipinos, are known for being hospitable, and also...very determined people. If we want it, we get it! Thats how it rolls. If we have strong faith in something, we stick to believing it. Okay, you may say we may be stubborn...but is it determination is the key? We have faith in something and we are determined that what we believe in is true.*
>
> *Here's the thing...when you want to achieve something, all you need is strong will and good faith, and surely you will reach that thing that you aimed for!*
>
> *(Grade7 Essay3 of the 2010 Grades 7-9 Dataset)*

The surroundings of UPIS is in a state where it is both clean and dirty. In my opinion, I don't or am not particularly fond of this state of cleanliness because it shows that people aren't persistent when it comes to cleanliness and shows weakness of integrity of creation. The people in general share a mind-set in where its okay to destroy the environment where they think that the environment is below them and they can do anything with it, but there are some people with good intentions who try to prevent the destruction from happening.

UPIS tries their best to implement cleanliness campaigns and health programs. For cleanliness campaigns, they usually just remind the students about segregation and bringing their own containers to prevent the overuse of plastic. Another program or campaign is the conservation of energy so they can lessen the costs of electricity. For health programs, there are health appraisals, supervision of the environment, preventive measures, and reasonable pricing for the foods. UPIS is basically doing great in health programs but not much in cleanliness.

In my opinion, before you do cleanliness campaigns and health programs, make sure that the people you are trying to implement it on is disciplined enough to be able to accept the responsibility. The campaigns or programs wont work if discipline, obedience, and persistancy isn't observed. All have to say is discipline is key to success.

(Grade7 Essay14 of the 2014 Grades 7-9 Dataset)

My name is [redacted]*. I am ten years old. My family and I are super close. We like biking, swimming (except for my mom) and going on outings. We sometimes fight and tell jokes. My father and I always go on an adventures but now it is dangerous for him because he is injured. The happiest time / moment is when swim in the La Mesa Eco Park, April 13. We swim, tell jokes and we eat a lot-lot more. And we also sing because it is my mom's birthday. We also practice swimming, and diving. My sister can dive into 7 - 8 feet deep, while my baby brother* [redacted] *and I can only swim 3 - 4 feet only. But I am glad that we could spend more time together.*

(Grade3 Essay1 of the 2014 Grades 3-6 Dataset)

To understand the reasons behind the misclassifications of these essays, a two-part investigation was conducted on each of them corresponding to the two components of the system, namely, content and grammar. For each essay, we gathered the word and POS bi-gram tokens present and we identified which of these tokens are more prevalent in the predicted and actual grade level text samples.

For Essay3 of the 2010 Grades 7-9 dataset, we found that there are 141 word tokens and 121 POS bi-gram tokens. Among the 141 word tokens, 104 are found to be more prevalent in grade 9 text samples. Moreover, 119 out of 121 POS bi-gram tokens are also found to be more prevalent in this grade level (i.e. grade 9). With these, we can therefore speculate that the reason for the misclassification of this essay is that its features (i.e. both the grammar- and content-based features) are more likely to occur in the grade 9 samples than in its actual grade level classification (i.e. grade 7).

These more prevalent word tokens are listed in Table 5.8, while Table 5.9 shows all the POS bi-gram tokens present in Essay3. The ***bold italicised*** tokens in this table are the only ones more prevalent in the grade 7 class (i.e. only 2 tokens) and the rest are more prevalent in grade 9 as discussed above.

Table 5.8: Grade7 Essay3 Word Tokens which are More Prevalent in Grade 9 Essays

a	by	everything	her	living	people	strength	wake
achieve	can	example	here	marcos	person	strong	want
alive	certain	filipino	herself	may	president	that	way
all	come	first	him	me	reach	the	we
also	cory	for	how	means	say	them	what
and	day	free	i	my	see	thing	when
aquino	define	from	if	need	she	think	who
are	determination	get	in	new	should	times	will
be	determined	good	is	not	simple	to	woman
before	did	great	it	of	someday	true	word
being	do	had	known	okay	something	truly	yet
believe	dreams	hard	leader	only	still	up	you
but	enough	have	lives	our	stop	very	your

Table 5.9: Grade7 Essay3 POS Bi-gram Tokens

cc jj	in nns	nn cc	nnp nnp	prp nn	rb prp$	vb wp	vbp nnp
cc nnp	in prp	nn dt	nnp nns	prp nnp	rb rb	vbd in	vbp prp
cc prp	in prp$	nn in	nnp prp	***prp nnps***	rb vb	vbd nns	vbp rb
cc rb	in rb	nn md	nnp vbd	prp prp	rb vbn	vbd prp	vbp to
cc vb	in vbg	nn nn	nnp vbz	prp to	rb vbz	vbd vbn	vbp vbn
cc vbd	in wp	nn nnp	***nnps vbp***	prp vbd	to dt	vbg dt	vbz dt
cc vbz	jj cc	nn nns	nns cc	prp vbp	to prp$	vbg in	vbz in
dt jj	jj in	nn prp	nns in	prp vbz	to vb	vbg jj	vbz jj
dt nn	jj nn	nn rb	nns to	prp$ jj	vb cc	vbg prp	vbz nn
dt nns	jj nnp	nn to	nns vbp	prp$ nn	vb dt	vbn in	vbz nnp
dt prp	jj nns	nn vbg	nns wp	prp$ nns	vb in	vbn nns	vbz prp
dt vbz	jj rb	nn vbz	prp cc	rb in	vb jj	vbn prp	vbz rb
in dt	jj to	nnp cc	prp dt	rb jj	vb nn	vbp in	wp prp
in nn	md rb	nnp in	prp in	rb nn	vb prp	vbp jj	wp vbd
in nnp	md vb	nnp nn	prp md	rb prp	vb rb	vbp nn	wp vbz
							wrb prp

Essay14 of the 2014 Grades 7-9 dataset has a total of 130 word tokens and 98 POS bi-gram tokens. Considering the word tokens for the content component of the system, the results of the investigation are inconclusive since, out of the 130 total number of word terms in Essay14, only 54 of these are more prevalent among grade 9 essays than in the grade 7 class (see Table 5.10), while 66 out of 130 Essay14 terms are more commonly found in the grade 7 class. However, it can be argued that this is to be expected since the topic for the Grade7 essays is different from the Grade9 essays, and essays with the same topic are more likely to share more vocabulary terms or word tokens. As we move on to the grammar component of the investigation, we discovered that 65 of the 98 POS bi-gram terms of Essay14 are more prevalent among grade 9 essays, while only 24 out of these terms (i.e. Table 5.11's entries in ***bold italics***) are more prevalent in the grade 7 class. Therefore, even though Essay14's content features represented by its word terms are more closely related to the grade 7 class, most of its grammar-related features represented by its POS bi-gram tokens are more fitting closely to that of the grade 9 samples. With this, we can speculate that the misclassification happened because Essay14 is grammatically more similar to the essays in the grade 9 level than in the grade 7 level. Since the word and POS bi-gram features are weighted equally, its 65 out of 98 or 66.33% grade 9 POS bi-gram features prevalence rating is significantly higher than its 66 out of 130 or 50.77% grade 7 word features prevalence rating. Thus, it is still misclassified as grade 9.

Table 5.10: Grade7 Essay14 Word Tokens which are More Prevalent in Grade 9 Essays

a	able	about	accept	all	am	another	are	basically
be	because	best	both	but	can	comes	don't	for
from	good	have	i	if	in	intentions	is	isn't
it	just	make	mind	much	my	not	okay	on
opinion	or	own	people	say	share	shows	their	them
they	think	this	usually	when	where	who	with	you

Table 5.11: Grade7 Essay14 POS Bi-grams Tokens

cc jj	***in dt***	jj nns	nn wrb	prp cc	rb to	vbd nn	vbp to
cc nn	in jj	jj to	nnp vbz	prp in	rb vbd	***vbg jj***	vbp vbg
cc nns	***in nn***	jj wrb	***nns cc***	prp md	rb vbg	***vbg prp$***	vbz dt
cc prp	***in nnp***	md vb	nns ex	prp rb	***rb vbn***	vbg to	vbz in
cc rb	in nns	nn cc	***nns in***	prp vbp	to nn	***vbn dt***	vbz jj
cc vbg	in prp	nn dt	***nns nn***	prp vbz	***to vb***	vbn rb	vbz nn
cc vbp	in prp$	***nn in***	nns nnp	prp$ jj	***vb dt***	vbp cc	vbz prp$
cc vbz	in vbg	nn jj	nns prp	prp$ jjs	vb jj	vbp dt	vbz rb
dt jj	in vbz	nn nn	nns to	prp$ nn	***vb nn***	vbp in	vbz to
dt nn	in wrb	***nn nns***	nns vbd	rb in	vb prp	vbp jj	vbz vbn
dt nns	jj cc	nn prp	nns vbp	rb nn	vb vbz	***vbp nn***	wp vbp
dt vbp	jj in	nn vbz	nns wp	***rb rb***	vbd dt	vbp rb	wrb prp
ex vbp	***jj nn***						

For Essay1 of the 2014 Grades 3-6 dataset, there are only 85 word tokens, among which 41 are more prevalent in grade 5 (refer to Table 5.12) and 38 are more prevalent in grade 3. Since there is just a slim difference in the prevalence of these word tokens between the predicted (i.e. grade 5) and actual (i.e. grade 3) grade levels, we still cannot conclude that this could be the reason for its misclassification. However, among the 68 POS bi-gram tokens Essay1 has, 61 are more prevalent in grade 5 (i.e. see Table 5.13 entries in regular font) and only 7 are more prevalent in grade 3 (i.e. see Table 5.13 entries in ***bold italics*** font). With this big difference in the prevalence of POS bi-gram tokens between the predicted and actual grade levels, together with the slim difference in the prevalence of the word tokens discussed earlier, we can speculate that in both the content and grammar components, Essay1 has features which are more likely to occur in grade 5 essays. Thus, it is misclassified as grade 5.

Table 5.12: Grade3 Essay1 Word Tokens which are More Prevalent in Grade 5 Essays

4	an	can	going	is	my	sing	that
8	and	could	he	it	now	sometimes	the
a	are	eat	him	like	on	super	time
also	because	fight	i	lot	only	swimming	when
always	but	for	in	more	practice	tell	while
am							

Table 5.13: Grade3 Essay1 Bi-gram Tokens

cc nn	dt nns	jj nn	nn md	nnp nnp	prp rb	vb cc	vbp dt
cc prp	in cd	***jj prp$***	nn nn	nns cc	prp vbp	vb in	vbp in
cc rb	in dt	jjr nn	nn nnp	***nns jj***	prp vbz	vb jjr	vbp jj
cc vb	in in	***jjs nn***	nn prp	nns prp	prp$ nn	vb nns	vbp vb
cc vbg	in nn	md rb	nn prp$	nns prp$	rb cc	vbg cc	vbz jj
cd nns	in nns	md vb	nn rb	nns rb	rb prp	vbg in	vbz prp$
dt jjs	in prp	nn cc	nn vbg	prp in	rb vb	vbn dt	vbz vbn
dt nn	in prp$	nn in	nn vbz	prp md	rb vbp	***vbp cd***	***vbz wrb***
dt nnp	jj in	nn jj	nnp cc				

Table 5.14 provides the summary of statistics on the prevalence of the content- and grammar-based features for the predicted and actual grade level classifications of the aforementioned O-type essays.

Table 5.14: Statistics on the Prevalence of the Content- and Grammar-based Features in the Predicted and Actual Classes for Grade7 Essay3, Grade7 Essay14 and Grade3 Essay1

CATEGORY DESCRIPTION	CONTENT	GRAMMAR
Essay3's Total No. of Features	141	121
No. of Features More Prevalent in Gr9 (predicted class)	104	119
No. of Features More Prevalent in Gr7 (actual class)	31	2
No. of Features Equally Prevalent in Gr7 and Gr9	6	-
Essay14's Total No. of Features	130	98
No. of Features More Prevalent in Gr9 (predicted class)	54	65
No. of Features More Prevalent in Gr7 (actual class)	66	24
No. of Features Equally Prevalent in Gr7 and Gr9	10	9
Essay1's Total No. of Features	85	68
No. of Features More Prevalent in Gr5 (predicted class)	41	61
No. of Features More Prevalent in Gr3 (actual class)	38	7
No. of Features Equally Prevalent in Gr3 and Gr5	6	-

5.2.4.2 U-type Error Investigation

In this section, we will analyse 3 U-type error examples, namely, Essay167 of the 2010 Grades 7-9 dataset, Essay148 of the 2014 Grades 7-9 dataset, and Essay163 of the 2014 Grades 3-5 dataset. These essays have been consistently classified as two grade levels lower than their actual grade level by their corresponding best-performing systems using CI+POS bi-grams. Essay167 and Essay148 are essays written by grade 9 students which are classified as grade 7 samples by the system. Similarly, Essay163 is written by a grade 5 student which is tagged by the system as a grade 3 essay. Below are the actual essays written by the students.

