Computer Assisted Learning Dictionary for Rural Students

A thesis submitted in
partial fulfillment of the requirements for the degree of
Master of Science in Computer Science & Engineering of
BRAC University

by
Annajiat Alim Rasel

in supervision of
Dr Amitabha Chakrabarty

May 2018

Declaration

This is to certify that the research work titled "Computer Assisted Learning Dictionary for Rural Students" is submitted by Annajiat Alim Rasel (ID: 16266004) to the Department of Computer Science and Engineering (CSE), School of Engineering and Computer Science (SECS), BRAC University in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering. The contents of this thesis report have not been submitted anywhere else for the award of any degree. I hereby declare that this thesis is my original work based on the results I have calculated. The materials or work found by other researchers and sources have been properly acknowledged by giving credit where due, through appropriate copyright notices or marks, and through references, etc. on best effort basis. Rights of all other cited work are copyright by their respective authors. For example, all rights of the board text book are reserved by the publisher NCTB itself. It was my honor to have carried out my work under the supervision of Dr. Amitabha Chakrabarty.

Dated: 15 May 2018

Signature of Supervisor

_____

Dr. Amitabha Chakrabarty
Assistant Professor and
Post Graduate Coordinator
CSE Department
BRAC University

Signature of Author

_____

Annajiat Alim Rasel
Student ID: 16266004
M.Sc. in CSE
BRAC University

Committee Members

_____

Dr Amitabha Chakrabarty, Committee Chair

_____

Prof Dr Md Abdul Mottalib, Committee Member

_____

Dr Matin Saad Abdullah, Committee Member

_____

Dr Amina Hasan Abedin, Committee Member

_____

Professor Mohammad Zahidur Rahman, Committee Member

Dedication

To everyone I learnt something from.

Table of Contents

Table of Contents

Table of Contents

List of Abbreviations

ASCII	American Standard Code for Information Interchange
AWL	Academic Word List
BANBEIS	Bangladesh Bureau of Educational Information and Statistics
BD	Bangladesh
BN	Bangla / Bengali Language
BNC	British National Corpus
CFG	Context Free Grammar
COCA	Corpus of Contemporary American English
GDEX	Good Dictionary Examples
HSC	Higher Secondary School Certificate
ICT	Information and Communication Technology
ICT4D	ICT for Development
IDP	International Development Program, Co-Owner of IELTS
IELTS	International English Language Testing System
JSC	Junior School Certificate
NLP	Natural Language Processing
OBE	Outcome Based Education
OCR	Optical Character Recognition
POS	Parts of Speech
PSC	Primary School Certificate
SSC	Secondary School Certificate
TOEFL	Test of English as a Foreign Language
UNICODE	Universal Character Set / Code Point
VST	Vocabulary Size Test

List of Figures

List of Tables

List of Publications

1. **A.A. Rasel,** M.S. Abdullah, A. Chakrabarty (2017) English to Bangla Learning Dictionary for Secondary School Textbook in Bangladesh, 10th Annual International Conference of Education, Research and Innovation (ICERI 2017) Proceedings, pp. 4342-4345, Seville, Spain. 16-18 November, 2017.

2. **A.A. Rasel,** M.S. Abdullah, A. Chakrabarty (2018), Learning Dictionary for Higher Secondary School Textbook in Bangladesh, 7th International Conference on Informatics, Electronics & Vision (ICIEV) & 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 25-29 June, 2018 **(Accepted for Publication)**

Abstract

It is observed that rural students (the target demography) are falling far behind in the countrywide standard tests like SSC and HSC. As high as up to 80% of students from rural areas are failing in national English examinations due to lack of quality and extra-care received by urban students. Training up school and college teachers from one village at a time will take too long to achieve. Furthermore, the extra-care received by urban students like private tuition, the surrounding environment which uses English a lot would be too expensive. It would be almost infeasible to replicate the urban environment in rural areas.

English language skills are usually tested in four key areas, listening, reading, writing and speaking. Public examinations in Bangladesh only tests reading and writing skills. This is done in two parts through English 1st paper Examination and English 2nd paper Examination. These parts focus mostly on reading and writing skills respectively. Both of these skills heavily rely on stock vocabulary. Vocabulary stock depends on how much vocabulary students could learn by reading. On the other hand, being able to read requires some vocabulary. Due to limited vocabulary, it becomes difficult for the students to read and write in English. Sentence construction and being able to express in English is a more advanced task compared to reading. This work focuses on building vocabulary for academic reading for the English 1st part examination. The goal is to assist the rural students in acquiring sufficient vocabulary so that they can read the board book prescribed by the National Curriculum & Textbook Board (NCTB) as a Textbook.

The more words students know, the more they can read. On the other hand, the more they read the more words they will know. It appears to be a chicken and egg problem. Due to an extremely small size of vocabulary known by students from rural areas of Bangladesh, it becomes tough for them to increase their vocabulary size and to be able to read. It leads to nationwide failure in English examination in addition to Mathematics. This work investigates what approach student may follow to acquire more vocabulary to make their reading smoother.

Despite the availability of resources for this purpose, according to our investigation, no empirical research has yet identified the learning environments and the unique learning requirements of the target demography. This work aims to explore the possibilities of automated morphosyntactic tagging of the educational material for these courses and automated extraction of linguistic patterns from the tagged corpora.

“In the name of Allah, most gracious, most merciful”

Acknowledgments

I would like to thank the Almighty Allah for giving me the strength and support to put together this research work.

I gladly thank BRAC University for giving me the opportunity to work as a research intern more than a decade back which stirred my curiosity and enlightened me about the field of natural language processing (NLP). The research internship experience combined with broad-based education of BRAC University greatly motivated me to undertake this interdisciplinary research utilizing Computational Linguistics, ICT for Educational Development (ICT4D), Lexicography, Outcome-Based Education (OBE), etc.

I would like to thank my thesis supervisor for his countless motivation, continuous monitoring, overall guidance and support towards the completion of this research.

Chapter 1: Introduction

Most of the Bangladeshi students reside outside the Dhaka, the capital city of Bangladesh. For a long time, there were two national examinations, namely Secondary School Certificate (SSC) and Higher Secondary School Certificate (HSC). SSC is the examination at the completion of 10th grade and HSC is the examination at the completion of 12th grade. Gradually two more examinations were introduced. Junior School Certificate (JSC) was introduced at the end of 8th grade and Primary School Certificate (PSC) was introduced at the end of 5th grade. There are also similar examinations conducted by Madrasa and Technical Education Board.

Each of the above examinations consists of individual examinations of multiple subjects. Usually, there is a very high failure rate in the English subject. As failing in one subject leads to failure in the whole examination, the failure in English drags down overall result of the examinees [1], [2]. This greatly endangers and raises the significant difficulty for the examinees for their future student life as well as professional endeavors. Poor scores in English examination is not only affecting the examination results, it is also suppressing the whole country's advancement. Any minor improvement in English education may have an enormous effect on the whole nation.

1.1 Motivation

There is a significant difference in the result obtained by rural students and students from urban areas. Rural students experience higher failure rate than urban students. There is a massive qualitative difference between teachers from rural areas and teachers from urban areas [2], [3].

Table 1: Qualifications of English Teachers in Secondary Level in Bangladesh

Background	Quantity	Percentage	Remarks
BA with compulsory 100 marks in English	41,416	52.82%	
BA with 300 marks in English	9,800	12.50%	
BA honors in English	3,452	4.40%	
Masters in English	5,398	**6.88%**	**Too Low**
Bachelor without English	16,104	**20.54%**	**Critical**
HSC Pass	2,245	**2.86%**	**Critical**
Total Teachers	**78,417**	**100.00%**	

Bangladesh Education Statistics 2016 report published by Bangladesh Bureau of Educational Information and Statistics (BANBEIS) shows that 97% of teachers do not have a masters level education in English, 88% of teachers did not study English as a core subject in

their BA/MA, 3% of teachers studied English last during their own HSC, 20% of teachers have completed graduation in subjects other than English [3],[4]. On top of the overall status of teachers' qualification being in such unfortunate status, it may be assumed that majority share of qualified teachers chooses to remain in urban areas. It is demonstrated in the Table 1: Qualifications of English Teachers in Secondary Level in Bangladesh.

Rural students are at a significant disadvantage and falling behind in national examinations because they do not have access to the quality and extra care facilities received by students from urban areas.

1.2 Implications of Dictionary

Due to lack of utilizable helping hands in the surrounding environment, students do not find anyway but to utilize dictionary and guidebooks in an effort to comprehend the given textbook. The monolingual dictionary uses English only to teach English. While it may be suitable for learners who have some stock of vocabulary to build on, it is often impractical for learners with a reduced subset of stock vocabulary. A bilingual dictionary is usually prepared for a large scale of the audience often for a continent or sub-continent [5]. This generic dictionary may not be able to portrait meaning in all senses that the learner needs to know, is unable to indicate which of the meaning is best fit or appropriate for the line that the student is trying to comprehend or the sentence that the learner is trying to construct. Furthermore, vocabulary set available in the dictionary may not sufficiently overlap with the vocabulary set required for the textbook used by the students. Human language is a living mechanism. Word meanings change from time to time as the language evolves. How a phrase or word is used in current times is most relevant for the learners. A regular dictionary may not be able to serve this purpose for the students.

1.3 Difficulty in Learning

In computer science, machines can compress files and text so that it takes a much smaller space to store, transfer and it is convenient. It can be restored to full length anytime. However, when it comes to human, it is impractical to memorize (store) whole text or even the words to human memory. Students find difficulty in remembering all the words let alone the entire book. Difficulty in remembering the words also leads to difficulty in recalling the examples they have seen in the text. Students may learn word meanings from the dictionary, guidebooks, and translations that was done by teachers in classrooms on board, often in native language Bangla orally or classroom boards. As the words are being learned a bit separately from the text, while

learning, the knowledge is not being acquired cohesively. Thus students sometimes face difficulty in understanding the same words found later in the text. It is troublesome for them to recall the meaning from the word. Even if they can remember the individual meaning, it becomes hard to relate the meaning to the sentence at hand or current context. Similarly, it becomes difficult for them to prepare an example of using the word in a sentence (sentence construction).

1.4 Research Contribution

No empirical research has been undertaken with the textbook for English subject for Bangladesh national examinations in order to assist the rural learners of Bangladesh. It is an endeavor to find out how to look at the textbook from a different angle. Extraction of a subset of the tokens was done from the text that may assist students to learn the text quicker with less difficulty and effort. Thus comprehension of the text as well as construction of new sentences based on the words exposed in the text would accelerate greatly. As a proof of concept, mostly the SSC textbook was analyzed and also briefly looked at HSC text book. This is the very first attempt at computational lexicography to build a learning dictionary for rural students, the target audience (demography).

This research signifies making learning process smoother for rural students. In order to do so, it strives to facilitate maximum initial learning in minimum time and effort barriers. This aims to strengthen national education backbone and progress it further. It utilizes ICT for Educational Development as a key part of ICT for Development (ICT4D). If the benefits of the research reach rural students of every corner of the country, then it may have a countrywide very large-scale impact.

To summarize this work,

- Contributes a corpus from the NCTB textbook which will facilitate other researchers to extend this work further in order to enhance and improve Education of Bangladesh. A partial draft version has been made available online at http://bit.ly/2HDLx6K to help future researchers.
- Produces a list of useful words to learn
- Formulates how to learn, what sequence rural students should follow without overwhelming themselves
- Produce a list of group-words that cannot be translated directly through collocations

- Produce concordance that shows how the same words may have separate meanings based on their usage in different sentences

1.5 Assumptions

This work primarily focuses on the SSC textbook for English 1st paper examination. The assumptions are as follows:

- Students are failing in SSC just for English subject mostly
- Target demography aim rural students only
- Rural students do not have access to expensive gadgets and learning materials
- Rural students have very limited vocabulary stock and exposure

1.6 Volume of Work

The NCTB text book is studied by students for over two years. The book consists of 243 pages. The text is not directly available in plain text format from NCTB. On top of that, there are many invisible tokens in each page of the text data extracted from the book. It required semi-automated cleaning many times repeatedly. Then more than 40,000 visible tokens of the book had to be manually inspected to find errors in conversion. There were thousands of special characters which required close inspection. Furthermore visible characters were from different character UNICODE code points, for example, in case of quotes there were many variations, ‘, ’, ', “, ”, '', and then there were ... (a special character with three dots), . . . (three dots separately), --, ---, more repetitions of “-“, ASCII control characters, headers, footers, page numbers, unit (chapter) number, lesson numbers etc. There were the same words written differently, for example, e.t.c. and etc. Such words had to be located and merged from thousands of other words.

1.7 Thesis Outline

Remaining part of this thesis report has been organized as follows:

- Chapter 2 briefly reviews some of the related concepts from the literature and concerns for learning words and preparing the dictionary.
- Chapter 3 discusses the components of the proposed system, its design and semi-automated implementation.
- Chapter 4 explores results found from different approaches taken towards extracting useful tokens from the textbook, sample snippets for the proposed dictionary.
- Chapter 5 synopsizes the whole thesis and suggests future potential derivative work for further research.

Chapter 2: Literature Review

As students are directly or indirectly utilizing dictionaries, it is important that there be a strong overlap between the words available in the dictionary and the textbooks. It is often taken for granted. However, it may not always be true. For example, the word 'arsenicosis' was used in 2010 in the textbook for SSC [6]. Unfortunately, it was not available in the Pocket Oxford Dictionary (Software version) of 1995 [7]. Oxford is very reputed in producing a good dictionary and has a well-founded root [8]. Even though Oxford Learners Dictionary of 2017 [9] have specifically tried to cater for regional needs of the Asian sub-continent [5], still unfortunately, that word was not available. Hence it is important to produce relevant dictionaries suitable for the linguistic market [10] of Bangladesh.