A Hero of Mine

A Hero must be influential and has a great mind that can help others with their problems. A hero doesn't need to be popular or some legends. Usually heroes are just typical everyday workers. These heroes you can rely on even if you don't know them.

For me my Hero of mine is my mother because since birth when my daddy is at work my mother is around to feed me, care for me, and even change my diaper. And when at night my mother is very restless because I'm crying for a bottle of milk or she doesn't know why I'm crying. Until now my mother usually do my chores. She washes my clothes, iron my clothes, cook for me in breakfast or snacks. For me she's a real hero because without her I can't do my school work or I can't enjoy my vacation or weekends. Usually or sometimes I help her but she rejects because my mother said that I am just a problem and it wastes time. I really appreciate what my mother does for me. And that is what a true hero for me.

(Grade9 Essay167 of the 2010 Grades 7-9 dataset)

SELFIE

Selfie can be also a cause of bullying because some have ugly photos and all of the people seeing it so they tease people, the person who bully he/she lover self esteam, so photos like this are important so that you will know who is your true friend.

(Grade9 Essay148 of the 2014 Grades 7-9 dataset)

MY ACHIEVEMENTS

Last quarter, our PE topic was table tennis. Now, our topic is track and field. From 1st quarter until our 2nd quarter, I felt that our topics is not for me. Last year, in grade 4, I was accepted at a volleyball team, but when summer started, we stopped until this quarter. Our coach said that we will stop until next year. Now, in grade 5, 3rd quarter to 4th quarter, out topic is track and field. We are already in long jumps then triple jump. Since I am not part of the team, I am going to try-out for track and field varsity. I attained this position by eating a lot of healthy food and trying hard.

Eventhough I am very active in sports, I am also active in my studies. My grades this year are higher than my last year's grades, because I study hard and recite especially in science and math because those two subjects are my worst subjects since grade 3.

I can help people with my skills because I can teach people and if I am varsity I can compete in the Olympics and win for the Philippines! I will study hard and do my best to achieve these dreams! I'll award my self a most athletic and smart student award.

(Grade5 Essay163 of the 2014 Grades 3-6 dataset)

For U-type errors, we conducted a similar investigation as we did in the last section to be able to compare the behaviour of the error types. Our first observation is that, unlike the O-types, the word and POS bi-gram tokens of Essay167, Essay148 and Essay163 (i.e. U-type errors) are found to be more prevalent in their actual grade levels as shown in Table 5.15.

Note that content- and grammar-related features are represented by the word and POS bi-gram tokens, respectively. If the prevalence of these features is the only basis

Table 5.15: Statistics on the Prevalence of the Content- and Grammar-based Features in the Predicted and Actual Classes for Grade9 Essay167, Grade9 Essay148 and Grade5 Essay163

CATEGORY DESCRIPTION	CONTENT	GRAMMAR
Essay167's Total No. of Features	97	88
No. of Features More Prevalent in Gr9 (predicted class)	96	87
No. of Features More Prevalent in Gr7 (actual class)	1	1
No. of Features Equally Prevalent in Gr7 and Gr9	-	-
Essay148's Total No. of Features	41	39
No. of Features More Prevalent in Gr9 (predicted class)	33	23
No. of Features More Prevalent in Gr7 (actual class)	6	14
No. of Features Equally Prevalent in Gr7 and Gr9	2	2
Essay163's Total No. of Features	114	82
No. of Features More Prevalent in Gr5 (predicted class)	111	78
No. of Features More Prevalent in Gr3 (actual class)	1	3
No. of Features Equally Prevalent in Gr3 and Gr5	2	1

for classification, these essays would have been tagged correctly since both feature sets (i.e. content- and grammar-related feature sets) are indeed more prevalent in their actual grade levels.

For Essay167, 96 out of 97 of its word tokens are more prevalent in the grade 9 class (i.e. its actual grade level) than in the grade 7 class (i.e. its predicted grade level). A similar case holds for 33 out of 41 word tokens in Essay163. Moreover, 111 out of 114 of Essay148's word tokens are also found to be more prevalent in its actual grade level (i.e. grade 5) than in its predicted grade level (i.e. grade 3). Tables 5.16–5.18 provide the lists of these more prevalent word tokens of the aforementioned essays in their respective actual grade levels.

Table 5.16: Grade9 Essay167 Word Tokens which are More Prevalent in Grade 9 Essays

am	care	enjoy	in	mother	problems	sometimes	very
and	change	even	influential	must	real	that	washes
appreciate	chores	everyday	iron	my	really	their	wastes
are	clothes	feed	is	need	rejects	them	weekends
around	cook	for	it	night	rely	these	what
at	crying	great	just	now	restless	time	when
be	daddy	has	know	of	said	to	why
because	diaper	help	legends	on	school	true	with
birth	do	her	me	or	she	typical	without
bottle	does	hero	milk	others	since	until	work
breakfast	doesn't	heroes	mind	popular	snacks	usually	workers
but	don't	if	mine	problem	some	vacation	you
can							

Table 5.17: Grade9 Essay148 Word Tokens which are More Prevalent in Grade 9 Essays

all	bullying	friend	is	lover	seeing	tease	ugly
are	can	have	it	people	self	they	who
be	cause	he	know	person	selfie	this	you
because	esteam	important	like	photos	she	true	your
bully							

Table 5.18: Grade5 Essay163 Word Tokens which are More Prevalent in Grade 5 Essays

1st	athletic	eventhough	in	now	self	team	trying
2nd	attained	felt	is	of	since	tennis	two
3rd	award	field	jump	olympics	skills	than	until
4th	because	food	jumps	our	smart	that	varsity
accepted	best	for	last	out	sports	the	very
achieve	but	from	ll	part	started	these	volleyball
achievements	by	going	long	pe	stop	this	was
active	can	grade	lot	people	stopped	those	when
already	coach	grades	math	philippines	student	to	will
also	compete	hard	me	position	studies	topic	win
am	do	healthy	most	quarter	study	topics	with
and	dreams	help	my	recite	subjects	track	worst
are	eating	higher	next	said	table	triple	year
at	especially	if	not	science	teach	try	

In Tables 5.19–5.21, we present the POS bi-grams present in Essays 167, 148 and 163. For Essay167, only 1 out of 88 POS bi-gram tokens is more prevalent in grade 7 (i.e. its predicted class) than in grade 9 (i.e. its actual class). A similar case holds for Essay148 with only 14 out of 39 POS bi-gram tokens more prevalent POS bi-gram tokens in its predicted grade level (i.e. grade 7). Moreover, for Essay163, only 3 out of 82 POS bi-gram tokens are more prevalent in its predicted class (i.e. grade 3) than in its actual class (i.e. grade 5). These POS bi-gram tokens which are more prevalent in their predicted grade level than in their actual grade level are shown in Tables 5.19–5.21 in ***italics bold*** font.

Table 5.19: Grade9 Essay167 POS Bi-gram Tokens

cc dt	in dt	jj nns	nn wrb	nns vbp	prp$ nns	to vb	vbp prp
cc nns	in in	md vb	nnp dt	prp cc	rb cc	vb in	vbp prp$
cc prp	in nn	nn cc	nnp in	prp in	rb dt	vb jj	vbp rb
cc rb	in nnp	nn in	nnp md	prp md	rb in	vb nns	vbp vb
cc vbz	in prp	nn nn	***nnp vbg***	prp prp$	rb jj	vb prp	vbz dt
cc wrb	in prp$	nn prp	nns dt	prp rb	rb nns	vb prp$	vbz in
dt jj	in rb	nn prp$	nns in	prp vb	rb prp	vb to	vbz nn
dt nn	jj cc	nn rb	nns nn	prp vbp	rb prp$	vb wp	vbz prp$
dt nnp	jj in	nn vbd	nns prp	prp vbz	rb to	vb wrb	vbz rb
dt nns	jj jj	nn vbz	nns rb	prp$ nn	rb vb	vbd in	vbz wp
dt vbz	jj nn	nn wdt	nns vb	prp$ nnp	rb vbp	vbg in	wdt md

Table 5.20: Grade9 Essay148 POS Bi-gram Tokens

cc dt	***in dt***	***jj nn***	***nn rb***	***nns dt***	prp rb	rb prp	vbp nns
dt in	in in	jj nns	nn wp	***nns in***	prp vbp	vb rb	vbz prp
dt nn	***in nn***	md vb	nnp md	nns vbg	prp$ jj	vb wp	vbz prp$
dt nns	in prp	***nn in***	***nnp nnp***	prp md	***rb dt***	vbg prp	wp vbz
dt vbp	jj in	nn nn	***nns cc***	prp nn	rb nns	vbp jj	

With these dissimilar results between O-type and U-type errors, we can say that the nature of the cause of these error types is also different from each other. Our

Table 5.21: Grade5 Essay163 POS Bi-gram Tokens

cc in	in nns	nn cd	nn vbp	nns vbz	rb jj	vb jj	vbg rb
cc jj	in prp	nn dt	nn vbz	prp jj	rb nn	vb nns	vbg to
cc nn	in prp$	nn in	nnp nns	prp md	rb prp	vb prp$	vbn in
cc vb	in vbg	nn nn	nns cc	prp vbd	rb prp$	vb rb	vbp jj
cc vbg	jj cc	nn nns	nns dt	prp vbp	rb rb	vbd dt	vbp nn
cd nns	jj in	nn prp	nns in	prp$ jj	to jj	vbd in	vbp prp$
dt nn	jj nn	nn prp$	***nns jj***	prp$ jjs	to nn	vbd nn	vbp rb
dt nns	jj nns	nn rb	nns prp	prp$ nn	to vb	vbd prp	vbp vbg
in dt	jjr in	nn to	nns prp$	prp$ nns	vb dt	vbd vbn	***vbz nn***
in jj	md vb	nn vbd	nns vbp	rb in	vb in	vbg dt	vbz rb
in nn	nn cc						

speculation is that, on one hand, O-type errors occur because of the presence of predicted-grade-level-specific features (i.e. more complex or distinctive features of the higher predicted class) in these essays in the lower grade levels. On the other hand, U-type errors happen because they lack those features which are distinct to their actual grade level class. Moreover, although the features present in these essays are indeed more prevalent in their actual class, these can also found in the their predicted lower grade level class. Therefore, we can speculate that the absence of the actual-grade-level-specific features, together with the sufficiency of the essays' features to belong to a lower predicted class, could result in a U-type error.

5.3 Chapter Summary

This study has four experimental phases. In Phase 1, we conducted the baseline experiments in which we individually used the feature sets, LSI, CI and POS. Then, we integrated the POS-based feature sets separately into LSI and CI to produce the LSI+POS and CI+POS feature sets, respectively. In Phase 2, the SS discussed in Section 4.6 was not implemented yet (i.e. SI=1.0) and we simply added the POS-based features into LSI and CI. To further optimise the combination process, we conducted Phase 3 in which the SS was applied to the datasets with SI values ranging from 0.5 to 0.9. Lastly, in Phase 4, we took a closer look at the wrongly classified documents in Phase 3's highest performing system configuration (i.e. CI+POS bi-grams at SI=0.9) to be able to speculate on the probable causes of these remaining errors.

Table 5.22 summarises the highest MEAA achieved per phase on each dataset. It also presents the feature set(s) which achieved these MEAAs. For comparison purposes, we also applied the readability formulas discussed in Section 2.2 on our datasets and derive their corresponding accuracies as presented in Table 5.23.

In Phase 1, the CI-based systems dominated the 2010 and 2014 Grades 7-9 datasets which are only composed of documents from the secondary school levels, while POS-based systems dominated the datasets which include essays from the primary school levels, i.e. 2014 Grades 3-6 and Grades 3-9 datasets. With these, we can infer that in the lower grade levels grammatical structures, which are approximated by the POS-based feature sets, are more representative. This can be explained by the fact that language learning in lower grade levels is more focused on the fundamentals of correct grammatical structures rather than rich semantics. As the grade level increases, however, content becomes more significant. Thus, the classification gears

towards the CI-based feature sets in the secondary levels.

In Phase 2, LSI+POS and CI+POS almost equally performed well. Note that in the datasets with essays from the primary levels (i.e. 2014 Grades 3-6 and Grades 3-9 Datasets), the integration of POS tri-gram features yielded the highest MEAAs. This further strengthens our claim that grammatical structures are more important in these levels since POS tri-grams are the most indicative of grammar among n-grams. Also, in this phase, overall highest MEAA of the systems decreased in 2 out of the 4 datasets compared to Phase 1's highest outputs, with no MEAA change for the 2014 Grades 3-6 dataset. This degradation in performance can be interpreted as an indication that directly adding features together could result in lower system performance. Therefore, we can also argue that integrating several distinct feature sets together in modelling the grade levels does not guarantee better classification.

In Phase 3 wherein the SS was implemented, POS bi-grams systems dominated on all the datasets. With SI=0.6 to SI=0.9, they consistently yielded the highest MEAAs across the different values of the *dim* and *sub* parameters. As shown in Tables 5.3 and 5.4, the highest MEAA values are never achieved at SI=1.0 (i.e. without the SS). This proves that the SS essentially enhanced the classification process by removing the sparse terms which only contributed noise in the system. Comparing the results of Phase 1 and Phase 3, we found that the MEAA values achieved in the latter (i.e. Phase 3) are significantly higher in 3 out of the 4 datasets for both the LSI- and CI-based systems, with the 1 remaining dataset still reaching higher MEAAs but with questionable statistical significance, i.e., p-values>0.05. In addition to that, all the MEAA values achieved in this phase are significantly higher than those achieved in Phase 2 for all datasets. With these, we can say that among the 3 phases, Phase 3 yielded the best results.

Table 5.22: Summary of Highest MEAA Achieved Per Phase on each Dataset

Experiment	**2010 GR 7-9**	**2014 GR 7-9**	**2014 GR 3-6**	**2014 GR 3-9**
Phase 1	CI **0.90**	CI **0.93**	POS **0.90**	POS **0.83**
Phase 2	LSI+POS tri CI+POS bi **0.85**	LSI+POS bi **0.87**	LSI+POS tri CI+POS tri **0.90**	CI+POS tri **0.82**
Phase 3	LSI+POS bi, SI=0.8 **0.90**	CI+POS bi, SI=0.8 **0.95**	LSI+POS bi, SI=0.9 CI+POS bi, SI=0.9 **0.92**	LSI+POS bi, SI=0.8 CI+POS bi, SI=0.8 **0.86**

Table 5.23: Summary of Accuracies Achieved by Prominent Readability Formulas Discussed in Section 2.2 on each Dataset

Formula	**2010 GR 7-9**	**2014 GR 7-9**	**2014 GR 3-6**	**2014 GR 3-9**
DC (3000 word list)	0.27	0.35	0.06	0.19
FK	0.10	0.19	0.01	0.09
FRE	0.37	0.23	0.09	0.20
FOG	0.18	0.08	0.22	0.16
SMOG	0.10	0.16	0.02	0.08

Lastly, in Phase 4, we analysed two types of errors, namely, *Overestimation* (i.e. O-type) and *Underestimation* (i.e. U-type) errors, in terms of the prevalence of content- and grammar-related features. In our investigation, we found that essays tagged as O-types have more prevalent distinctive features belonging only to the higher grade levels in which they were classified. On the contrary, essays tagged as U-types have more prevalent features belonging to their actual grade level class. However, these features can also be found in their predicted lower grade level class. Therefore, we can say that this investigation led us to a simple asymmetry issue, wherein low-level features are shared among all grade levels, while high-level features can only be found in higher level grade levels. Consequently, we speculate that O-type errors occur because of the extensive presence of these high-level features found in higher grade levels, while U-type errors occur because of the lack of these features.

Chapter 6

Conclusion and Future Work

This chapter concludes our study. Section 6.1 presents the overall summary of our work. In this section, we will revisit the research questions and hypotheses we presented in Chapter 3. Then, in Section 6.2 we will provide probable future directions for other researchers who would like to continue working on this topic.

6.1 Summary of the Study

Reading is a prerequisite in learning and the process of learning how to read varies for each person. In a typical classroom setting, we cannot expect students to have the same motivation, preference, knowledge and attitude towards learning. Thus, there is no "one size fits all" language learning program that we can easily implement for our learners and that is what makes it a challenging field of study.

Technology can play a vital role in language learning. With the advances in the NLP area, specifically in the TRA domain, we can now develop systems which can be powerful tools to promote self-directed language learning and to optimise rigorous processes involved in the selection of appropriate instructional materials for learners.

There have been several studies in the TRA domain, including the use of readability formulas, such as FOG, SMOG and Flesch-Kincaid, and Machine Learning techniques as used by authors like Si and Callan of the Expectation Maximisation-based

system, Schwarm and Ostendorf of the SVM-based system, and Collins-Thompson and Callan of the Multinomial Naive Bayes-based system. These were discussed in Chapter 2 of this thesis.

Research Question 1
How can we create an easily retrainable reading ability estimation system using ML strategies?

In this study, we developed a novel approach to reading ability estimation of English language learners using concepts and strategies in the TRA domain. Actual written essays from students in the primary and secondary levels were used to approximate their reading ability and calibrate our system. In our implementation, as discussed in Chapter 4, we did not use raw text features, such as sentence length and word tokens, which were commonly used in previous research. Instead, we utilised content-based similarity features between the student essays and reference materials. These similarity features were derived from the LSI and the CI algorithms discussed in Section 4.5. One advantage of our proposed system is that it will never expire unlike the formula-based methods. To update the system, we only need to 1.) collect new essays and reference materials, and 2.) retrain the system using these new materials.

Research Hypothesis 1
The combination of content- and grammar-based text features yields better performing systems.

Research Question 2
Which feature set or feature set combinations are most relevant and effective in modelling each school grade level in the datasets?

As discussed in Chapter 5, we investigated several feature set combinations using LSI-, CI- and POS-based features. In Phase 1, we conducted isolated experiments on LSI, CI and POS which serve as our baseline. Next, we directly combined POS uni-, bi-, and tri-grams features into LSI and CI in Phase 2 (without the SS). In this phase, we were able to achieve our highest MEAAs using either LSI+POS bi-grams or CI+POS bi-grams on datasets with secondary school levels (i.e. 2010 and 2014 Grades 7-9 datasets) and either LSI+POS tri-grams or CI+POS tri-grams on datasets involving primary school levels (i.e. 2014 Grades 3-6 and Grades 3-9 datasets). Finally, in Phase 3, we performed the SS on the feature set combinations of Phase 2. Results show that the combined content- and grammar-based features, LSI+POS bi-grams and CI+POS bi-grams, generally yields the highest MEAA values in this phase which validates our first research hypothesis stated above.

Research Hypothesis 2

Optimisation of the feature set combination process yields better performing systems.

Research Question 3

How can we efficiently combine and/or augment the content-based features from CI or LSI with the grammar-based features represented by the POS n-grams?

In this study, we have provided evidence that simply adding feature sets together can result in a decline in system performance as shown in the results of our Phase 2 experiments. This also implies that having several features in a language model does not guarantee a higher-performing system. Therefore, researchers in this field should be more cautious in combining the feature sets to achieve optimal results.

The SS discussed in Section 4.6 played a vital role in Phase 3. Using this strategy, we were able to further enhance our system's overall performance by eliminating sparse feature vectors which are prevalent in Phase 2. It served as an optimisation step for us to achieve our best-performing systems with MEAA values ranging from 0.86 to 0.95 for all datasets. In this context, we can say that we have also validated our second hypothesis.

Research Question 4

How can we create a learner-focused reading ability estimation system to be able to recommend reading materials to students in each grade level and to promote self-directed learning?

In Section 2.1.2, we established the close connection between writing and reading abilities. We utilised this connection in this study by initially calibrating our reading ability estimation system using actual written essays by students. By doing this, we are able to gather actual information on the current status of the writing abilities of students in different school grade levels, which inherently allows our system to have better approximation of their corresponding reading abilities. With this, we can say that our system is learner-focused since it is based on real and actual student abilities.

Using our best-performing systems, we also conducted error analysis to have better understanding of our data and our system. Details of this were discussed in Section 5.2.4. We classified the errors into two types, the O-type and U-type error types. An O-type error is defined as a type of misclassification in which the predicted grade level of a document is 2 or more levels higher than its actual grade level. The opposite is true about a U-type error which is defined to occur when the document's predicted grade level is 2 or more levels lower than its actual grade level. In our investigation, we were able to end up with speculations that: 1.) O-type errors occur because of the prevalence of high-level features which are distinct to higher grade level text samples; 2.) U-type errors occur because of the lack of these high-level features which can distinguish them from the lower grade level text samples. These error documents, however, can be interpreted as outliers of their respective actual grade level classifications which could be manifestations of students who have extremely high or

extremely low reading abilities compared to most of the other students in their class. In real school scenarios, outliers such as these happen. With our system mapping them to a different grade level, we can say that it was able to detect these anomalies and that it essentially recommends lower or higher level reading materials to these students which would be ideal for them.

Research Question 5
What performance metrics can we use to validate the effectiveness of the systems?

In all our experiments, MEAA, as explained in Section 4.8, is used as the performance metric of the systems. To validate our output comparisons, we performed statistical significance tests using the Wilcoxon Matched Pairs Signed-Rank Test (Hollander, Wolfe and Chicken, 2013) with a significance threshold of p-value=0.05. With these, we were able to validate the effectiveness of our proposed systems.

6.2 Future Work

With the success of this study, there are still questions left unanswered and options left unexplored. Results of our investigations have also paved the way to new research directions. Hence, in this section, we will present some ideas which future researchers can pursue in relation to this study.

Future researchers can explore the effects of the pre-processing techniques presented in Section 4.4

The researcher can conduct an investigation into the effects of *stemming* and *stop-words removal* (refer to Section 4.4). In Razon (2010), these pre-processing techniques

were implemented and were found to enhance the performance of the classification system. Note that in our study, we did not implement these processes.

Future researchers can investigate the different weighting schemes for the matrix representations of the training, test and reference sets discussed in Section 4.5.1

A comparative study on different weighting schemes applied on the matrices created in Section 4.5.1 can also be interesting. In our implementation, we only used the normalised raw term frequency (i.e. normalised TF) weighting scheme on all our matrices. However, there are other weighting schemes to explore. One of the most popular schemes is the normalised term frequency-inverse document frequency (i.e. normalised TF-IDF). In this scheme, the value of each cell in a matrix is calculated using:

$$TF * IDF = \frac{tf_i * log\frac{N}{n_i+1}}{\sqrt{\sum(tf_i * log\frac{N}{n_i+1})^2}}$$

where tf_i is the raw term frequency of the word token t_i, N is the total number of documents in the datasest, and n_i is number of documents where the token t_i appears Razon (2010). Less content-rich word tokens like articles (i.e. 'a', 'an', 'the') and conjunctions (e.g. 'and', 'or') are given low scores in this scheme even though they most frequently appear in the entire dataset.

Future researchers can look into other sub-clustering algorithms for Section 4.5.2.2

CI's dimensionality reduction step called CD involves the use of the *K-means* clustering algorithm as discussed in Section 4.5.2.2. This step is vital to the algorithm's performance since the concept vectors are created from the output clusters of the

K-means algorithm per grade level. Thus, enhancing this step will have a significant effect on the system's performance. One of the clustering algorithms that the future researcher can consider is the *Fuzzy C-Means* which is utilised in Razon et al. (2010).

Investigation of different kernel functions for the SVM classifier can also be done.

In Section 4.7, we presented the configuration of our SVM classifier. RBF is the only kernel function we used in all the experiments. It will be very interesting to know how the system performs using the *polynomial* and *sigmoid* functions included in R Software's `e1071` package.

Exploration of different feature set combinations is recommended.

It would be interesting to find out what happens if we use the combined POS n-gram feature sets (i.e. 1. uni-grams and bi-grams, 2. bi-grams and tri-grams, 3. uni-grams and tri-grams and 4. uni-grams, bi-grams, tri-grams) together with CI- or LSI-based features. The researcher can also try using combined LSI-, CI- and POS-based features (i.e. LSI+CI+POS). We have not explored these feature sets at all in our experiments.