2.1 The Scale of the Problem

As millions of examinees are attempting the examinations, the enormous amount of candidates renders the problem to be very challenging. On the other hand, there is a great lack of English teachers both in terms of quantity and quality to meet the needs of the target audience. Hence it is an instance of chicken and egg problem. A very high number of teachers do not have proper and sufficient training to effectively teach English. A similar situation exists in other countries [11]. This problem is severe in Bangladesh due to the very large scale of the problem.

2.2 Demographic Limitation

Professional development training program for the teachers from one village at a time may take an impractical amount of time to achieve complete training for all teachers. Moreover, the extra-care facilities that are accessible to and often received by urban students, for example, in-institution after-school coaching, private tuition, coaching centers, model tests, a surrounding atmosphere which utilizes English in everyday life, etc. would be very cost-inefficient. Reproducing urban environment in rural settings is impractical. Classroom multi-media, media or language labs with audio-visual aids, or even a family member to practice with is often unavailable. As a consequence of lack of exposure, students, teachers, parents [12] have significantly small vocabulary set.

2.3 Building Dictionary

Early efforts of building dictionary included the COBUILD project which tries to produce suitable and tailored to learners limited domain through lexical computing [13]. A corpus-based approach to learning languages has opened a new dimension in language learning

[14]. It allows to examine and experiment with actual examples from real data and analyze instead of having to rely on assumptions, well-known patterns, and grammar. Thus linguistic competence [15] of the learners is greatly enhanced. Collocational examples clarify word meaning significantly, gives ideas of different usage of the word, bringing in examples from different part so the text [14], [16]–[18].

2.4 Choosing Examples for Dictionary

While reading a regular dictionary, several meanings can be found. Occasionally there are few examples showing usages. It may not be practical for a rural student to comprehend a contextual meaning of a word from one example. Furthermore, the example sentences should utilize more known words and try to avoid words that the students have not seen yet. These examples are called good dictionary examples (GDEX) [16]. It may not be feasible to avoid all unknown words, however, using more known and avoiding less known words may be the key.

Other approaches include Substitutable Defining Formats which is alternative of full sentence definitions [19] and utilizing co-occurrence comparison techniques [20] to learn the words.

It is well known that the more senses are engaged while learning and the more interlinked the learning materials are, study materials are better comprehended [21], [22]. While learning each word, if student can be facilitated related examples directly from the text, they can learn the words better with related examples. Looking at the examples side by side they can relate the word and its usage in a sentence. Reading both at the same time without delay in between may improve learning. Later it becomes easier for them to construct similar sentences as they can unconsciously learn the sentence formulation pattern as well.

2.5 Language Modalities

According to Euro Lingual Teaching Methods [23], languages may be evaluated in four different modalities as illustrated in Figure 2.1. Out of the four skills: listening, reading, writing, speaking, only two of the skills are directly relevant to national examinations of Bangladesh. Question category-wise individual marking and other detailed data is not available for national examinations relating to different subskills of the students. However, the gap can be inferred from the analysis of post HSC English related activities.

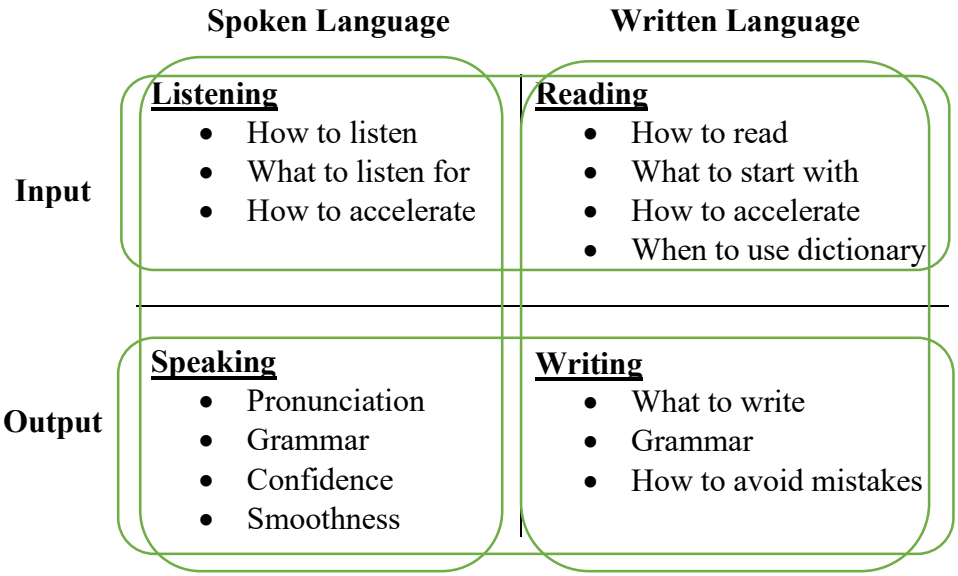


Figure 2.1: Four Quadrants of English Skills

2.6 Post HSC English Related Activities

After HSC, many students go for university admission coaching where English is taught as well. In different universities, there is a varying number of courses, for example:

- Foundation for the very weak students
- Remedial English Course for previous lackings
- Regular University English Course I
- Regular University English Course II
- Advanced English Courses as Elective Courses

Many of the university textbooks and class materials are in English. After graduation, many students go for TOEFL/IELTS coaching hoping to apply for Ph.D. / Masters / Bachelors level education abroad. Even after that much coaching and practice, the similar weaknesses that are found in national examinations, are prevalent in international examinations attempted from Bangladesh as well.

2.7 Inferring modality-wise skills from IELTS Research

For reference, IELTS Test taker performance 2016 report [24] from Teaching and Research Division of IELTS can be analyzed. There are two modules of the IELTS test namely Academic Test for the purpose of higher studies and General Training for the purposes of migration, job, etc.

Table 2: IELTS Bengali Speakers Academic Test, 2016

First Language	Listening	Reading	Writing	Speaking	Overall
Bengali	6.3	**6**	**5.9**	6.2	6.2
Average of 141 First Language Speakers	6.5	**6.4**	**6**	6.6	6.4
Difference from Average	0.2	**0.4**	**0.1**	0.4	0.2
Actual Difference	30 %	**33 %**	**34 %**	31 %	31 %

Table 2 shows average scores of all Academic Module test takers who identified Bangla as their first language on the first row. The next row shows an average of all 141 first language speakers. The third row shows the difference in scores obtained by Bangla speakers compared to all language speakers. All individual skills of Bangla as a first language speakers are weaker than the rest of the first language speakers. Consequently, overall average score is weaker. Writing is the weakest skill, followed by reading skills. It may be inferred that weaker reading skills may have led to weaker writing skills.

Table 3: IELTS Bengali Speakers General Training Test, 2016

First Language	Listening	Reading	Writing	Speaking	Overall
Bengali	5.9	6.4	6.0	6.4	6.2
Average of 134 First Language Speakers	6.3	5.9	6.0	6.6	6.3
Difference	0.4	-0.5	0.0	0.2	0.1
Actual Difference	34 %	**28 %**	**33 %**	28 %	31 %

Table 3 shows average scores of all General Training Module test takers who identified Bangla as their first language on the first row. The next row shows average of all 134 first language speakers. The third row shows the difference in scores obtained by Bangla speakers compared to all language speakers. All individual skills of Bangla as a first language speakers are weaker than the rest of the first language speakers except Reading skills. This may be due to professional exposure of generally aged nature of the test takers of this general training test module. Writing is the weakest skill, followed by listening, reading and speaking skills. It may be inferred that slightly higher reading skills may have led to slightly better writing skills which are equal to the world average of non-native English users. However, the overall average score is still weaker compared to the rest of the users of the language.

Table 4: IELTS Bengali Speakers Average, 2016

Category	Listening	Reading	Writing	Speaking	Overall
Academic	6.3	**6**	**5.9**	6.2	6.2
General Training	5.9	**6.4**	**6**	6.4	6.2
Average	6.1	**6.2**	**6**	6.3	6.2
Actual Difference	30 %	**33 %**	**34 %**	31 %	31 %

Table 4 shows average scores of all individual skills. In average, writing is the weakest skills closely followed by listening and reading. So, focusing on reading and writing is still important. It also demonstrates that overall, both test takers of both modules have shown similar performance.

For another perspective, characteristics of the test takers who have identified themselves as being from the country, Bangladesh can be observed.

Table 5: IELTS Academic Test of Bangladeshi Candidates, 2016

Place of Origin	Listening	Reading	Writing	Speaking	Overall
Bangladesh	6.27	**5.94**	**5.82**	6.19	6.12
Average of 230 Countries	6.48	**6.28**	**5.91**	6.66	6.39
Difference	0.21	**0.34**	**0.09**	0.47	0.27
Actual Difference	30 %	**34 %**	**35 %**	31 %	32 %

Table 5 shows average scores of all Academic Module test takers who identified Bangladesh as their place of origin on the first row. The next row shows average scores of candidates from all 230 countries. The third row shows the difference in scores obtained by Bangladeshi candidates compared to candidates from all other countries. All individual skills of Bangladeshi candidates are weaker than the rest of the candidates. Consequently, the overall average score is weaker. Writing is the weakest skill, followed by reading skills. It may be inferred that weaker reading skills may have led to weaker writing skills. After speaking, Bangladesh is very far from the world average in terms of reading skills.

Table 6 shows average scores of all General Training Module test takers who identified Bangladesh as their place of origin on the first row. The next row shows average scores of candidates from all 213 countries. The third row shows the difference in scores obtained by Bangladeshi candidates compared to candidates from all other countries. All individual skills of Bangladeshi candidates are weaker than the rest of candidates. Consequently, the overall

average score is also weaker. Writing is the weakest skill, followed by reading skills. It may be inferred that weaker reading skills may have led to weaker writing skills. After speaking, Bangladesh is very far from the world average in terms of reading skills.

Table 6: IELTS General Training Test of Bangladeshi Candidates, 2016

Place of Origin	Listening	Reading	Writing	Speaking	Overall
Bangladesh	6.25	**5.73**	**5.93**	6.34	6.13
Average of 213 Countries	6.28	**5.95**	**6**	6.67	6.29
Difference	0.03	**0.22**	**0.07**	0.33	0.16
Actual Difference	30 %	**36 %**	**34 %**	29 %	31 %

Table 7: IELTS Average of Bangladeshi Test Takers, 2016

Categories	Listening	Reading	Writing	Speaking	Overall
Academic	6.27	**5.94**	**5.82**	6.19	6.12
General Training	6.25	**5.73**	**5.93**	6.34	6.13
Average	6.26	**5.84**	**5.88**	6.27	6.13
Actual Difference	30 %	**34 %**	**35 %**	31 %	32 %

Table 7 average scores of all individual skills. On average, reading is the weakest skills closely followed by writing. So, focusing on reading and writing has to be the key priority for Bangladesh. It also demonstrates that overall, both test takers of both modules have shown similar performance.

2.8 Emerged Patterns and Focus Area

Several patterns emerge from the statistics above:

- Reading and writing skills are in general poorer than other skills.
- Test takers who speak Bangla as their first language, have weaker overall skills especially reading and writing skills compared to speakers of other 130+ languages. General Training reading skills are higher, 6.4 out of band score 9, which is still not very high, only 71%.
- Test takers from Bangladesh have greater weakness compared to other 210+ countries. Reading and writing skills are weakest.

- General Training test takers are usually aged, have job experience, looking forward to migration and job abroad. As they have experienced more examples and have gone through more reading, they perform better in “reading” part of the test compared to academic test takers.
- Both academic group and general training group have weaknesses in "writing" part of the test than the rest of the countries as well as the rest of the people speaking their first language, Bangla.

Therefore, from the reliable data from IELTS research [24], it can be inferred that reading and writing needs focus for Bangladesh.

2.9 Cognitive Model of Reading

It is needless to say that writing requires acquiring knowledge through reading first. One of the models for reading assessment is a cognitive model [25]. They have given the model, Figure 2. which tries to describe many details that contribute to reading comprehension for a learner. One of the central ideas for reading is to have a stock vocabulary among many other contributing factors. Other factors including concept for print text, general purposes reading, special purposes reading, strategic knowledge for reading, general strategies, awareness of phonology, decoding skills, word sight knowledge, context, and fluency, recognizing word automatically, having background and structural knowledge in addition to vocabulary leading to language comprehension and consequently reading comprehension.