Application of the proposed approach on larger datasets and on languages other than English can also be very interesting tasks.

One of the powerful features of the proposed approach is *re-trainability*. Although the system is just tested on the English language, we speculate that, with sufficient training materials, it can also be applied on other languages. The researcher would

just have to train the system using text samples and POS tags of the new target language. This has been successfully done by Ong in Ong (2011), where he applied a similar algorithm from Razon (2010) on the Filipino language. We also highly recommend the application of the proposed system on larger datasets to further test the system's performance and to validate our outputs.

Appendix A

Experiments on SVM Parameters as referred in Section 4.7

A.1 Phase 1: Exploratory SVM Parameters Grid Search

This section presents the parameters grid search conducted on the datasets using R's *tune.svm*() function to determine suitable values for the C and γ parameters of the SVMs. Tables A.1 to A.5 summarise the results of this search for each feature set (i.e. LSI, CI, POS uni-grams, POS bi-grams, and POS tri-grams), along with the corresponding Mean Squared Error (MSE) yielded for each random set.

Table A.1: Summary of the SVM Parameters Grid Search for LSI

Dataset	**C**	γ	**MSE**
2010 Gr 7-9	1	0.1	0.5059
	1	0.1	0.5071
	10	1	0.3995
	10	0.1	0.4133
	10	0.01	0.4879
	10	0.01	0.5115
	10	0.1	0.5230
	10	0.01	0.5304
	100	0.01	0.4058
	100	0.01	0.4201
2014 Gr7-9	1	1	0.4866
	10	0.1	0.4264
	10	0.1	0.4265
	10	1	0.4323
	10	0.1	0.4336
	10	1	0.4911
	10	0.1	0.4916
	10	0.1	0.5214
	10	0.1	0.5673
	100	0.01	0.4208
2014 Gr3-6	10	0.1	0.6153
	10	0.1	0.6454
	10	0.1	0.6477
	10	0.1	0.6486
	10	0.01	0.6530
	10	0.1	0.6599
	10	0.01	0.6658
	10	0.01	0.6866
	10	1	0.6937
	10	0.1	0.7057
2014 Gr3-9	10	0.001	2.6105
	10	0.001	2.6194
	10	0.001	2.6369
	10	0.001	2.6482
	10	0.001	2.6520
	10	0.001	2.6707
	10	0.001	2.6742
	10	0.001	2.6745
	10	0.001	2.6810
	10	0.001	2.6837

Table A.2: Summary of the SVM Parameters Grid Search for CI

Dataset	**C**	γ	**MSE**
2010 Gr 7-9	10	1	0.0399
	10	0.1	0.0401
	10	1	0.0430
	10	0.1	0.0431
	10	0.1	0.0433
	10	0.1	0.0452
	10	0.1	0.0464
	10	1	0.0606
	10	0.1	0.0608
	10	0.1	0.0759
2014 Gr7-9	10	1	0.0638
	10	1	0.0711
	10	1	0.0752
	10	1	0.0912
	10	1	0.0937
	10	1	0.0937
	10	1	0.0968
	10	1	0.1016
	10	1	0.1387
	10	1	0.1514
2014 Gr3-6	10	0.1	0.1941
	10	0.1	0.2191
	10	0.1	0.2233
	10	0.1	0.2251
	10	0.1	0.2278
	10	0.1	0.2279
	10	0.1	0.2311
	10	0.1	0.2376
	10	0.1	0.2424
	10	0.1	0.2461
2014 Gr3-9	10	0.1	1.4140
	10	0.1	1.4245
	10	1	1.4382
	10	0.1	1.4451
	10	0.1	1.4749
	10	0.1	1.5135
	10	1	1.5270
	10	0.1	1.5515
	10	0.1	1.5686
	10	1	1.5949

Table A.3: Summary of the SVM Parameters Grid Search for POS-Unigrams

Dataset	**C**	γ	**MSE**
2010 Gr 7-9	1	0.1	0.3840
	10	0.1	0.2642
	10	0.1	0.2707
	10	0.01	0.2719
	10	0.01	0.2720
	10	0.01	0.2871
	10	0.01	0.3230
	10	0.01	0.3388
	100	0.1	0.2631
	100	0.1	0.2973
2014 Gr7-9	1	0.001	0.4975
	1	0.001	0.5410
	1	0.01	0.5892
	10	0.001	0.3467
	10	0.001	0.3604
	10	0.001	0.3823
	10	0.001	0.4070
	10	0.001	0.4239
	10	0.01	0.4952
	10	0.01	0.5446
2014 Gr3-6	1	0.01	0.3273
	1	0.01	0.3920
	1	0.01	0.4544
	10	0.001	0.2985
	10	0.01	0.3218
	10	0.01	0.3347
	10	0.001	0.3587
	10	0.001	0.4146
	10	0.01	0.4348
	10	0.001	0.4530
2014 Gr3-9	1	0.01	1.5232
	1	0.1	1.6937
	1	0.01	1.7783
	1	0.1	1.8193
	10	0.001	1.3548
	10	0.001	1.4462
	10	0.001	1.4654
	10	0.001	1.5722
	10	0.01	1.6112
	10	0.01	1.6636

Table A.4: Summary of the SVM Parameters Grid Search for POS Bi-grams

Dataset	**C**	γ	**MSE**
2010 Gr 7-9	10	0.001	0.1417
	10	0.001	0.1604
	10	0.001	0.1640
	10	0.001	0.1662
	10	0.001	0.1703
	10	0.001	0.1741
	10	0.001	0.1798
	10	0.001	0.1810
	10	0.001	0.1921
	10	0.001	0.1943
2014 Gr7-9	1	0.001	0.2746
	1	0.001	0.3103
	10	0.001	0.2622
	10	0.001	0.2736
	10	0.001	0.2807
	10	0.001	0.2959
	10	0.001	0.3483
	10	0.001	0.3485
	10	0.001	0.3514
	100	0.001	0.3465
2014 Gr3-6	10	0.001	0.1960
	10	0.001	0.2133
	10	0.001	0.2150
	10	0.001	0.2245
	10	0.001	0.2386
	10	0.001	0.2399
	10	0.001	0.2447
	10	0.001	0.2509
	100	0.001	0.2094
	100	0.001	0.2543
2014 Gr3-9	10	0.001	0.8917
	10	0.001	0.9039
	10	0.001	0.9045
	10	0.001	0.9055
	10	0.001	0.9173
	10	0.001	1.0009
	10	0.001	1.0169
	10	0.001	1.0335
	100	0.001	0.9535
	100	0.001	0.9761

Table A.5: Summary of the SVM Parameters Grid Search for POS Tri-grams

Dataset	**C**	γ	**MSE**
2010 Gr 7-9	10	0.001	0.1813
	10	0.001	0.1935
	10	0.001	0.1941
	10	0.001	0.1958
	10	0.001	0.1997
	10	0.001	0.2010
	10	0.001	0.2054
	10	0.001	0.2111
	10	0.001	0.2151
	10	0.001	0.2243
2014 Gr7-9	10	0.001	0.3312
	10	0.001	0.3338
	10	0.001	0.3392
	10	0.001	0.3519
	10	0.001	0.3707
	10	0.001	0.3742
	10	0.001	0.3763
	10	0.001	0.3789
	10	0.001	0.3907
	10	0.001	0.4024
2014 Gr3-6	10	0.001	0.2340
	10	0.001	0.2393
	10	0.001	0.2452
	10	0.001	0.2680
	10	0.001	0.2720
	10	0.001	0.2770
	10	0.001	0.2959
	100	0.001	0.2253
	100	0.001	0.2838
	100	0.001	0.3057
2014 Gr3-9	10	0.001	1.0815
	10	0.001	1.1048
	10	0.001	1.1120
	10	0.001	1.2296
	10	0.001	1.2637
	100	0.001	1.0255
	100	0.001	1.0797
	100	0.001	1.0824
	100	0.001	1.1323
	100	0.001	1.1336

A.2 Phase 2: SVM Preliminary Experiments for γ

In the previous phase, we have established that 10 is the most frequently occurring C value for all datasets in each feature set. In this phase, we set C to 10 and perform preliminary experiments to determine the final values of the γ parameter. For these experiments, we derived the candidate γs to be all the γ values paired with C=10 in the Phase 1 results. However, since there is only 1 value of γ, 0.001, for the POS bi-grams and POS tri-grams feature sets, tests were no longer conducted on these sets.

In Tables A.6-A.8, we present the results of the preliminary tests to derive the final values of γ for the LSI, CI and POS uni-grams feature sets when C=10.

Table A.6: Summary of the EAA Values from the SVM Preliminary Experiment on the LSI Feature Set using C=10

Random Sets	**Candidate γ Values for C=10 in each Dataset**										
	2010 Gr 7-9			**2014 Gr 7-9**			**2014 Gr 3-6**			**2014 Gr 3-9**	
	0.01	0.1	1	0.01	0.1	1	0.01	0.1	1	0.1	0.001
1	0.634	0.662	0.549	0.813	0.734	0.734	0.679	0.654	0.615	0.482	0.475
2	0.535	0.620	0.577	0.766	0.703	0.641	0.628	0.705	0.590	0.525	0.504
3	0.620	0.662	0.620	0.750	0.813	0.672	0.628	0.641	0.590	0.482	0.468
4	0.704	0.690	0.634	0.719	0.703	0.656	0.667	0.590	0.577	0.574	0.546
5	0.592	0.634	0.549	0.797	0.750	0.719	0.641	0.654	0.615	0.504	0.504
6	0.592	0.606	0.521	0.766	0.813	0.703	0.615	0.654	0.628	0.504	0.496
7	0.634	0.718	0.606	0.750	0.734	0.703	0.564	0.628	0.577	0.504	0.539
8	0.704	0.704	0.662	0.797	0.750	0.703	0.603	0.628	0.615	0.574	0.582
9	0.606	0.648	0.662	0.734	0.828	0.734	0.628	0.628	0.551	0.489	0.496
10	0.690	0.704	0.606	0.766	0.734	0.703	0.590	0.628	0.538	0.511	0.475
MEAA	**0.631**	**0.665**	**0.599**	**0.766**	**0.756**	**0.697**	**0.624**	**0.641**	**0.590**	**0.515**	**0.509**

p-values	0.1/0.01:	0.0156	0.1/0.01:	0.6328	0.1/0.01:	0.1875	0.1/0.001:	0.3359
	0.1/1:	0.0039	0.1/1:	0.0039	0.1/1:	0.0020		

Final γ	**0.1**	**0.1**	**0.1**	**0.1**

Table A.7: Summary of the EAA Values from the SVM Preliminary Experiment on the CI Feature Set using C=10

Random Sets	**Candidate γ Values for C=10 in each Dataset**							
	2010 Gr 7-9		**2014 Gr 7-9**		**2014 Gr 3-6**		**2014 Gr 3-9**	
	0.1	1	0.1	1	0.1	1	0.1	1
1	0.901	0.901	0.984	0.953	0.821	0.846	0.702	0.610
2	0.930	0.887	0.969	0.984	0.872	0.897	0.688	0.695
3	0.972	0.986	0.953	0.938	0.782	0.795	0.681	0.681
4	0.887	0.873	0.953	0.969	0.846	0.846	0.723	0.702
5	0.972	0.944	0.969	0.969	0.808	0.808	0.660	0.610
6	0.958	0.958	0.938	0.938	0.821	0.808	0.674	0.702
7	0.901	0.887	0.953	0.984	0.833	0.833	0.681	0.638
8	0.873	0.887	0.969	0.984	0.846	0.833	0.660	0.631
9	0.958	0.972	0.938	0.953	0.795	0.795	0.702	0.674
10	0.915	0.901	0.953	0.984	0.782	0.808	0.674	0.617
MEAA	**0.927**	**0.920**	**0.958**	**0.966**	**0.821**	**0.827**	**0.684**	**0.656**

p-values	0.1/1:	0.250	0.1/1:	0.2656	0.1/1:	0.2812	0.1/1:	0.0430

Final γ	**0.1**	**0.1**	**0.1**	**0.1**

Table A.8: Summary of the EAA Values from the SVM Preliminary Experiment on the POS Uni-grams Feature Set using C=10