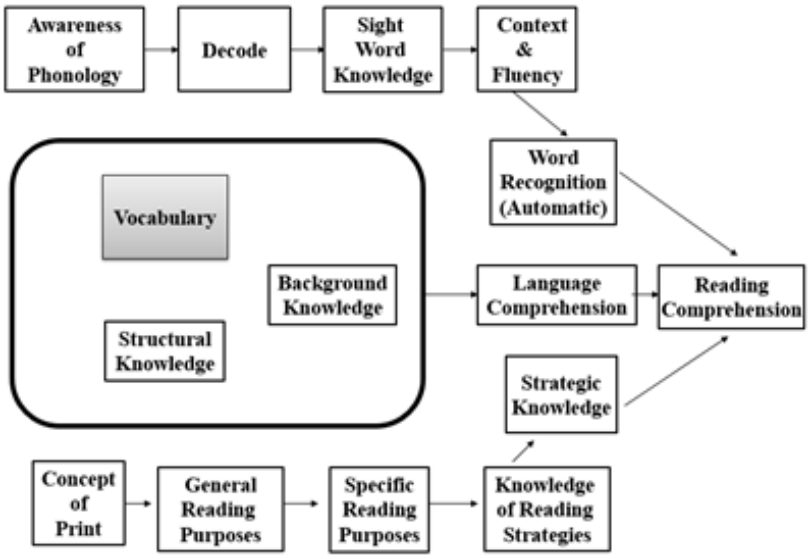


Figure 1: Reading Assessment Cognitive Model

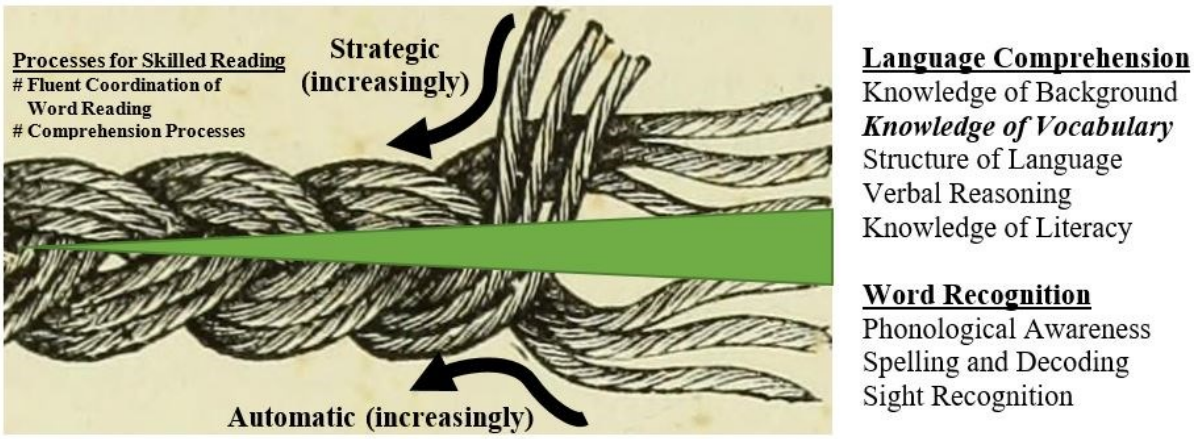


Figure 2.2: Skilled Reading: Many Strands That are Woven into

2.10 Multiple Strands of Skilled Reading

Reading skills is heavily multi-faceted. It is naturally acquired over many years of practice and instructional guidance. Skillful reading has been discussed in [26]. Processes for Skilled Reading involve fluent coordination of word reading and comprehension processes. These are the result of the weaving of many different strands of skills and knowledge. Language comprehension involves knowledge of background, vocabulary, structure of language, verbal reasoning, literacy knowledge etc. Word recognition involves awareness of phonology, the ability to spell and decode, sight recognition etc. Language comprehension increasingly becomes strategic and word cognition increasingly becomes automatic. The fine mixture of both results in smooth and speed reading. This has been demonstrated in Figure 2.2.

2.11 Virtuous Cycle for Reading

Learning individual words, exposure to a linguistically rich oral language, knowledge about word generation, etc. helps in increasing vocabulary. Increasing amount of vocabulary assists in better reading comprehension. Reading comprehension by itself assists in building comprehension strategies, background knowledge, the accuracy of decoding text, improving fluency, etc. Together these three practices form a virtuous cycle for reading as demonstrated in Figure 2.3.

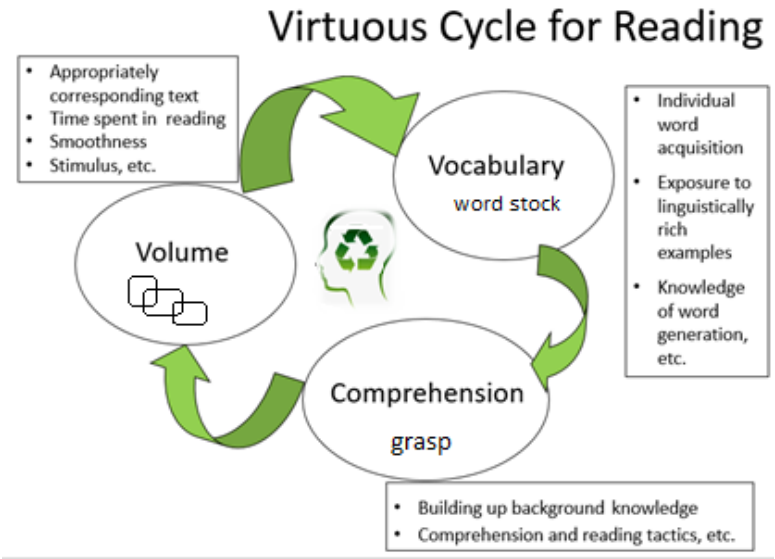


Figure 2.3: Virtuous Cycle for Reading

2.12 Vocabulary Instruction

National Reading Panel (NPR) has published several findings on vocabulary instruction in [27]. According to them, Vocabulary should be taught:

- both directly and indirectly
- by presenting terms in linguistically rich contexts
- through engaging students actively
- through several methods including computer technology
- with re-iteration and exposure to words in several viewpoints
- using task rearrangement

It is clear that the more the students read, the more they will get better at reading. The better they are reading, the more they can read. However, the challenges are determining where to start from, what to learn, how to learn, etc.

2.13 Words in the Oxford English Dictionary

The 2nd edition of the Oxford English Dictionary contains about 0.2 million words. Though the majority of the words are used today in the English language, over 40 thousand words have become archaic. Over 9 thousand words are derivatives. Around 50% are nouns, around 25% are adjectives, 14% are verbs, remaining are exclamations, conjunctions, prepositions, suffixes, etc. There are at least 2,50,000 words and about 7,50,000 words counting different senses and discounting technical and regional words [28].

2.14 Vocabulary Size

According to testyourvocab.com [29], non-native adult English speakers typically know 4,500 words. Their vocabulary size may grow up to around 10,000 words by leaving abroad learning 2.5 words per day on average. Whereas native English speaking of 4 years of age know around 5000 words which 500 words more than a typical non-native adult English user. A native English user crosses adult non-native speakers at the age of just 8 years. Native adult English users knows two to three times more words that are usually learned by non-native English users.

2.15 Word Selection for Vocabulary

There are many other measures of vocabulary. Word selection for vocabulary may be categorized into different tiers as discussed in [30], [31]. They are illustrated in Table 8: Word Selection Tiers for Vocabulary.

Table 8: Word Selection Tiers for Vocabulary

Tier 1	Tier 2	Tier 3
easy, high frequency, known to everyone	words for mature users, suitable in different circumstances	infrequent and subject specific
believe, when, catch,	endure, essential, sinister, benevolent,	tsunami, isotope

In the beginning, the student needs to learn high-frequency tier 1 words. However, the problem arises which ones to learn. It may not be possible for the learner to learn all of the high-frequency words. To be able to comprehend our selected text, SSC textbook [32], at least tier 2 list of words are needed to be able to apply in different situations and environments. Different researchers suggested different lists of words to follow and to test vocabulary. There are over 2000 root words shortlisted from the Dale-Chall list of 3000 words that students should learn before grade 4 [33]–[36]. There is another list of 2000 Academic Word List (AWL) suggested by [37]. The top words cover two-third of the usual texts. There are other many word lists such as 14,000 and 20,000-word lists for learning and vocabulary size test (VST), headwords from British National Corpus (BNC) / Corpus of Contemporary American English (COCA), text coverage etc. [17], [38]–[41].

Chapter 3: System Design and Implementation

A word general word list may not be suitable for the target demography due to greatly limited nature of stock vocabulary combined with resource-scarce environment of the rural students. Although there are traditional learning materials like guidebooks, available for the target demography no empirical research has yet identified the vocabulary learning requirement for these different examination levels. These concerns bring up the following questions:

- Whether there is a need for such a text analysis?
- What is the optimal amount of vocabulary that would have the broadest coverage?

3.1 Design Considerations

Based on the concerns above, following system design concerns may be considered:

- What should be the measure of relevance be for text corpora to be incorporated for content extraction for the target demography?
- How to identify the colocation needed to comprehend texts in these levels of examinations?
- What would be the model of optimal content extraction and pedagogical relevance measurement from relevant corpora?

3.2 Design Objectives

In the circumstance where students are failing in the subjects, they are not being able to grasp the whole book. In alignment with outcome-based education (OBE), students are facilitated with a sub-subset of vocabulary required by majority part of the textbook. Some of the words have separate meaning when used together with some other words compared to a simpler usage of the word. These words need special attention as they do not carry their usual meaning. Their meaning changes based on the collocation of the words. The outcome is students will be able to read more sentences with increased comprehension as a consequence of increased vocabulary size.

3.3 Implementation Algorithm

The implementation was done through the steps as summarized below:

Step 1: **Initial setup:** The setup of utilized tools requires manatee based corpus management tool, bonito, Apache, GNU Bison. Java, Python, PCRE library, and notepad++.

Step 2: **Text Data Extraction:** For extracting text data from rich file formats, Optical Character Recognition (OCR) is used.

Step 3: **Data Cleaning:** Data was automatically and partially pre-processed for cleaning by custom java program and regular expressions. Then data was manually inspected to clean most of the remaining noise as much as possible in a limited duration of time.

Step 4: **Tagging:** Manatee based corpus manager was used to tag the corpus.

Step 5: **Token Extraction:** Python-based tools and libraries was used to query the corpus and extract tagged information.

Step 6: **Data Visualization:** Extracted data was analyzed using python based tools exported to a spreadsheet to generate visualizations and statistics.

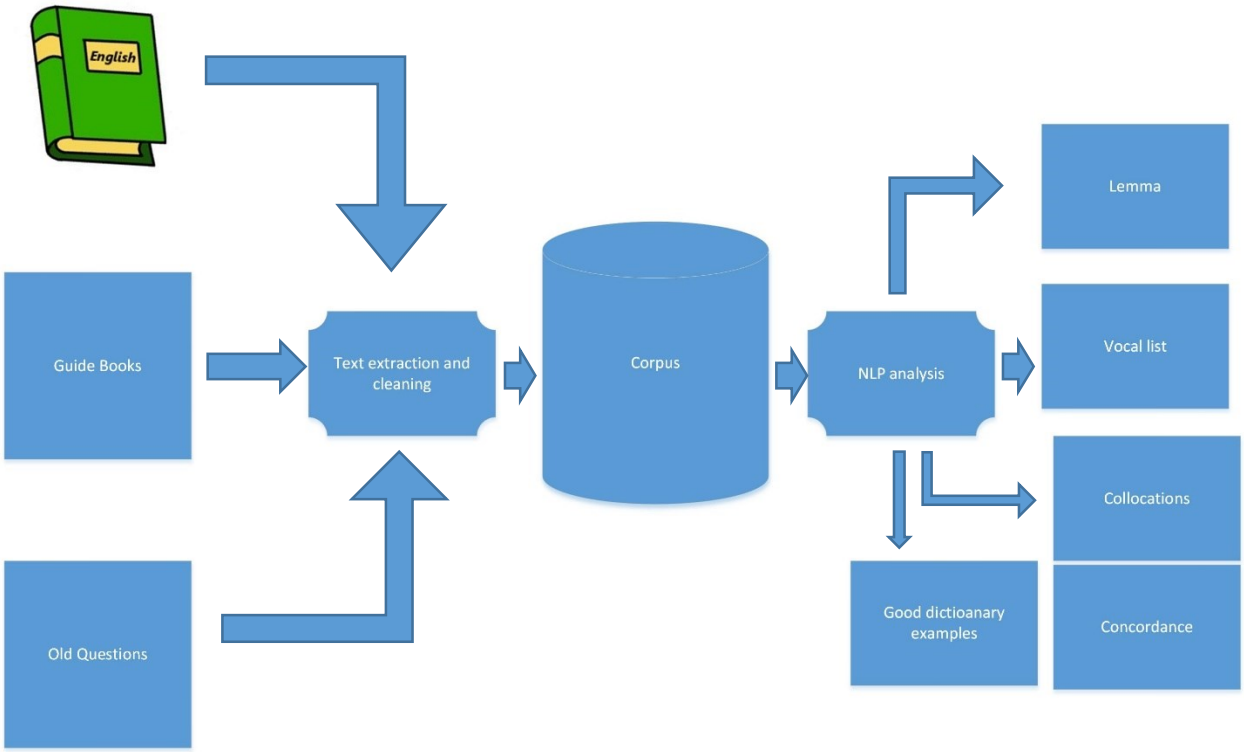


Figure 3.1: System Architecture

3.4 System Components

All steps of the above system are described below as follows:

3.4.1 Data Source

SSC English Textbook, Guidebooks, Old Questions from Test Papers, Model Tests etc. materials relevant to the students of target level are considered as input data source for the system. The download link for the book was yet to be made available in the e-book project website (ebook.gov.bd) at that time. Furthermore, the project provides Adobe Flash animated interface for reading the book which was not compatible for text data extraction. Due to the scarcity of resources, time and other resource constraints, only the book prescribed by the National Curriculum & Textbook Board (NCTB) as a Textbook from the academic session 2013 was used as the data source as shown in Figure 3.2. It is a pre-print version of 2013 textbook. Other textbooks were not available at the time of initiating this research.