Random Sets	**Candidate γ Values for C=10 in each Dataset**											
	2010 Gr 7-9			**2014 Gr 7-9**			**2014 Gr 3-6**			**2014 Gr 3-9**		
	0.001	0.01	0.1	0.001	0.01	0.1	0.001	0.01	0.1	0.001	0.01	0.1
1	0.634	0.606	0.577	0.906	0.719	0.344	0.718	0.782	0.705	0.638	0.631	0.248
2	0.648	0.718	0.620	0.906	0.781	0.344	0.756	0.756	0.462	0.716	0.652	0.603
3	0.732	0.775	0.704	0.750	0.766	0.344	0.705	0.744	0.731	0.567	0.539	0.546
4	0.704	0.620	0.704	0.859	0.703	0.344	0.744	0.769	0.436	0.681	0.610	0.610
5	0.746	0.761	0.732	0.875	0.781	0.344	0.718	0.731	0.679	0.603	0.660	0.227
6	0.803	0.831	0.789	0.719	0.750	0.344	0.769	0.769	0.654	0.723	0.617	0.546
7	0.746	0.648	0.718	0.875	0.719	0.344	0.744	0.744	0.474	0.674	0.546	0.539
8	0.746	0.732	0.704	0.906	0.750	0.344	0.846	0.756	0.526	0.681	0.652	0.645
9	0.803	0.831	0.775	0.734	0.734	0.344	0.692	0.705	0.628	0.674	0.560	0.631
10	0.690	0.662	0.662	0.859	0.672	0.344	0.808	0.744	0.449	0.709	0.624	0.582
MEAA	0.725	0.718	0.699	0.839	0.738	0.344	0.750	0.750	0.574	0.667	0.609	0.518

p-**values**	0.001/0.01:	0.9121	0.001/0.01:	0.0234	0.001/0.01:	0.5469	0.001/0.01:	0.0137
	0.001/0.1:	0.0039	0.001/0.1:	0.0020	0.001/0.1:	0.0059	0.001/0.1:	0.0020

Final γ	**0.001**	**0.001**	**0.001**	**0.001**

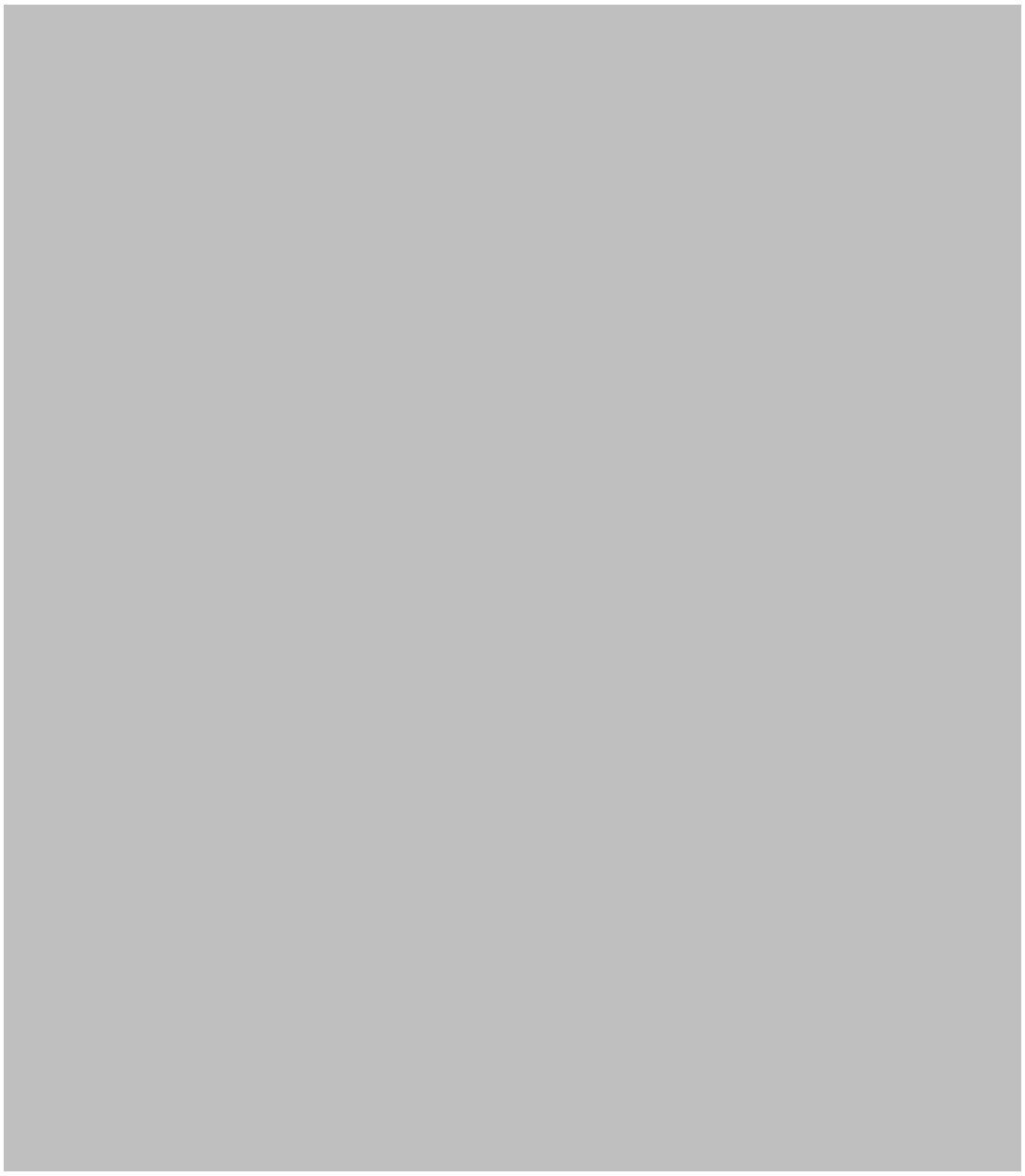
Appendix B

Sample Reference Documents

Excerpt from: The Colors of Us by Karen Katz

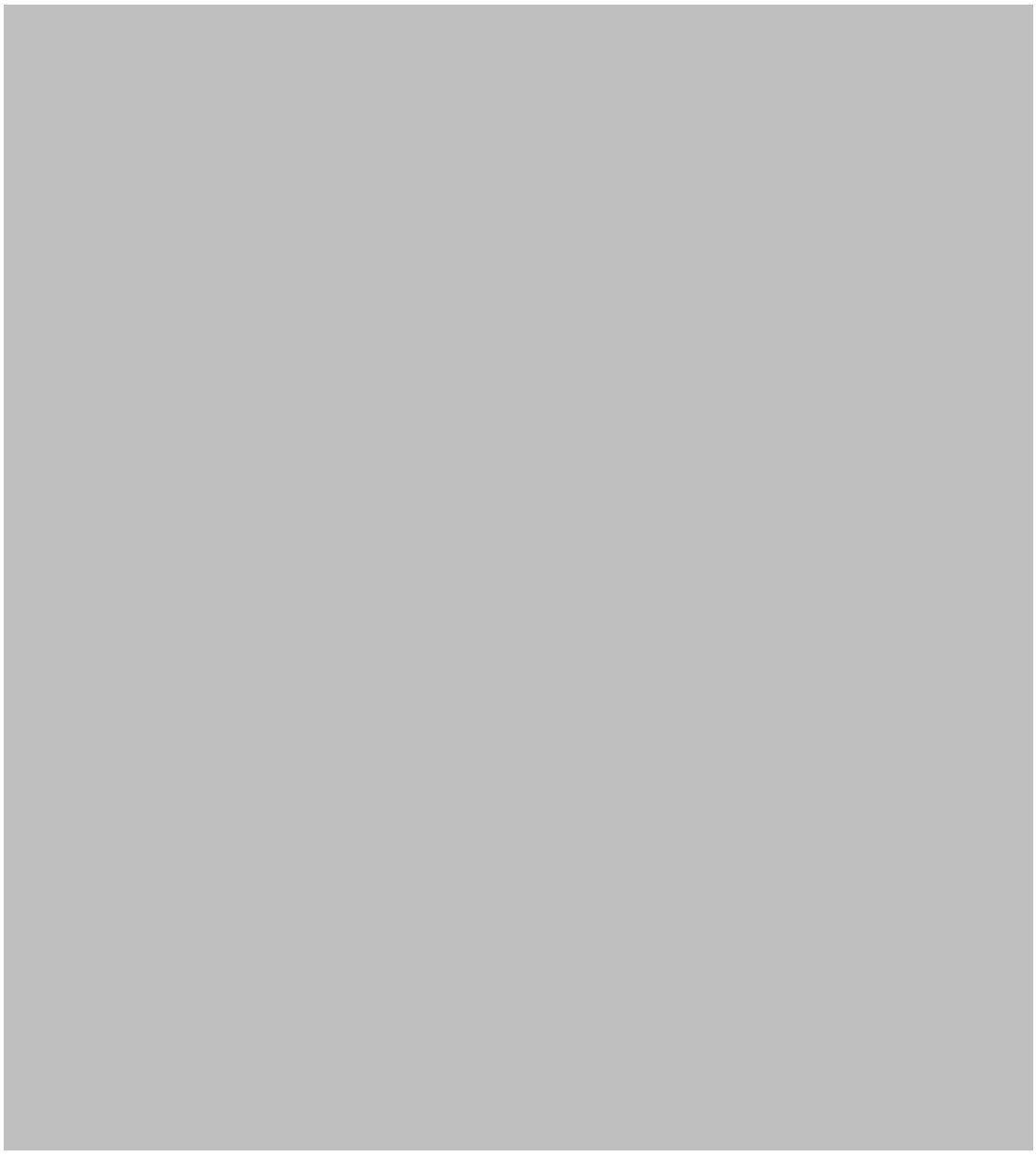
(Grade3 Reference Document 1)

Excerpt from: The Little New Year (Adapted) by Ellen Robena Field



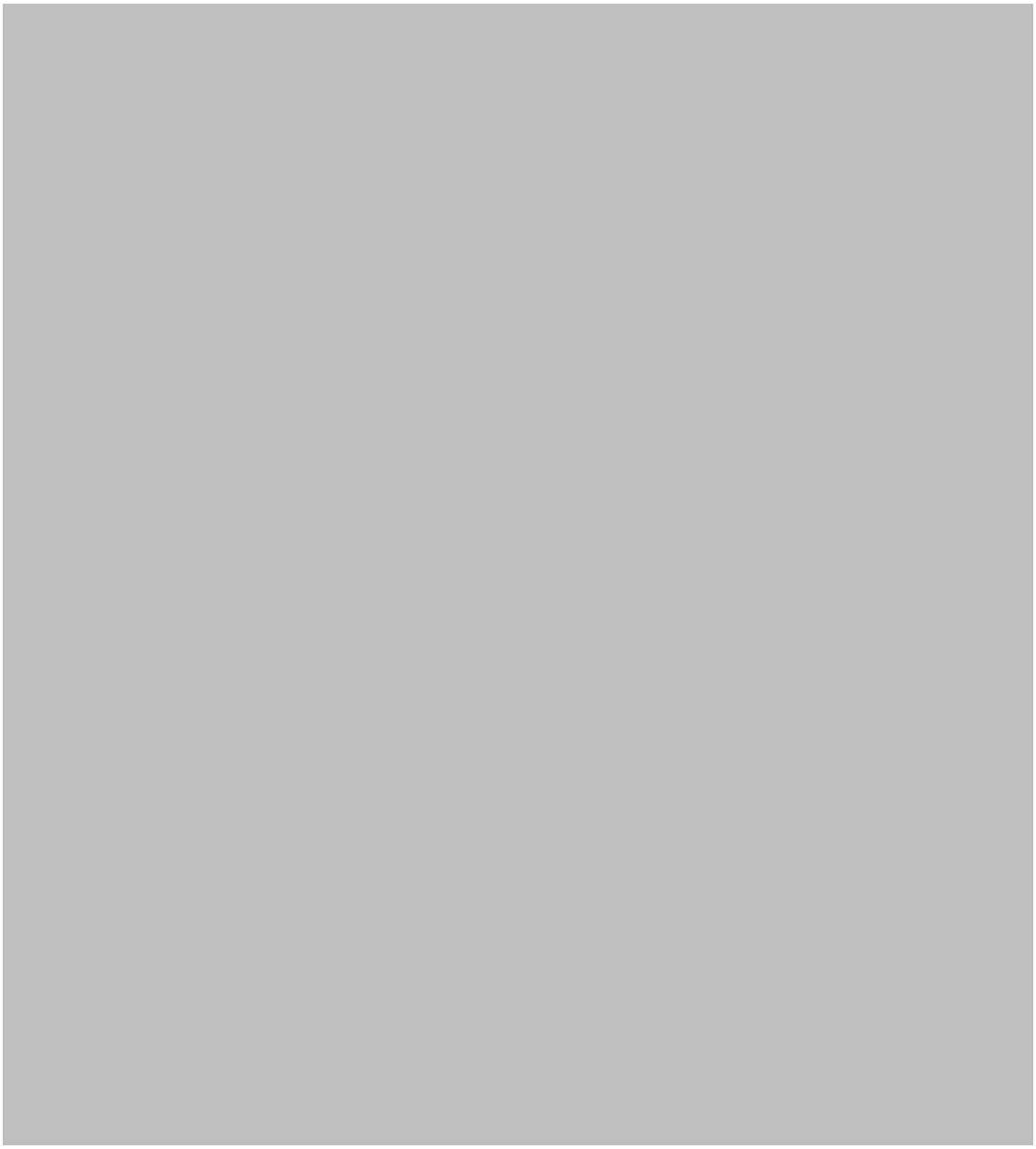
(Grade3 Reference Document 2)