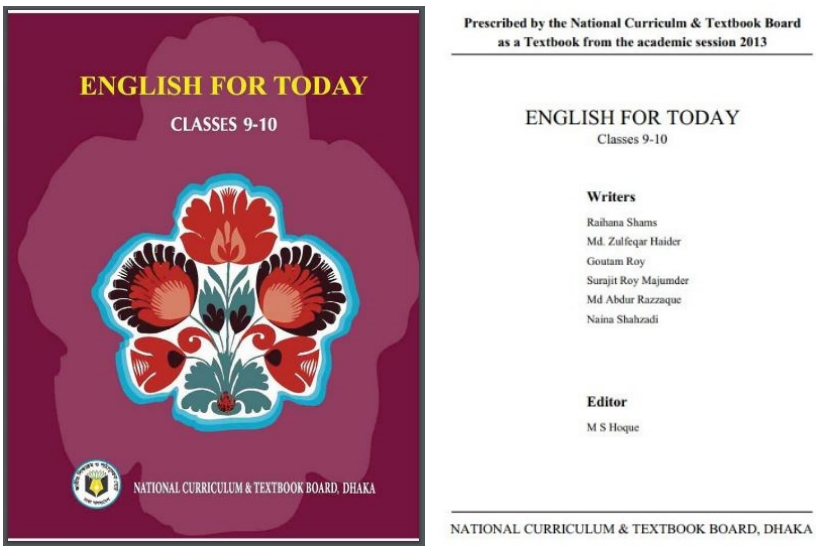


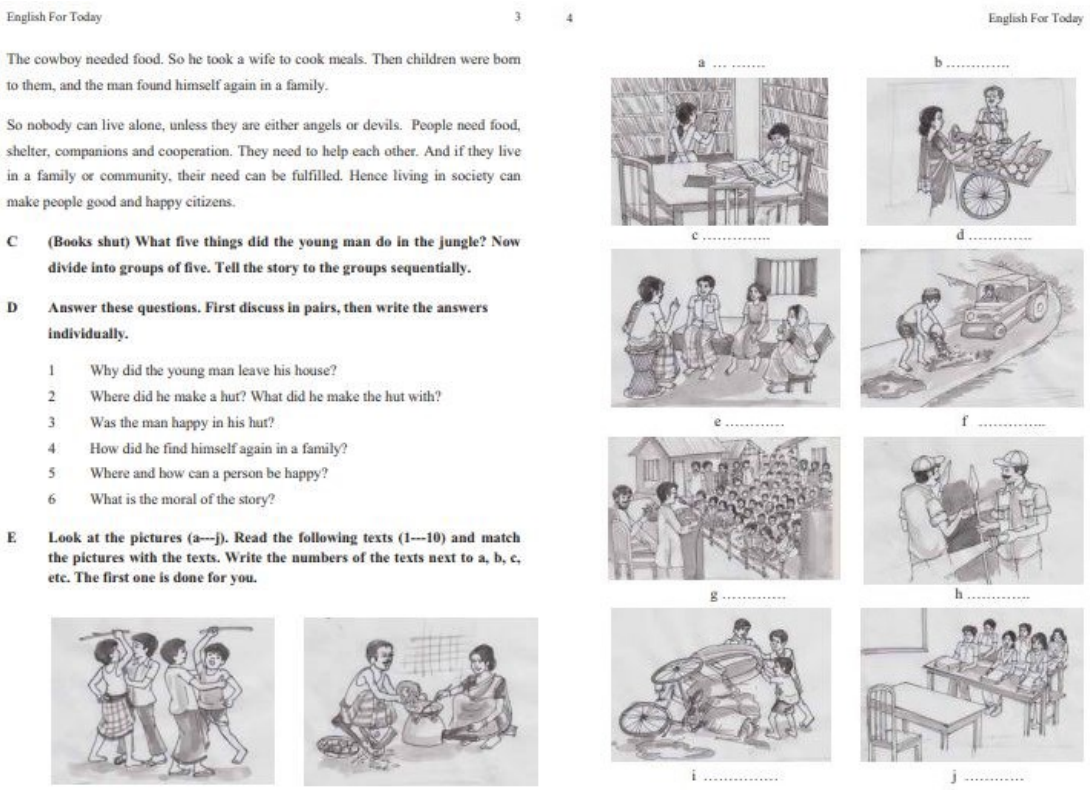
Figure 3.2: Data Source

3.4.2 Data Pre-Processing

First text data is extracted from data sources. All rich file formats are converted to plain text using Optical Character Recognition (OCR). There are many choices of OCRs. Some include Google's & HP's Tesseract, Adobe Acrobat's built-in OCR, Google Chrome Browser's built-in OCR, etc. Initially, built-in Acrobats OCR and Chrome was employed to extract text data from PDF. Later text version of PDF was acquired and the text was extracted through Adobe Acrobat.

Then the extracted data from data source is not readily usable as it is not ASCII or just alphanumeric. It contains lots of formatting, conversion errors, errors introduced by the word processing software used to prepare the data source, and many other UNICODE characters etc. Then data is processed through semi-automatic cleaning. Then manual cleaning is done for errors that do not have a visible pattern. It is skimmed by human eye manually in an effort to remove the irrelevant parts from the data. More than 40,000 visible tokens of the book had to be manually read one by one and carefully inspected to find special characters surrounding the words conversion. There were thousands of special characters which required close inspection. Furthermore, there were multiple variations of visible characters. Each of the variations was from different character UNICODE code points, for example, in case of quotes there were many variations, ‘, ’, ', “, ”, ", quotes and then there were ... (a special character with three dots), . . . (three dots separately), --, ---, more repetitions of "-", ASCII control characters, headers, footers, page numbers, unit (chapter) number, lesson numbers etc. There were the same words written

differently, for example, e.t.c. and etc. Such words had to be located, and merged from thousands of other words. As demonstrated in Figure 3.3, there is "English For Today" header at left-hand side, there is a variable changing page number on right-hand side. To make matter more complicated, odd pages had the book name at left and page number at right, but even numbered pages had these in opposite order, that is page number at left-hand side and book name at right-hand side. There were Activity section numbers A, B, C D, question number 1, 2, 3, picture captions a, b, c, fill in the blanks in multiple forms ___, ..., ..., ..., ranges, --, --- many other formatting, etc.



English For Today 3

The cowboy needed food. So he took a wife to cook meals. Then children were born to them, and the man found himself again in a family.

So nobody can live alone, unless they are either angels or devils. People need food, shelter, companions and cooperation. They need to help each other. And if they live in a family or community, their need can be fulfilled. Hence living in society can make people good and happy citizens.

C **(Books shut) What five things did the young man do in the jungle? Now divide into groups of five. Tell the story to the groups sequentially.**

D **Answer these questions. First discuss in pairs, then write the answers individually.**

1 Why did the young man leave his house?
2 Where did he make a hut? What did he make the hut with?
3 Was the man happy in his hut?
4 How did he find himself again in a family?
5 Where and how can a person be happy?
6 What is the moral of the story?

E **Look at the pictures (a---j). Read the following texts (1---10) and match the pictures with the texts. Write the numbers of the texts next to a, b, c, etc. The first one is done for you.**

4 English For Today

a ..... b .....
c ..... d .....
e ..... f .....
g ..... h .....
i ..... j .....

Figure 3.3: Excerpts from Data Source

Initial attempt at semi-automated cleaning and tokenization utilized java language. It seemed that most of the typing the book was done in Microsoft Word, hence reduced number of errors justified using the Cp1252 encoding. A partial screenshot of the code is shown in Figure 3.4

```
            System.out.println("output file: "+outputFile);
            Path file = Paths.get(inputFile);
            try (
                 Scanner in= new Scanner(file, "Cp1252")){
                String word = null;
                while (in.hasNext()) {
                    word=in.next();
                    process(word.trim());
                }
            } catch (IOException x) {
                System.err.format("IOException: %s%n", x);
                x.printStackTrace();
            }
            out.close();
            unglued.close();
            foundChars.close();
        }
    }
    public static void process(String word) {
        boolean foundSpecial=false,foundLetter=false,foundDigit=false,skippedWord=true;
        for(int i=0;i<word.length();++i){
            if(!Character.isJavaIdentifierPart(word.charAt(i))){//isLetter/digit/$/_
```

Figure 3.4: Beta Version of Initial Semi-Automated Cleaning

In addition to cleaning, sentences were tokenized into words as shown in Figure 3.5.

```
    p(out,word.substring(0,word.length()-2));
    p(out,"<g/>");
    p(out,word.substring(word.length()-2,word.length()));
    return;
}else if(Pattern.matches("^[0-9]+(s)$",word)){//1990s
    skippedWord=false;
    System.out.println("SuffixS broke "+word+" into "+word.substring(0,word.length()-1)+" <g/> "+
    p(out,word.substring(0,word.length()-1));
    p(out,"<g/>");
    p(out,word.substring(word.length()-1,word.length()));
    return;
}else if (Pattern.matches(fpRegex, word)){//number with comma/dot
    //PREV: if(!foundLetter && foundDigit)
    //http://docs.oracle.com/javase/8/docs/api/java/lang/Double.html#valueOf-java.lang.String-
    //Double.valueOf(word); // Will not throw NumberFormatException
    skippedWord=false;
    p(out,word);
    return;
}else if(Pattern.matches("^[=\\—\\-\\-\\.\\...]{3,}$",word)){//==========
    skippedWord=false;
    p(out,word);
    return;
```

Figure 3.5: Initial Version of the Tokenizer

The output from the tokenizer is a vertical text file. It means it contains only one column of data. All tokens from the original text appear into different lines in a single column as demonstrated

in Figure 3.6. Here the <g> tag means glue tag. It presents that food and dot are together. <s> tag represents separate sentence.

TokenNo	Token		TokenNo	Token
926	cowboy		**955**	man
927	needed		**956**	found
928	food		**957**	himself
929	<g/>		**958**	again
930	.		**959**	in
931	</s>		**960**	a
932	<s>		**961**	family
933	So		**962**	<g/>
934	he		**963**	.
935	took		**964**	</s>
936	a		**965**	<s>
937	wife		**966**	So
938	to		**967**	nobody
939	cook		**968**	can
940	meals		**969**	live
941	<g/>		**970**	alone
942	.		**971**	<g/>
943	</s>		**972**	,
944	<s>		**973**	unless
945	Then		**974**	they
946	children		**975**	are
947	were		**976**	either
948	born		**977**	angels
949	to		**978**	or
950	them		**979**	devils
951	<g/>		**980**	<g/>
952	,		**981**	.
953	and		**982**	</s>
954	the			

Figure 3.6: Vertical Text Tokens

3.4.3 **Tagger**

Data is tagged through a Parts of Speech (POS) tagger, which identifies and tags each token of sentences. Tagging can be done manually through brat, an open source (MIT license) rapid annotation tool. It provides an online environment for collaborative text annotation. It can also be done using automatically using other taggers. Python's Natural Language Toolkit (NLTK) also provides a tagger. NLTK tagger was experimented with. Due to the volume of text, budget constraints, an automatic tagger was chosen for our initial prototype, called TreeTagger. Both of the taggers use Penn Treebank Tagset [42], [43].

TokenNo	Token	Tag	TokenNo	Token	Tag
926	cowboy	NN	**955**	man	NN
927	needed	VVD	**956**	found	VVD
928	food	NN	**957**	himself	PP
929	<g/>		**958**	again	RB
930	.	SENT	**959**	in	IN
931	</s>		**960**	a	DT
932	<s>		**961**	family	NN
933	So	RB	**962**	<g/>	
934	he	PP	**963**	.	SENT
935	took	VVD	**964**	</s>	
936	a	DT	**965**	<s>	
937	wife	NN	**966**	So	RB
938	to	TO	**967**	nobody	NN
939	cook	VV	**968**	can	MD
940	meals	NNS	**969**	live	VV
941	<g/>		**970**	alone	RB
942	.	SENT	**971**	<g/>	
943	</s>		**972**	,	,
944	<s>		**973**	unless	IN
945	Then	RB	**974**	they	PP
946	children	NNS	**975**	are	VBP
947	were	VBD	**976**	either	RB
948	born	VVN	**977**	angels	NNS
949	to	IN	**978**	or	CC
950	them	PP	**979**	devils	NNS
951	<g/>		**980**	<g/>	
952	,	,	**981**	.	SENT
953	and	CC	**982**	</s>	NN
954	the	DT			

Each of the tags represents different parts of speech and its various forms. For example, CC means coordinating conjunctions, CD means cardinal number etc. A snapshot of the tagset is demonstrated in Figure 3.7.

3.4.4 **Corpus Management Tool**

Corpus management tool simplifies the toolchain for importing data, tagging, and simplifies queries running. There are a few alternatives available:

BNCweb, a web user interface (UI) for the British National Corpus (BNC), BYU-BNC allows querying the BNC and others corpus from Brigham Young University, CQPweb which

POS Tag	Description
CC	coordinating conjunction
CD	cardinal number
DT	determiner
EX	existential there
FW	foreign word
IN	preposition/subord. conj.
LS	list marker
MD	modal
NN	noun, singular or mass
NNS	noun plural
NP	proper noun, singular
NPS	proper noun, plural
PP	personal pronoun
RB	adverb
SENT	end punctuation
SYM	symbol
TO	to
VBD	verb be, past
VBG	verb be, gerund/participle
VBN	verb be, past participle
VBZ	verb be, pres, 3rd p. sing
VBP	verb be, pres non-3rd p.

Figure 3.7: Example of Tagset

is a web UI facilitates studying different corpora, KonText extended and modified NoSketch Engine by replacing the web UI Bonito, NoSketch Engine facilitates a free open-source corpus management system combining Manatee (back-end) and Bonito (web UI), Sketch Engine provides trial user account for text corpus management and analysis, WordSmith Tools is a collection containing a software package primarily for linguists, etc. Free and open source (FOSS) manatee based sketch engine trial user and noSketch engine augmented by some handwritten tools and libraries written python and Java etc. were used for the prototype. After compiling the corpus, all of the tags are identified and ready for query. A partial snapshot of first draft of the compiled corpus is shown in. Figure 3.8 It is available online at http://bit.ly/2HDLx6K

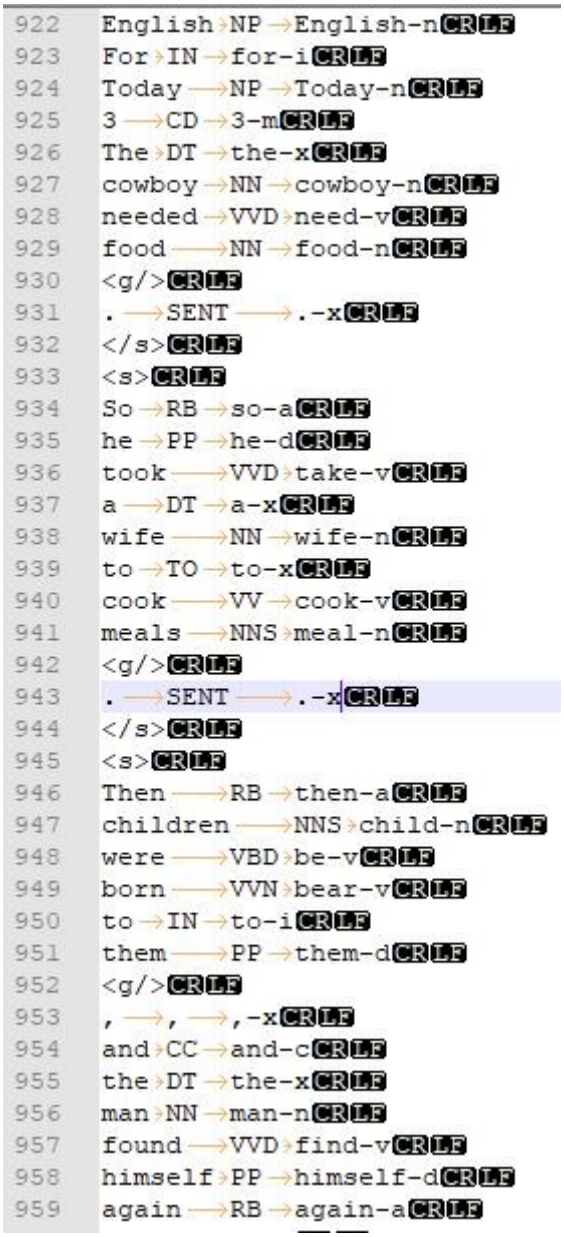


Figure 3.8: Snapshot View of Compiled Corpus

3.5 NLP Analysis

Different tokens and snippets were extracted from the corpus to find and measure suitable components for the desired learning dictionary. It includes lemma, vocabulary list, collocations, concordance and good dictionary examples.

3.5.1 **Lemma**

In computational linguistics, lemmatization process finds root words from a different form of the word. Each root word is called lemma. It is also called dictionary form. It helps in further analyzing groups of many different forms derived from a single lemma. Lemma gives the students quicker access to a broader range of words.

3.5.2 Vocabulary list

List of various types of tokens and words are generated. A subset of the listed is curated for inclusion in the dictionary. This list is larger than the list of lemmas.

3.5.3 Collocations

Multiple words when situated together changes the meaning significantly versus if they were individually used. It is difficult to comprehend the meaning from individual words without significant knowledge and experience. This module produces a list of collocations that students can begin with.

3.5.4 Concordance

Looking at multiple examples of usage of each word facilitates a better comprehension of the meaning of the word. It also assists in learning scenarios where the word may be used. Moreover, it might also help in sentence construction at later phases when students will try to answer questions.

3.5.5 Good Dictionary Examples

If examples for the dictionary are taken from the textbook itself, it will make accelerate reading the textbook. Moreover, if the example uses too many difficult words the students have not stumbled upon yet, it becomes difficult to estimate the meaning of a word. If examples can be chosen such that it ranks sentences with more unknown words lower and brings up easier examples first, it will increase the quality of the dictionary. Such examples are called Good Dictionary Examples (GDEX). This module tries to produce such a list. There may be different approaches to this. Examples can be generated or re-ordered based on the sequence of appearance in the text. This may be computationally expensive. Due to time and resource constraints, this part is yet to be implemented.

Chapter 4: Result and Discussion

Preliminary studies reveal that there are 40,000+ tokens, 6000+ words. Among the words, there are only 5000+ unique words. These words form 2000+ sentences. Further analysis shows that those 6000+ words are rooted in 4000+ lemma. The comparison is demonstrated in Figure 6.

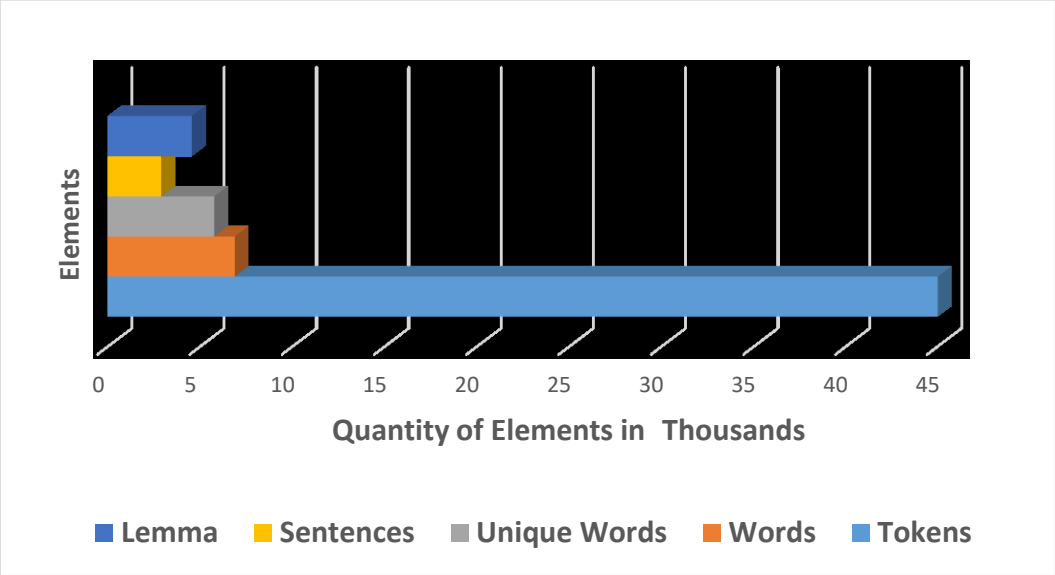


Figure 4.1: Initial Corpus Analysis

4.1 Token Distribution

There are many different types of token in the corpus. They have been identified and tagged using POS tagger. The tokens found includes: adjective, adverb, cardinal number, conjunction, determiner, foreign word, infinitive 'to', interjection, list marker, modal, noun, particle, pronoun, Symbol, that as subordinator, verb, etc. Token distribution has been illustrated in Figure 8. Based on token distribution, it can be observed that major parts of the pie are noun, verb, adjective, and adverb. Therefore, the learning dictionary should mostly focus on these token types.

4.2 Token Coverage

It was observed that covering nouns covers almost more than 40% of the tokens. learning verbs increases it greatly to over 70% of the tokens. Next major increment in coverage are found through adjectives and adverbs which covers about 90% and 95% of the tokens respectively. It gives signals that few tokens types will be major players in the learning dictionary.

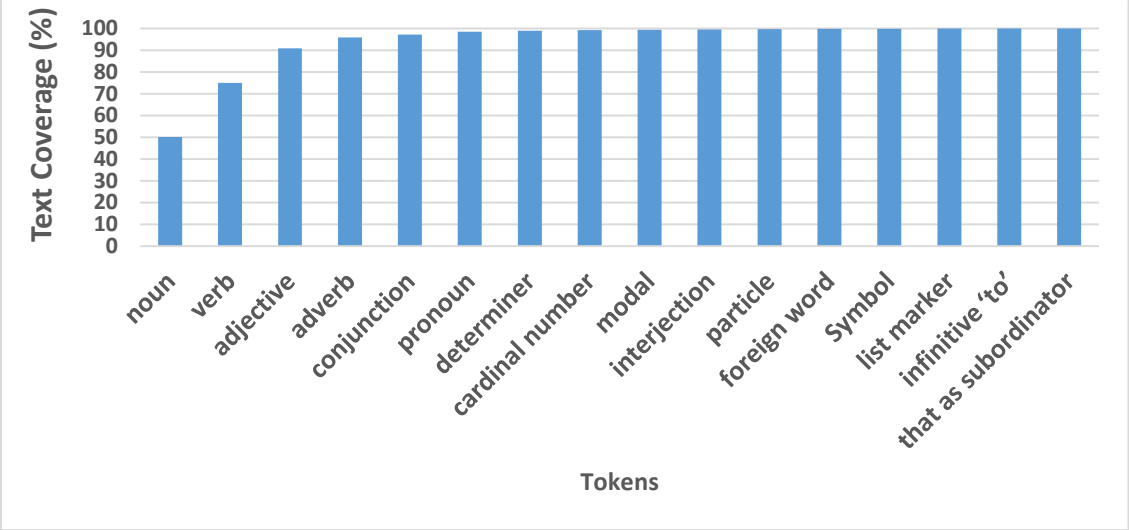


Figure 4.2: Token Coverage

4.3 Text Coverage

Each type of token occupies some portion of the corpus. Learning each type of token will increasingly cover a different part of the text. However, it is needed to choose which type of token to begin with, what sequence to follow and how much of the text is covered by each token. Result of this analysis is shown in Figure 8. It shows that the first six tokens have steadily increased coverage. The sixth one is pronouns which is a closed-set of words. Hence, covering five tokens will have significant return on investment (ROI).

4.4 Token Coverage

Now individually, learning a particular type of token will provide certain amount of benefits. Approaching each types of parts of speech (POS) and other tokens will gradually ensure that all types of POS and other forms are learnt in coherent and well-balanced manner. Majority of the tokens are noun, followed by verb, adjective, adverb, conjunction, pronoun, determiner, cardinal number, modal, interjection, particle, foreign word. Remaining parts are Symbol, list marker, infinitive 'to', and that as subordinator which may be ignored for the learning part initially.

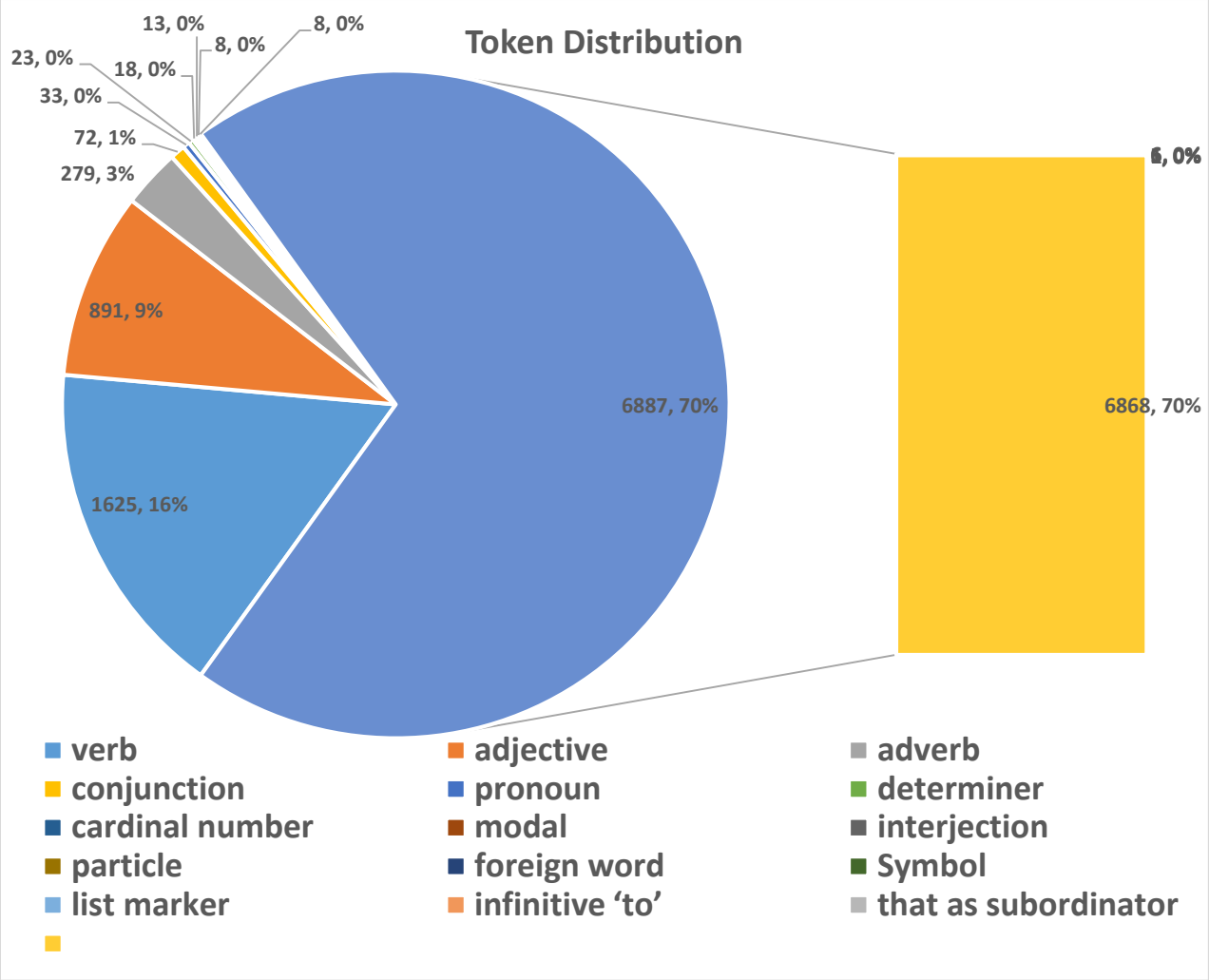


Figure 4.3: Corpus Token Distribution

4.5 Text Coverage

Each of the tokens have multiple appearances throughout the text. Learning each type of token lead to covering some part of the text. It was found that learning nouns covers about 30% of the text. Learning verbs raises the coverage to more than 45%. Adjectives increases the coverage by 10% to about 55%. Learning adverbs consequently will lead to total 60% coverage of whole text. Conjunctions builds up the text coverage to over 70%. Pronouns and determiners expands it further to 80% and 90% respectively. Other types of tokens do not change the coverage significantly due to their low frequency in the text.