How the Lanzones Became Edible



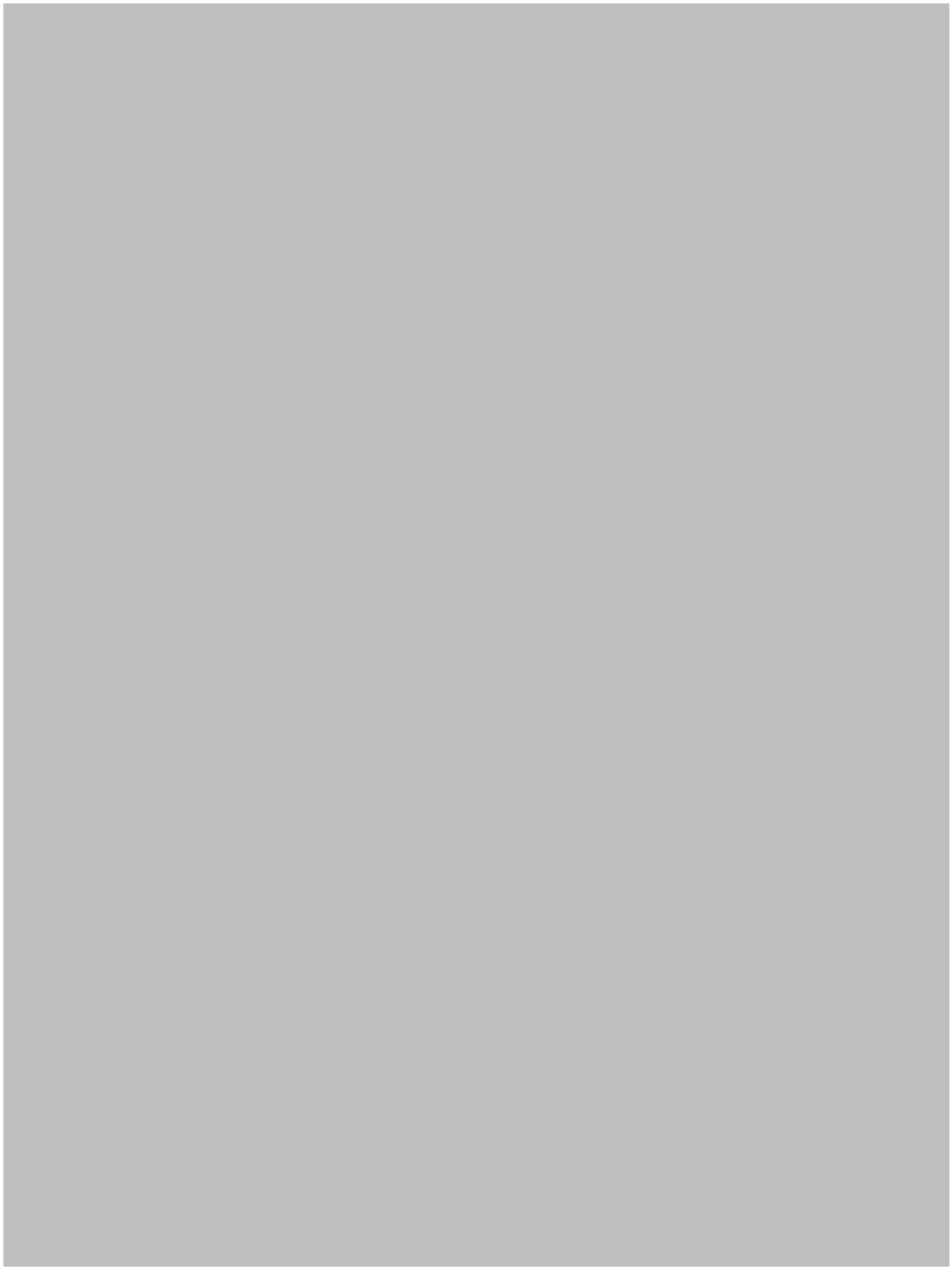
(Grade4 Reference Document 1)

Two Friends, One World - Antonio's Story



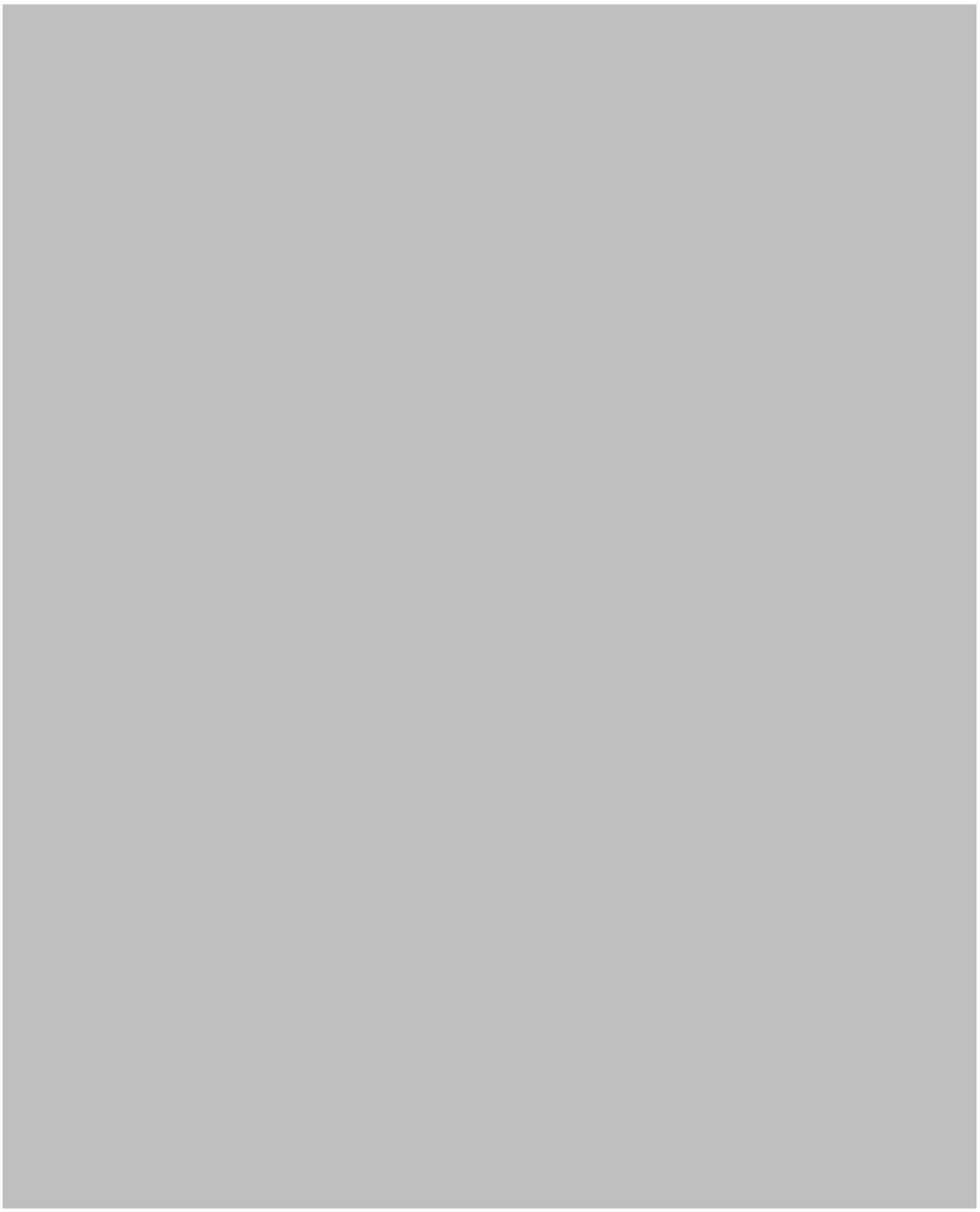
(Grade4 Reference Document 2)

The Earthquake and the Great Wave



(Grade5 Reference Document 1)

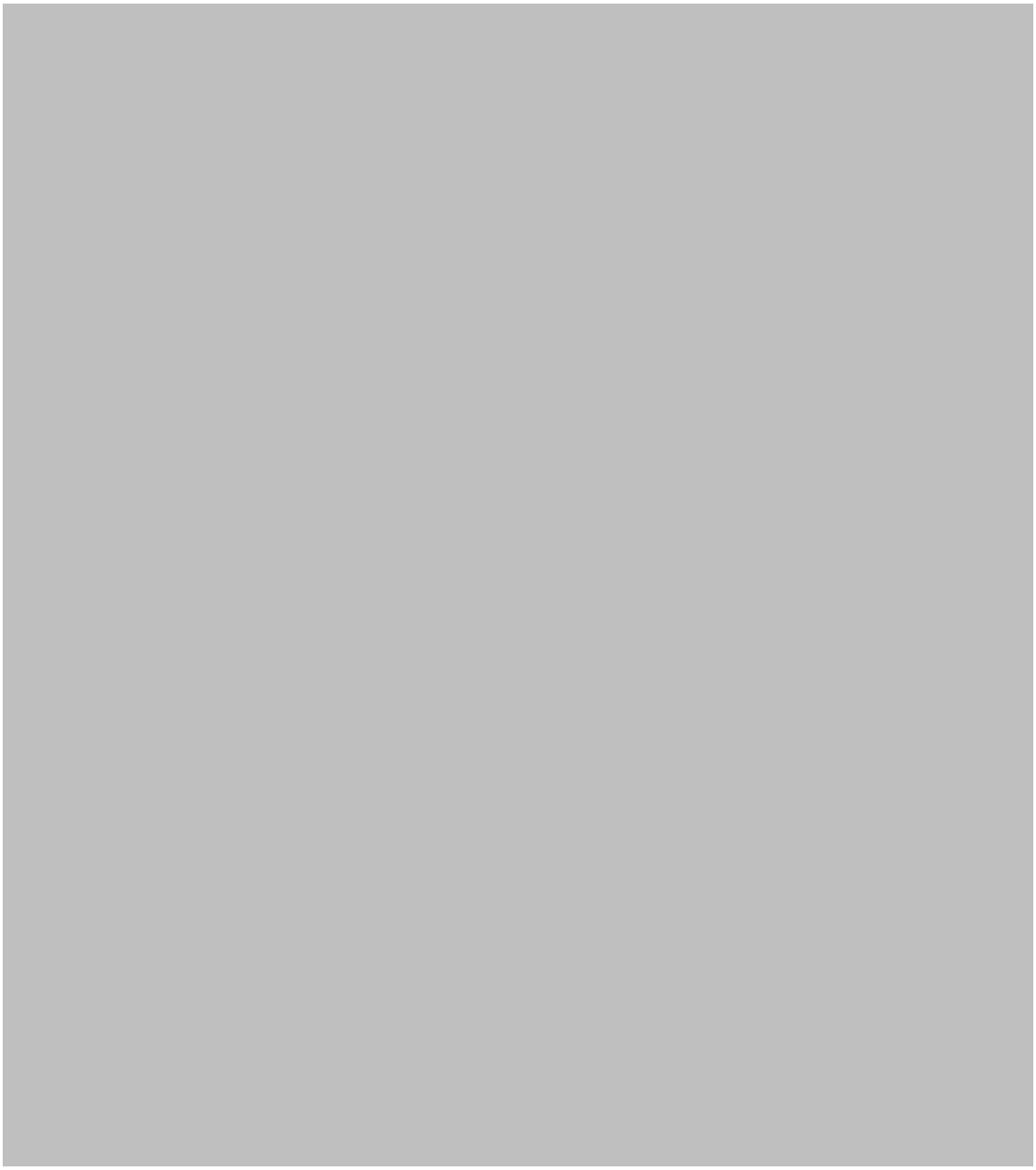
Excerpt from: EQUAL PAY FOR EQUAL WORK



(Grade5 Reference Document 2)

Excerpt from: THE STORY OF A PIECE OF COAL

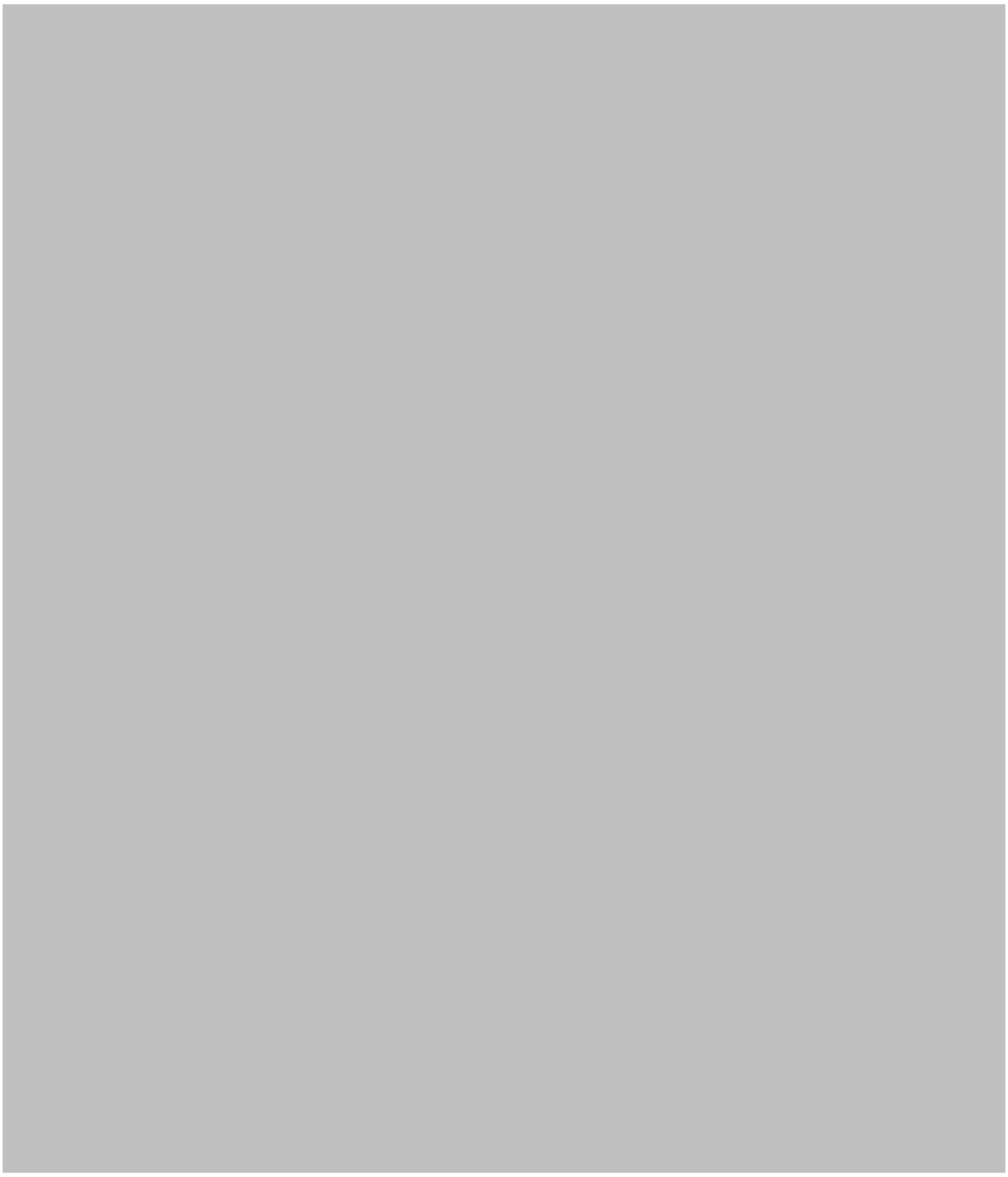
by James P. Moran S.J.



(Grade6 Reference Document 1)

Excerpt from: WHY PLANTS ARE WHAT THEY ARE

by Ray Gesulgen



(Grade6 Reference Document 2)

Story of Maykapal

(Grade7 Reference Document 1)

Excerpts from: Reproductive health bill: Facts, fallacies

(Grade7 Reference Document 2)

Belief in Supreme God

(Grade7 Reference Document 3)

Excerpt from: THE TIGER

(Grade8 Reference Document 1)

Excerpt from: "MY GOD! WHAT HAVE WE DONE?"

(Grade8 Reference Document 2)

Excerpt from: Bound Feet

(Grade8 Reference Document 3)

Excerpt from: ANGLO-SAXON INVASION OF BRITAIN

(Grade9 Reference Document 1)

Excerpt from: THE COMING OF GRENDEL

(Grade9 Reference Document 2)

Excerpt from: The Grapes of Wrath

(Grade9 Reference Document 3)

Appendix C

Part-of-Speech Tag List POS Tag List (2003)

Source:

http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

TAG	DESCRIPTION
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existentialthere
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Appendix D

R Software Packages Used

Package 'e1071'

August 5, 2015

Version 1.6-7

Title Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien

Imports graphics, grDevices, class, stats, methods, utils

Suggests cluster, mlbench, nnet, randomForest, rpart, SparseM, xtable, Matrix, MASS

Description Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, ...

License GPL-2

LazyLoad yes

NeedsCompilation yes

Author David Meyer [aut, cre],
Evgenia Dimitriadou [aut, cph],
Kurt Hornik [aut],
Andreas Weingessel [aut],
Friedrich Leisch [aut],
Chih-Chung Chang [ctb, cph] (libsvm C++-code),
Chih-Chen Lin [ctb, cph] (libsvm C++-code)

Maintainer David Meyer <David.Meyer@R-project.org>

Repository CRAN

Date/Publication 2015-08-05 18:51:12

Source: `https://cran.r-project.org/web/packages/e1071/e1071.pdf`

Package 'koRpus'

August 29, 2016

Type Package

Title An R Package for Text Analysis

Depends R (>= 2.10.0),methods

Enhances rkward

Suggests testthat,tm,SnowballC,shiny

Description A set of tools to analyze texts. Includes, amongst others, functions for automatic language detection, hyphenation, several indices of lexical diversity (e.g., type token ratio, HD-D/vocd-D, MTLD) and readability (e.g., Flesch, SMOG, LIX, Dale-Chall). Basic import functions for language corpora are also provided, to enable frequency analyses (supports Celex and Leipzig Corpora Collection file formats) and measures like tf-idf. Support for additional languages can be added on-the-fly or by plugin packages. Note: For full functionality a local installation of TreeTagger is recommended. 'koRpus' also includes a plugin for the R GUI and IDE RKWard, providing graphical dialogs for its basic features. The respective R package 'rkward' cannot be installed directly from a repository, as it is a part of RKWard. To make full use of this feature, please install RKWard from https://rkward.kde.org (plugins are detected automatically). Due to some restrictions on CRAN, the full package sources are only available from the project homepage. To ask for help, report bugs, suggest feature improvements, or discuss the global development of the package, please subscribe to the koRpus-dev mailing list (https://ml06.ispgateway.de/mailman/listinfo/korpus-dev_r.reaktanz.de).

License GPL (>= 3)

Encoding UTF-8

LazyLoad yes

URL http://reaktanz.de/?c=hacking&s=koRpus

Version 0.06-5

Date 2016-06-05

RoxygenNote 5.0.1

Source: `https://cran.r-project.org/web/packages/koRpus/koRpus.pdf`

Package 'lsa'

May 8, 2015

Title Latent Semantic Analysis

Version 0.73.1

Date 2015-05-07

Author Fridolin Wild

Description The basic idea of latent semantic analysis (LSA) is,
that text do have a higher order (=latent semantic) structure which,
however, is obscured by word usage (e.g. through the use of synonyms
or polysemy). By using conceptual indices that are derived statistically
via a truncated singular value decomposition (a two-mode factor analysis)
over a given document-term matrix, this variability problem can be overcome.

Depends SnowballC

Suggests tm

Maintainer Fridolin Wild <f.wild@open.ac.uk>

License GPL (>= 2)

Encoding UTF-8

LazyData yes

BuildResaveData no

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-08 19:58:09

Source: `https://cran.r-project.org/web/packages/lsa/lsa.pdf`

Package 'NLP'

February 18, 2016

Version 0.1-9

Title Natural Language Processing Infrastructure

Description Basic classes and methods for Natural Language Processing.

License GPL-3

Imports utils

NeedsCompilation no

Author Kurt Hornik [aut, cre]

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

Depends R (>= 2.10)

Repository CRAN

Date/Publication 2016-02-18 15:39:47

Source: `https://cran.r-project.org/web/packages/NLP/NLP.pdf`

Package 'openNLP'

February 18, 2016

Encoding UTF-8

Version 0.2-6

Title Apache OpenNLP Tools Interface

Description An interface to the Apache OpenNLP tools (version 1.5.3).
The Apache OpenNLP library is a machine learning based toolkit for the processing of natural language text written in Java.
It supports the most common NLP tasks, such as tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and coreference resolution.
See <http://opennlp.apache.org/> for more information.

Imports NLP (>= 0.1-6.3), openNLPdata (>= 1.5.3-1), rJava (>= 0.6-3)

Suggests openNLPmodels.en

Additional_repositories http://datacube.wu.ac.at

SystemRequirements Java (>= 5.0)

License GPL-3

NeedsCompilation no

Author Kurt Hornik [aut, cre]

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

Repository CRAN

Date/Publication 2016-02-18 15:39:49

Source: `https://cran.r-project.org/web/packages/openNLP/openNLP.pdf`

Package 'openNLPdata'

June 24, 2015

Version 1.5.3-2

Title Apache OpenNLP Jars and Basic English Language Models

Description Apache OpenNLP jars and basic English language models.

Imports rJava (>= 0.6-3)

SystemRequirements Java (>= 5.0)

License GPL-3

URL http://opennlp.apache.org/,
http://opennlp.sourceforge.net/models-1.5/

NeedsCompilation no

Author Kurt Hornik [aut, cre],
The Apache Software Foundation [ctb, cph] (Apache OpenNLP Java libraries),
JWNL development team [ctb, cph] (JWNL Java WordNet Library)

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

Repository CRAN

Date/Publication 2015-06-24 18:12:14

Source: https://cran.r-project.org/web/packages/openNLPdata/openNLPdata.pdf

Package 'operator.tools'

May 26, 2015

Type Package

Title Utilities for Working with R's Operators

Version 1.4.4

Date 2015-05-25

Author Decision Patterns

Maintainer Christopher Brown <chris.brown@decisionpatterns.com>

Depends utils

Suggests operators, magrittr, testthat

Description These utilities allow for programatically working with R's operators: translating between an operator and its underlying function, inverting an operator, determining an operator's type, etc.

License GPL-2 | file LICENSE

URL https://github.com/decisionpatterns/operator.tools

BugReports https://github.com/decisionpatterns/operator.tools/issues

NeedsCompilation no

Repository CRAN

Date/Publication 2015-05-26 14:43:07

Source:
https://cran.r-project.org/web/packages/operator.tools/operator.tools.pdf

Package 'RWeka'

February 19, 2015

Version 0.4-24

Title R/Weka interface

Description An R interface to Weka (Version 3.7.12).
Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Package RWeka contains the interface code, the Weka jar is in a separate package RWekajars. For more information on Weka see <http://www.cs.waikato.ac.nz/ml/weka/>.