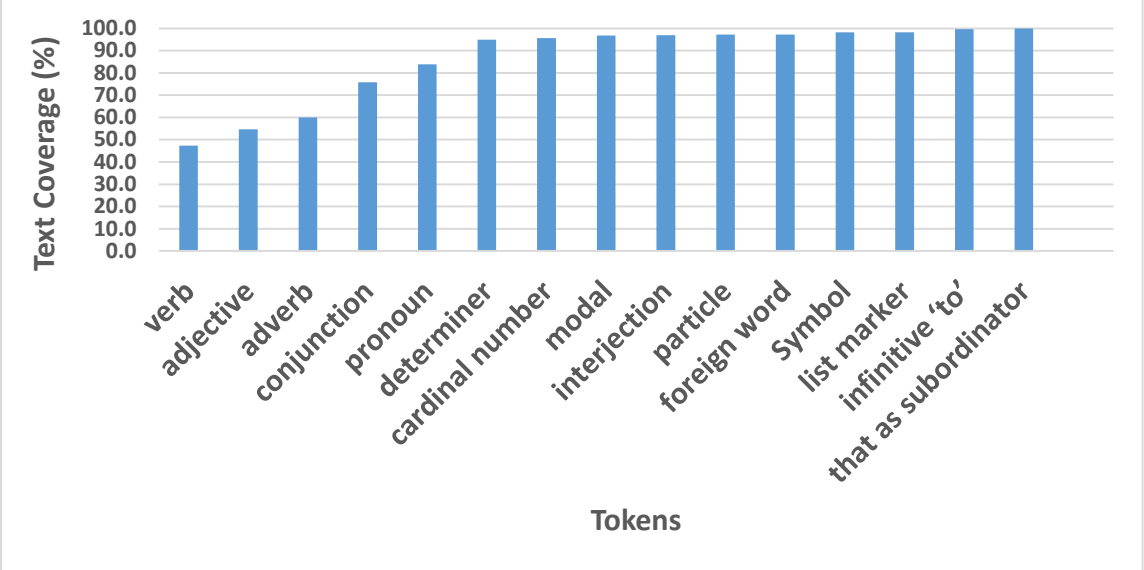


Figure 4.4: Text Coverage by Token

4.6 Token to Text Ratio

The whole text comprises of more than 40 thousand tokens. However, number of unique tokens significantly decreases to over 5 thousand unique tokens. Overall Token to Text ratio is only 15% as shown in Figure 4.5. Individual token ratios have been analyzed as well as shown in Figure 4.6. Adjectives have high text to token ratio rendering a number of adjective have been used many times throughout about 27% of the text. Nouns have 23% ratio which is 2nd highest. Verbs have more than 18% ratio shortly followed by adverbs, a little over 11%. Looking at the token to text ratio shows that in addition to noun, verb, adjective and adverb, some attention needs to be paid to foreign words, interjections, cardinal numbers and modals.

4.7 Token-Based Approach

Based on token coverage, we initially investigated token based approach and analyzed quantity and candidacy of tokens for significant coverage of the text as follows:

4.7.1 Choosing Unique Nouns

Initial analysis showed that there are more than 3.5 thousand nouns in the text. Further investigation revealed that more than 1 thousand out of those over 3.5 thousand nous are plural and singular proper nouns. This consists of mostly names of persons and places. Filtering names of persons and places greatly reduces number of nouns to be enlisted in the dictionary and to be learn by the students. The result is a bit over 2.5 thousand nouns which is a very large number for a rural student to learn. Coverage of

unique noun is illustrated in Figure 4.7. It shows that initially learning about 300 nouns may cover more than 50% of the nouns. As more nouns are learnt, there are diminishing returns. Learning 600 nouns will cover more than 70% of the nouns.

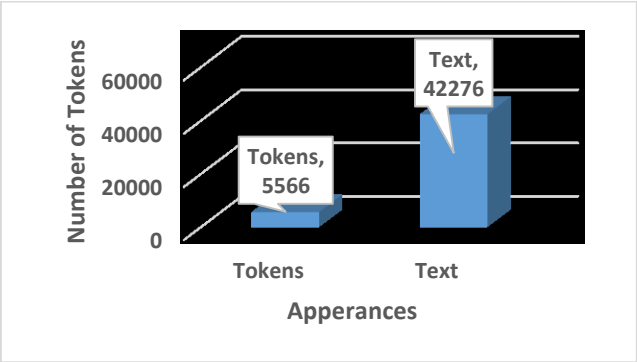


Figure 4.5: Overall Token to Text Ratio

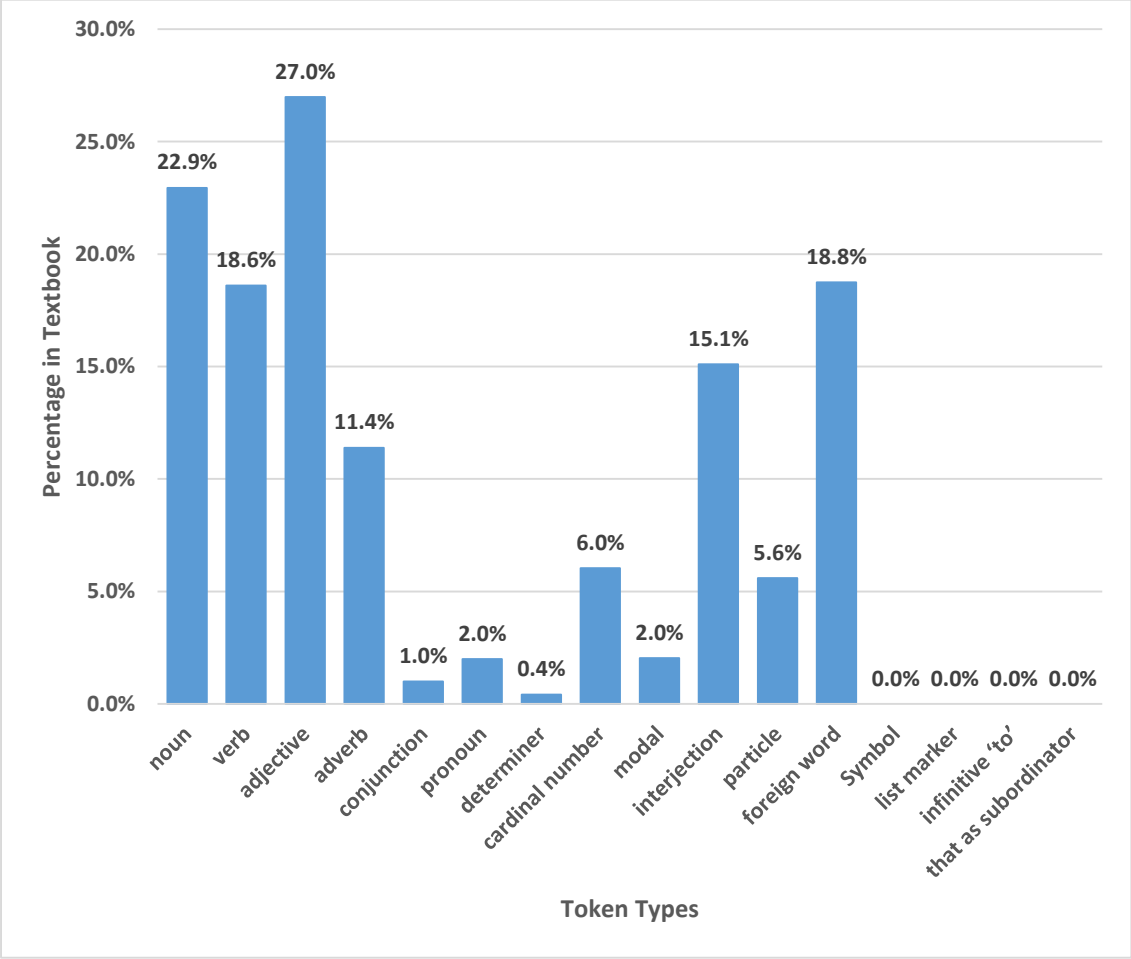


Figure 4.6: Specific Token to Text Ratio

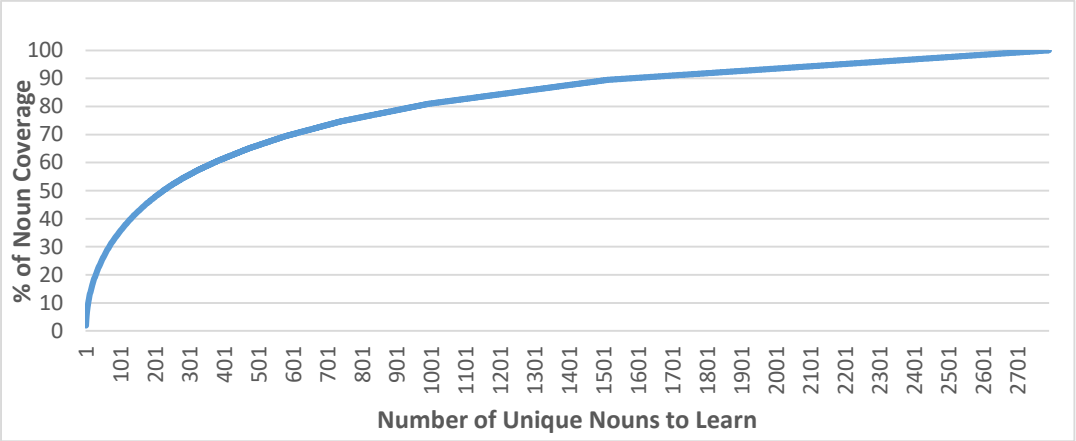


Figure 4.7: Coverage of Unique Nouns (%)

4.7.2 Choosing Unique Verbs

Compared to nouns, there are smaller number of verbs. There are more than 1300 unique verbs in the textbook. The return on learning verbs rises very quickly at the beginning compared to nouns. Learning first 100 verbs covers more than 60% of the verbs. Learning 2nd 100 verbs facilitates over 70% of the verbs. Similar 10% increase is also visible for the next 100 words. The coverage percentage of learning every 100 verbs is shown through Figure 4.8.

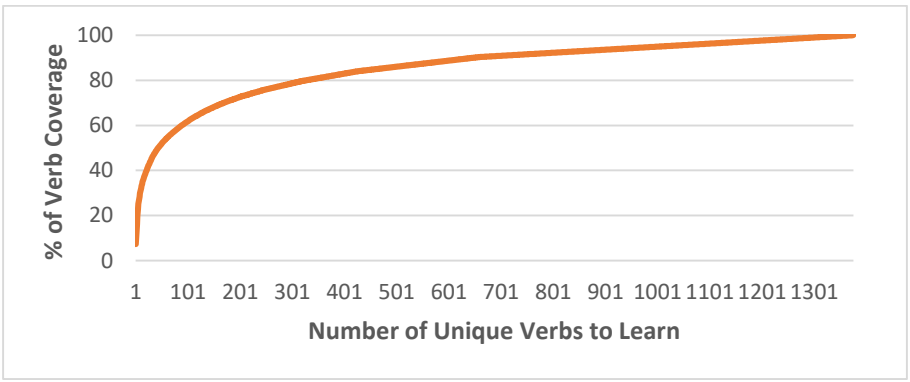


Figure 4.8: Coverage of Unique Verbs (%)

4.7.3 Choosing Unique Adjectives

The text book uses less than one thousand adjectives. Very close to the case of verbs, learning first 100 unique adjectives covers more than 50% of the adjectives. Learning 2nd 100 adjectives also have a very high return up to about 70% coverage. Then there are increasingly diminishing return for learning newer adjectives. It is illustrated through Figure 4.9.

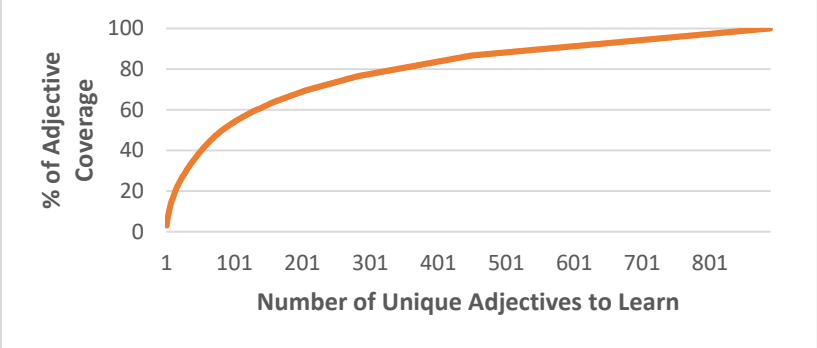


Figure 4.9: Coverage of Unique Adjectives (%)

4.7.4 Choosing Unique Adverbs

The corpus analysis showed that there are less than 300 adverbs only. Learning first 100 adverbs covers about 90% of the adverbs. Rest about 200 of the adverbs cover only about 10%. First 50 adverbs may be chosen for initial learning which covers about 80% of the adverbs. It is demonstrated in Figure 4.10.

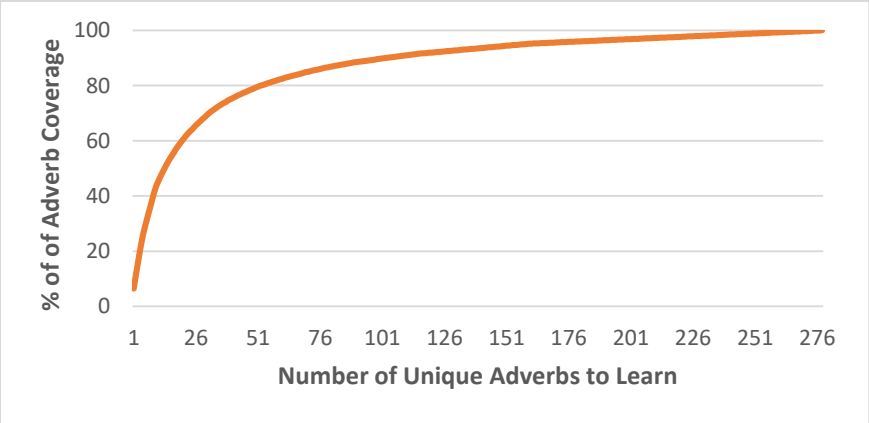


Figure 4.10: Coverage of Unique Adverbs (%)

4.7.5 Choosing Unique Conjunctions

Conjunctions have a very closed set. It means there are fixed number of conjunctions in the English language. There are less than 80 conjunctions in the text book. Learning 10 conjunctions covers about 80% of the conjunctions. As this is a short list, top 15 conjunctions may be chosen to have a 90% coverage of the text. Conjunction coverage is shown through Figure 4.11.

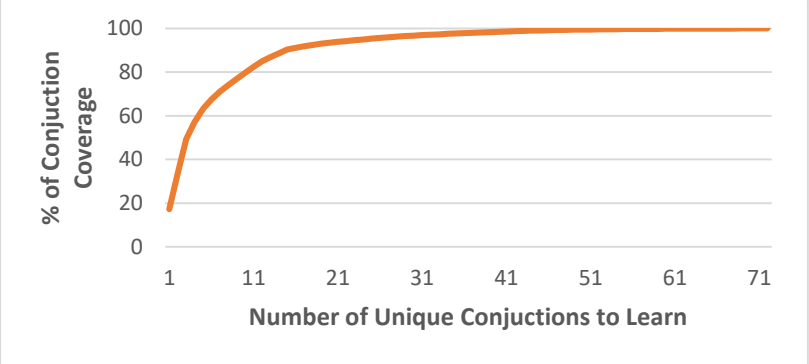


Figure 4.11: Coverage of Unique Conjunctions (%)

4.7.6 Choosing Pronouns

Pronouns are also a close set of words in English language like conjunctions. There about 30 pronouns used in the text book. 15 pronouns were used more than 100 times. 10 pronouns were used less than 10 times. 50% of the pronouns may be chosen for initial learning leading to coverage of 90% of the pronouns. Coverage enhancement for pronouns is shown via Figure 4.12.

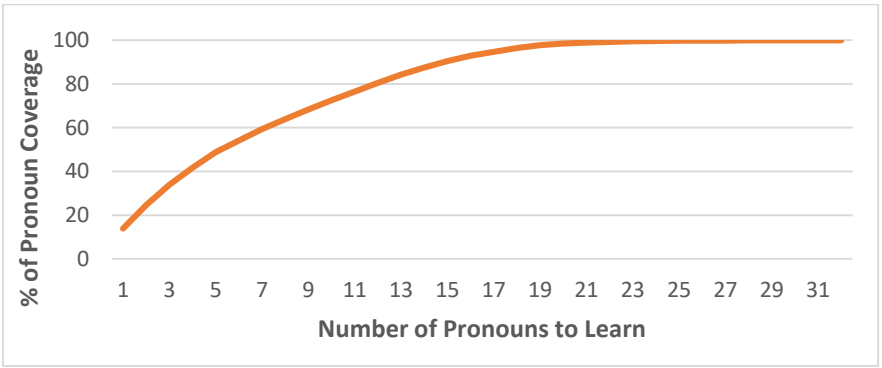


Figure 4.12: Coverage of Unique Pronouns (%)

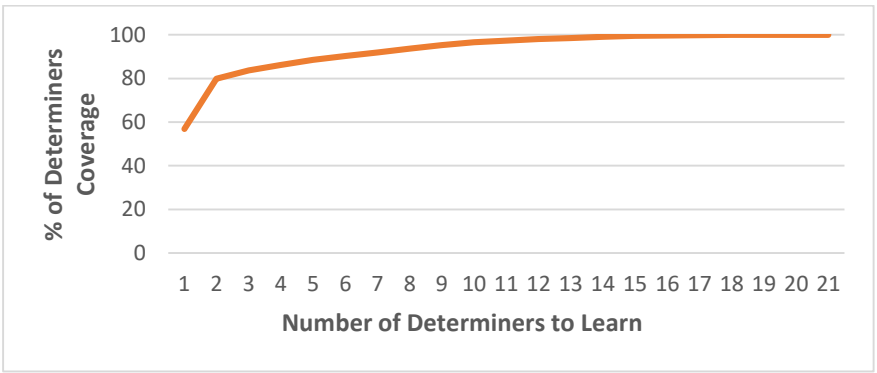


Figure 4.13: Coverage of Unique Determiners (%)

4.7.7 Coverage of Miscellaneous Tokens

There are several miscellaneous tokens which appear few times in the text. Symbols, list markers and infinitive-to are excluded from the discussion as they are not significant candidates for the dictionary. Other tokens are discussed as follows:

4.7.7.1 Determiners

There are about 20 determiners in the text. Most of them were used more than 10 times. Only two determiners were referred only once. They are “quite” and “half”. 50% of the determiners cover over 95% of the tokens as shown in Figure 4.13.

4.7.7.2 Cardinal Numbers

There are less than 20 cardinal numbers. Students at the target level SSC, already know most of the cardinal numbers. However, three cardinal numbers stood out from the list of tokens. Those candidates for the learning dictionary are “million”, “billion”, and “trillion”. Though they occur combined 15 times, their meaning is uncommon to rural students and the digit grouping is different in the subcontinent compared to the rest of the world as they use lac separator.

4.7.7.3 Modals

There are about 10 modal tokens. Due to being very small in number, all of them may be chosen as candidates for the learning dictionary. They are: will, would, could, may, must, should, might, shall etc.

4.7.7.4 Interjections

There are less than 10 interjections. All of them are candidates for inclusion in the dictionary. Standing out examples are: oh, ah, well.

4.7.7.5 Particles

Fewer than 10 particles are used in the text. They are: out, up, off, down, away, on, over, in. Most of the particles carry their usual meaning. However, when used in certain examples meaning changes to beginning part / ending part or left/right. For example, up-stream of the river, down the river, etc.

4.7.7.6 Foreign Words

There are six specific foreign words used in the text. They are: etc., i.e., e.g., de, ibid, and per. Due to minimum overhead to the vocabulary list, all of the foreign words were chosen for enlistment in the learning dictionary.

4.8 Lemma Based Approach

Token based approach reduced the word list and provided insights on the pattern of coverage changes (in percentage) based on number of token to be enlisted in the dictionary. However, though individually few hundreds of each of the token types does not seem much, combined they still put a heavy burden on rural students having very weak English background. To make the overwhelming process more manageable for the students, lemma based approach was considered. This strategy provides benefits for learning noun, verb, adjectives and adverbs only.

4.8.1 Choosing Unique Noun Lemmas

Moving to lemma based approach, 2200 unique noun lemma which reduces the original list of about 2700 unique nouns further by 500. Initially, first 100 unique noun lemmas cover about 40% of the lemmas. About 300 unique noun lemmas cover 60% of the unique noun lemmas. Coverage shows slower growth after words as visible in Figure 4.14.

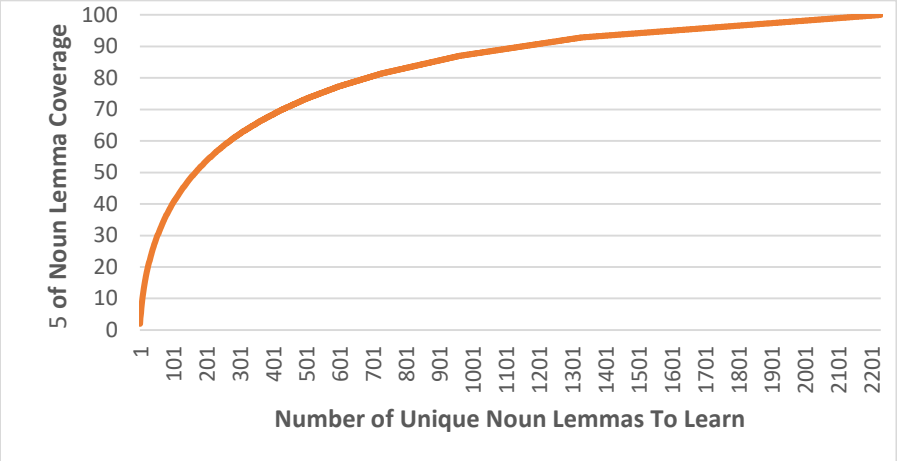


Figure 4.14: Coverage of Unique Noun Lemmas (%)

4.8.2 Choosing Unique Verb Lemmas

Inspecting verb lemmas reduced the list of verbs to almost 50% size. There were about 1400 unique verbs which can be traced back to about 800 unique verb lemma. Covering many of these unique verb lemma will ensure quicker learning of most of the verbs. Covering first 100 unique verbs provides access to about 75% of the verbs. Next 100 verbs raise the coverage to 85%. Covering 300 unique verb lemmas encompasses 90% of the verbs. Verb coverage is illustrated in Figure 4.15.

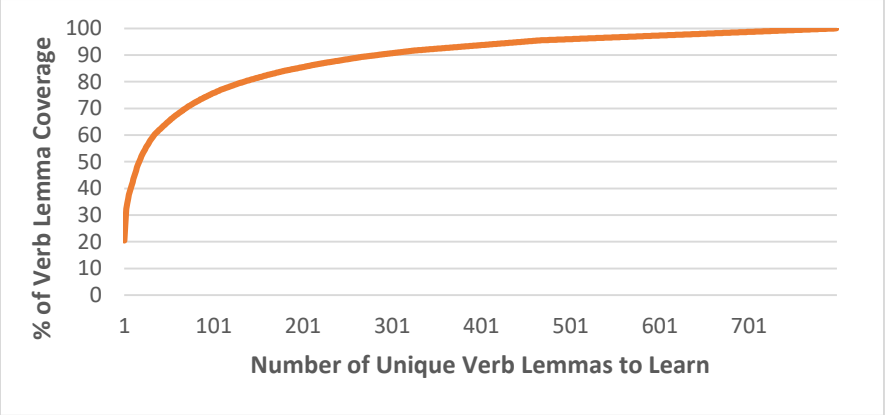


Figure 4.15: Coverage of Unique Verb Lemmas (%)

4.8.3 Choosing Unique Adjective Lemmas

There were about 890 unique adjectives. Lemma based approach reduces the list to about 850 unique adjective lemma. As the reduction is not significant, the learning coverage grows similar to unique adjectives but slightly improved. First 100 unique adjective lemmas cover more than 50% of the unique adjective lemmas. Next 100 unique adjective lemmas cover about 70% of all unique adjective lemmas. 300 unique adjective lemmas cover about 80% of the unique adjective lemmas. Coverage characteristics of unique adjective lemmas is shown in Figure 4.16.