Depends R (>= 2.6.0)

Imports RWekajars (>= 3.7.12), rJava (>= 0.6-3), graphics, stats, utils, grid

Suggests partykit (>= 0.8.0), mlbench, e1071

SystemRequirements Java (>= 6.0)

License GPL-2

Author Kurt Hornik [aut, cre],
Christian Buchta [ctb],
Torsten Hothorn [ctb],
Alexandros Karatzoglou [ctb],
David Meyer [ctb],
Achim Zeileis [ctb]

Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

NeedsCompilation no

Repository CRAN

Date/Publication 2015-01-28 14:40:41

Source: `https://cran.r-project.org/web/packages/RWeka/RWeka.pdf`

Package 'stringr'

April 30, 2015

Version 1.0.0

Title Simple, Consistent Wrappers for Common String Operations

Description A consistent, simple and easy to use set of wrappers around the fantastic 'stringi' package. All function and argument names (and positions) are consistent, all functions deal with ``NA''s and zero length vectors in the same way, and the output from one function is easy to feed into the input of another.

License GPL-2

Depends R (>= 2.14)

Imports stringi (>= 0.4.1), magrittr

Suggests testthat, knitr

VignetteBuilder knitr

NeedsCompilation no

Author Hadley Wickham [aut, cre, cph], RStudio [cph]

Maintainer Hadley Wickham <hadley@rstudio.com>

Repository CRAN

Date/Publication 2015-04-30 11:48:24

Source: `https://cran.r-project.org/web/packages/stringr/stringr.pdf`

Package 'tm'

July 3, 2015

Title Text Mining Package

Version 0.6-2

Date 2015-07-02

Depends R (>= 3.1.0), NLP (>= 0.1-6.2)

Imports parallel, slam (>= 0.1-31), stats, tools, utils, graphics

Suggests filehash, methods, Rcampdf, Rgraphviz, Rpoppler, SnowballC, tm.lexicon.GeneralInquirer, XML

SystemRequirements Antiword (<http://www.winfield.demon.nl/>) for reading MS Word files, pdfinfo and pdftotext from Poppler (<http://poppler.freedesktop.org/>) for reading PDF

Description A framework for text mining applications within R.

License GPL-3

URL http://tm.r-forge.r-project.org/

Additional_repositories http://datacube.wu.ac.at

NeedsCompilation yes

Author Ingo Feinerer [aut, cre],
Kurt Hornik [aut],
Artifex Software, Inc. [ctb, cph] (pdf_info.ps taken from GPL Ghostscript)

Maintainer Ingo Feinerer <feinerer@logic.at>

Repository CRAN

Date/Publication 2015-07-03 10:43:07

Source: `https://cran.r-project.org/web/packages/tm/tm.pdf`

References

Al-Khalifa, S. and Al-Ajlan, A., 2010. Automatic readability measurements of the Arabic text: An exploratory study. *The Arabian Journal for Science and Engineering*, 35(2C), pp.103–124.

Apache OpenNLP Development Community, 2004. Apache OpenNLP developer documentation. Accessed on 10 March 2016, <`https://opennlp.apache.org/documentation/manual/opennlp.html#tools.postagger`>.

Badgett, B.A., 2010. Toward the development of a model to estimate the readability of credentialing-examination materials. *UNLV Theses/Dissertations/Professional Papers/Capstones*, Paper 185. Accessed on 10 November 2016. <`http://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?article=1193&context=thesesdissertations`>.

Biddulph, J., 2002. Guided reading: grounded in theoretical understandings. *Steps to Guided Reading: A professional development course for grades 3.*

Boulis, C. and Ostendorf, M., 2005. Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams. *Proc. of the International Workshop in Feature Selection in Data Mining*, pp.9–16.

Braunger, J. and Lewis, J.P., 2005. *Building a knowledge base in reading 2nd edition.* International Reading Association.

Burdick, H., 2010. The origin of the Lexile Specification Equation. Accessed on 10 November 2016. <https://lexile-website-media-2011091601.s3.amazonaws.com/resources/materials/The_Origin_of_the_Lexile_Specification_Equation.pdf>.

Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), p.27.

Collins-Thompson, K. and Callan, J.P., 2004. A language modeling approach to predicting reading difficulty. *North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pp.193–200.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, 20(3), pp.273–297.

Cullinan, B.E., 2000. Independent reading and school achievement. *School Library Media Research*, 3(3).

Dale, E. and Chall, J.S., 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, 27(1), pp.11–28.

Dale, E. and Chall, J.S., 1949. The concept of readability. *Elementary English*, 26(1), pp.19–26.

Dale, E. and Chall, J.S. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books.

Dale-Chall easy word list text file. Accessed on 10 March 2016, <http://countwordsworth.com/download/DaleChallEasyWordList.txt>.

Deerwester, S.C., Dumais, S.T., Furnas, G.W., Harshman, R.A., Landauer, T.K., Lochbaum, K.E. and Streeter, L.A., Bell Communications Research, Inc., 1989. *Computer information retrieval using latent semantic structure.* U.S. Patent 4,839,853.

Dobša, J. and Dalbelo-Bašić, B., 2004. Comparison of information retrieval techniques: latent semantic indexing and concept indexing. *Journal of Information and Organizational Sciences*, 28(1-2), pp.1–15.

DuBay, W.H., 2004. *The principles of readability.* Impact Information.

DuBay, W.H., 2006. *The classic readability studies.* Impact Information.

Dulay, H.C. and Burt, M.K., 1974. Natural sequences in child second language acquisition. *Language learning*, 24(1), pp.37–53.

Garcia, E., 2006. Latent Semantic Indexing (LSI) a fast track tutorial. *Grossman and Frieders Information Retrieval, Algorithms and Heuristics.* Accessed on 10 November 2016. <http://www.apluswebservices.com/wp-content/uploads/2012/05/latent-semantic-indexing-fast-track-tutorial.pdf>.

Graham, S. and Hebert, M., 2010. *Writing to read: Evidence for how writing can improve reading: A report from Carnegie Corporation of New York.* Carnegie Corporation of New York.

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M., 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proc. of North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pp.460–467.

Heydari, P., 2012. The validity of some popular readability formulas. *Mediterranean Journal of Social Sciences*, 3(2), pp.423–435.

Hollander, M., Wolfe, D.A. and Chicken, E., 2013. *Nonparametric statistical methods.* John Wiley Sons.

Hornik, K., 2016. Package 'openNLP'. Accessed on 10 March 2016, <`https://cran.r-project.org/web/packages/openNLP/openNLP.pdf`>.

Horsmann, T., Erbs, N. and Zesch, T., 2015. Fast or accurate?–a comparative evaluation of pos tagging models. *Proc. of the International Conference of the German Society for Computational Linguistics and Language Technology*, pp.22–30.

Hsu, C.W. and Lin, C.J., 2002. A Comparison of Methods for Multiclass Support Vector Machines. *IEEE transactions on Neural Networks*, 13(2), pp.415–425.

Hsu, C.W., Chang, C.C. and Lin, C.J., 2003. A practical guide to support vector classification. Accessed on 10 November 2016. <`http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`>.

Jin, X. and Han, J., 2010. K-Means clustering. *Encyclopedia of Machine Learning*, pp.563–564.

Karypis, G. and Han, E., 2000. Concept Indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization (Technical Report No. 00-016, University of Minnesota).

Kireyev, K. and Landauer, T.K., 2011, June. Word maturity: Computational modeling of word knowledge. *Proc. of the 49th Annual Meeting of the Association for*

Computational Linguistics: Human Language Technologies-Volume 1, pp.299–308. Association for Computational Linguistics.

Klare, G.R., 1963. *The measurement of readability.* Ames, Iowa State University Press.

Landauer, T.K., Foltz, P.W. and Laham, D., 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3), pp.259–284.

Landauer, T.K., 2011. Pearson's text complexity measure. *Pearson's White Papers.* Accessed on 10 November 2016. <http://www.impact-information.com/PearsonsTextComplexity.pdf>.

Landauer, T.K. and Way, D., 2012. Improving text complexity measurement through the Reading Maturity Metric. *Annual meeting of the National Council on Measurement in Education.* Accessed on 10 November 2016. <http://images.pearsonassessments.com/images/tmrs/Word_Maturity_and_Text_Complexity_NCME.pdf>.

Larsson, P., 2006. *Classification into readability levels: implementation and evaluation* (MS thesis, The Chinese University of Hong Kong).

Lorge, I., 1944. Predicting readability. *Teachers College Record*, 45(6), pp.404–419.

Manning, C., Raghavan, P., and Schütze, H., 2009. *Introduction to information retrieval*, pp.32–34.

Mc Laughlin, G.H., 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8), pp.639–646.

Meyer, D., Hornik, K. and Feinerer, I., 2008. Text mining infrastructure in R. *Journal of statistical software*, 25(5), pp.1–54.

Milone, M., 2009. *The development of ATOS: The Renaissance readability formula*. Renaissance Learning, Incorporated.

Nelson, J., Perfetti, C., Liben, D. and Liben, M., 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. *Council of Chief State School Officers, Washington, DC*.

Ong, D.A., 2011. *Automated content scoring of Filipino essays using Concept Indexing* (MS thesis, University of the Philippines).

Ozasa, T., Weir, G. and Fukui, M., 2007. Measuring readability for Japanese learners of English. *Proc. of the 12th Conference of Pan-Pacific Association of Applied Linguistics*, pp.122–125.

Pang, L.T., 2006. *Chinese readability analysis and its applications on the internet* (Doctoral dissertation, The Chinese University of Hong Kong).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. RBF SVM parameters. Accesses on 10 March 2016, <`http://scikit-learn.org/0.17/auto_examples/svm/plot_rbf_parameters.html#example-svm-plot-rbf-parameters-py`>.

Penn Arts and Sciences, Department of Linguistics, 2003. Alphabetical list of part-of-speech tags used in the Penn Treebank Project. Accessed on 10 March 2016, <`https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`>.

Petersen, S.E. and Ostendorf, M., 2009. A machine learning approach to reading level assessment. *Computer speech language*, 23(1), pp.89–106.

Razon, A., 2010. *A new approach to automated essay content analysis using Concept Indexing* (MS thesis, University of the Philippines-Diliman).

Razon, A., Vargas, M.L., Guevara, R.C., and Naval, P.Jr., 2010. Automated essay content analysis using Concept Indexing with Fuzzy C-Means clustering. *Proc. of the 2010 IEEE APCCS*, pp.1167–1170.

Razon, A., Ledesma, R., Bosque, M.L., Loberas, H., Almario, A.R., Faune, R., Guevara, R.C., Naval, P.Jr., 2011. Readability analysis of grade school reading books using Concept Learning with K-Means clustering. Accessed on 10 November 2016. <https://www.researchgate.net/publication/260383371_Readability_Analysis_of_Grade_School_Reading_Books_using_Concept_Indexing_with_K-Means_Clustering>.

Razon, A., Barnden, J., 2015 A New Approach to Automated Text Readability Classification based on Concept Indexing with Integrated Part-of-Speech n-gram Features. *Proc. of the 2015 Recent Advances in Natural Language Processing*, pp.521–528.

Readability formulas. Accessed on 6 March 2016, <http://www.readabilityformulas.com/>.

Schwarm, S.E. and Ostendorf, M., 2005, June. Reading level assessment using support vector machines and statistical language models. *Proc. of the 43rd Annual Meeting on Association for Computational Linguistics*, pp.523–530. Association for Computational Linguistics.

Si, L. and Callan, J., 2001, October. A statistical model for scientific readability. *Proc. of the tenth international conference on Information and knowledge management*, pp.574–576. Association for Computational Linguistics.

Smith III, M., 2009. The reading-writing connection. *Reading*, 400, p.200L.

Snow, C., 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation. Accessed on 10 November 2016. <http://www.rand.org/content/dam/rand/pubs/monograph_reports/2005/MR1465.pdf>.

Stenner, A.J., Burdick, H., Sanford, E.E. and Burdick, D.S., 2007. The Lexile framework for reading (Technical Report). Metametrics.

Wang, Y., 2006, June. Automatic recognition of text difficulty from consumers health information. *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, pp.131–136. Institute of Electrical and Electronics Engineers.

What is a Lexile Measure? Accessed on 10 March 2016, <http://www.lexile.com/about-lexile/lexile-overview/>.