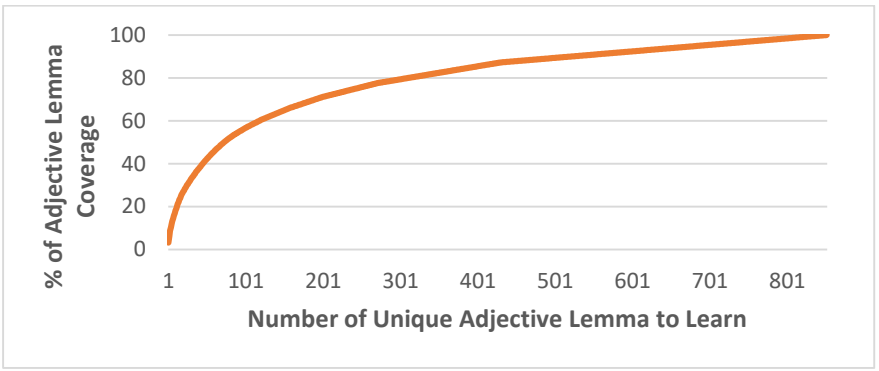


Figure 4.16: Coverage of Unique Adjective Lemmas (%)

4.8.4 Choosing Unique Adverb Lemmas

As in case of adjectives, adverbs show smaller improvement in lemma based approached. However, when students are failing, any improvement is a good improvement to help the students better. There were around 275 unique adverb words. Lemma based approach decreased it by about 3. First 50 unique adverb lemmas cover about 80% of the unique adverb lemmas. Learning 100 unique adverb lemmas leads to 90% coverage of all unique adverb lemmas. Further learning improves it slightly. The enhancement is visualized in Figure 4.17

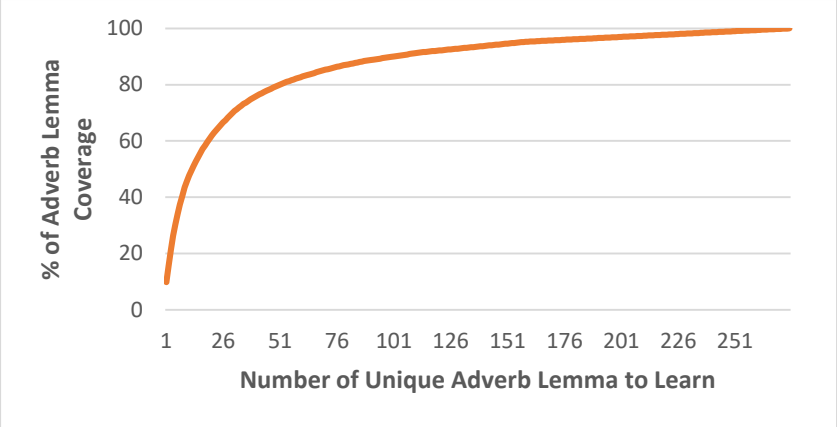


Figure 4.17: Coverage of Unique Adverb Lemmas (%)

4.9 Collocations

Let, make, get, have, these four verbs can replace many other verbs. Hence these introduces great difficulty in translation and understanding. For example, a sentence taken from the text book is “He took a wife to cook meals”. Here the verb lemma “take” is used as the verb “took” but means but it means the verb, “married”. Another example could be, “He is taking a break”. Here “break” does not mean something was “broken” or physical injury. It means momentary stop.

Stanford NLP parser was also evaluated which unfortunately did not work for our purpose to distinguish meaning of clauses. Some snippets from the found collocations are listed in Table 10.

4.10 Condordance

Looking at multiple usage of the same word shows different usages of the same word. It also facilitates discovery of multiple meaning based on usage scenario. For example: “make up” could have three different meaning different meanings. Some examples are shown in Table 9 for the word, “makeup”.

Table 9: Example of Concordance for the word, makeup for Learning Dictionary

Meaning	**Example**
cosmetics	He is selling make up ingredients.
covering/remedial	She applied for make-up examination.
imaginary	They made up stories for the Eid short movie.

Table 10: Example of Some Collocations from Learning Dictionary

Collocated words	Implied Meaning
blood pressure	রক্তচাপ
downy flake	তুষার পাত
first time	প্রথমবার
first true email system	প্রথম পূর্ণাঙ্গ ইমেইল ব্যবস্থা
last week	গত সপ্তাহে
little girl	সামান্য মেয়ে
making money	টাকা কামানো
next morning	পরের দিন সকালে
next time	পরের বার
old age	বার্ধক্য
other hand	অন্য দিকে
sophisticated hindu family	অত্যাধুনিক হিন্দু পরিবার
sort of clean technology	একরকম পরিবেশ বান্ধব প্রযুক্তি
state festival	রাষ্ট্রীয় উৎসব
taking care	দেখা শুনা করা
talk show	আলোচনা অনুষ্ঠান
weekly bazar day	হাট
young man	যুবক

4.11 Sample Vocabulary List

In addition to collocational examples and concordance examples, the dictionary includes word list with Bangla translation pre-seeded by google translate. Some example of nouns are shown in Table 11. List of selected words per parts of speech and token type are included in the learning dictionary.

Table 11: Sample Vocabulary List from the Learning Dictionary

Meaning	Example
shylock	নাছোড়বান্দা মহাজন
utmost	পরম
ventilation	বায়ুচলাচল
vow	ব্রত
wow	কি দারুন
wreath	জয়মাল্য
wreckage	ধ্বংসাবশেষ

4.12 Summary of Results

This work identifies top lemmas of each parts of speech to learn initial seed vocabulary to build on. In the end student will have to learn all new words from the text. However, learning sequence makes a key difference in learning strategy. Key lemmas followed by other lemmas, then key POS tokens followed by other POS tokens will ensure smooth coverage. However, not all words correspond to its dictionary meaning. Hence The dictionary contains collocational meaning as well which identifies group words that have separate meanings compared to the individual words. Finally, concordance shows how same word is used differently in different sentences. Meanings are initially seeded from google translate. It may be later refined using crowd sourcing. In short, derived results of the research include the followings:

- corpus built on words used in English 1st part NCTB text book
- list of words with meanings
- list of collocational words with non-trivial meaning
- list of concordance showing same words with different usages

Chapter 5: Conclusion and Future Work

This research project was to develop a better understanding of the English language skills needed by students at SSC level. The analysis of SSC materials enables to identify specific features of academic English, including the variances, settings, and stages. This will help students to be better prepared for reading the textbook. This may help resource scarce (both human and facilities) rural students to develop better understanding of the materials, will inform and improve the learning experience at SSC level. HSC textbook was briefly looked up but it was found that it will need greater optimization as HSC requires learning even more in less time.

This work tries to extract the academic word list in the domain of English 1st part text book by NCTB. As rural students have very limited vocabulary, they need a smoother learning curve to cope up. Initially, they are given key lemma following by less used other words along with their Bangla meaning. This not only seeds their vocabulary but also will give them some confidence. Then next level is learning more words including variations from lemmas. However, as meaning changes when words are grouped together, they are introduced to list of collocations or group words to learn. Finally, they learn different usage of the same word along sometimes with different meaning. In short, they learn to read the English 1st part book using a dictionary made out of the words in the same book including tricky meanings, group words and learning meaning from concordance examples.

Future Work

Due to time and resource limitations, following areas could not be explored. These areas remain as challenges for the future research work:

- In addition to learning words for English 1st paper, assist students in learning to write complete sentences, paragraphs, essays etc. for English 2nd paper.
- Incorporate essays, letters, applications, paragraphs, etc. to the corpus using OCR.
- Optimize the learning dictionary with good dictionary examples. This is a key area for future investigations to choose and rank example sentences from past lessons and lines in current lessons only.
- After words are presented in different contexts, empower students to visualize the words through usage of reading bookmarks.
- Explore synonyms and antonyms to help compare and contrast words.

References

[1] S. Islam, "SSC Exams Pass Rate Takes Hit for New Evaluation System," *bdnews24*, p. retrieved on 21 August, 2017, 05-May-2017 [Online]. Available: http://bdnews24.com/bangladesh/2017/05/05/ssc-exams-pass-rate-takes-hit-for-new-evaluation-system

[2] W. Bin Habib and T. S. Adhikary, "English, Maths Drag Results Down Again," *The Daily Star*, p. Front Page, 07-May-2018 [Online]. Available: https://www.thedailystar.net/frontpage/ssc-examination-result-2018-bangladesh-english-maths-drag-results-down-again-1572613

[3] Star Report, "Schools of 0, Students fail due to lack of qualified teachers," *The Daily Star*, Bangladesh, p. Front Page, 09-May-2018 [Online]. Available: https://www.thedailystar.net/frontpage/schools-0-1573576

[4] BANBEIS, "Bangladesh Education Statistics 2016," 2017. [Online]. Available: http://banbeis.gov.bd/data/index.php. [Accessed: 01-Jan-2018]

[5] Oxford University Press and British Council Bangladesh, "Launch of Oxford Advanced Learner's Dictionary 9th edition," 2015. [Online]. Available: https://www.britishcouncil.org.bd/en/about/press/launch-"oxford-advanced-learner's-dictionary-9th-edition. [Accessed: 13-Jun-2016]

[6] S. Shahzadi, N, Rabbani, F, and Tasmin, *English for Today Book 7 (for class IX and X)*. Bangladesh: National Curriculum & Textbook Board (NCTB) Text Book, 2010.

[7] OUP, "Pocket Oxford Dictionary, Software Version." Oxford University Press, 1995.

[8] Oxford University Press, "The Man Who Made Dictionaries." [Online]. Available: http://fdslive.oup.com/www.oup.com/elt/general_content/global/man_who_made_dictionaries/. [Accessed: 13-Jun-2017]

[9] Oxford University Press, "Oxford Learner's Dictionary," 2017. [Online]. Available: http://www.oxfordlearnersdictionaries.com/. [Accessed: 01-Mar-2017]

[10] M. O. Hamid, "The linguistic market for English in Bangladesh," *Current Issues in Language Planning*, vol. 17, no. 1, pp. 36–55, 2016.

[11] H. Hansen-Thomas, L. Grosso Richins, K. Kakkar, and C. Okeyo, “I do not feel I am properly trained to help them! Rural teachers’ perceptions of challenges and needs with English-language learners,” *Professional Development in Education*, vol. 42, no. 2, pp. 308–324, 2016.

[12] M. Lamb, “‘Your mum and dad can’t teach you!’: Constraints on agency among rural learners of English in the developing world,” *Journal of Multilingual and Multicultural Development*, vol. 34, no. 1, pp. 14–29, 2013.

[13] J. Sinclair, “Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary,” 1987.

[14] A. Kilgarriff, “Using corpora in language learning: the Sketch Engine,” *Journal of Optimizing the role of language in Technology-Enhanced Learning*, vol. 22, p. 21, 2007 [Online]. Available: https://cdn.uclouvain.be/public/Exports reddot/adri/documents/GRANGER-SYLVIANE-2007.pdf

[15] N. Chomsky, *Aspects of the Theory of Syntax*, vol. 11, no. no 11. 1965.

[16] A. Inumella, A. Kilgarriff, and V. Kovář, “ASSOCIATING COLLOCATIONS WITH DICTIONARY SENSES,” *Proceedings of 6th Biennial Conference of the Asian Association for Lexicography*, 2009.

[17] D. Shin and P. Nation, “Beyond single words: The most frequent collocations in spoken English,” *ELT Journal*, vol. 62, no. 4, pp. 339–348, 2008.

[18] F. Boers and S. Webb, “Teaching and learning collocation in adult second and foreign language learning,” *Language Teaching*, vol. 51, no. 1, pp. 77–89, 2018.

[19] M. Fakharzadeh and B. Mahdavi, “Full Sentence vs. Substitutable Defining Formats: A Study of Translation Equivalents,” *3L: Language, Linguistics, Literature®*, vol. 23, no. 2, 2017.

[20] V. Wiegand, A. Hennessey, C. R. Tench, and M. Mahlberg, “A cookbook of co-occurrence comparison techniques and how they relate to the subtleties in your research question,” in *The 9th International Corpus Linguistics Conference*, 2017 [Online]. Available: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper362.pdf

[21] Y.-J. Lan, N.-S. Chen, Y.-T. Sung, and T.-C. Liu, “Mind and Body Learn Together: Embodied Cognition and Language Learning,” in *2015 IEEE 15th International*

Conference on Advanced Learning Technologies, 2015, pp. 469–471.

[22] F. I. M. C. and R. S. Lockhart, "Levels of processing: A framework for memory research," *Journal of Verbal Learning and Verbal Behavior*, vol. 11, pp. 671–684, 1972.

[23] Euro-Lingual, "EuroLingual Teaching Methods," 2018. [Online]. Available: https://www.euro-lingual.com/english/our-teaching-method/. [Accessed: 01-May-2018]

[24] British Council, A. IDP, and Cambridge, "IELTS, Test taker performance 2016," *Teaching and Research*, 2016. [Online]. Available: https://www.ielts.org/teaching-and-research/test-taker-performance. [Accessed: 01-May-2018]

[25] M. C. McKenna and K. A. D. Stahl, *Assessment for reading instruction*. Guilford Publications, 2015.

[26] H. S. Scarborough, "Connecting Early Language and Literacy to Later Reading (Dis)abilities: Evidence, Theory, and Practice," in *Approaching Difficulties in Literacy Development: Assessment, Pedagogy and Programmes*, 2009, p. 320 [Online]. Available: http://books.google.com/books?hl=en&lr=&id=sfKpsYBGX2MC&pgis=1

[27] National Reading Panel, "Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction," *NIH Publication No. 00-4769*, vol. 7, p. 35, 2000 [Online]. Available: http://www.nichd.nih.gov/publications/nrp/upload/smallbook_pdf.pdf

[28] Oxford University Press, "How many words are there in the English language?," 2018. [Online]. Available: https://en.oxforddictionaries.com/explore/how-many-words-are-there-in-the-english-language. [Accessed: 01-May-2018]

[29] Johnson, "Vocabulary size: Lexical facts," *The Economist, blog*, 2013. [Online]. Available: https://www.economist.com/blogs/johnson/2013/05/vocabulary-size. [Accessed: 01-May-2018]

[30] I. L. Beck, M. G. McKeown, and L. Kucan, *Bringing Words to Life: Robust Vocabulary Instruction (Solving Problems in the Teaching of Literacy)*, 1 edition. Guilford Press, 2002 [Online]. Available: https://www.amazon.co.uk/Bringing-Words-Life-Vocabulary-Instruction/dp/1572307536

[31] I. L. Beck, M. G. McKeown, and L. Kucan, "Choosing words to teach," in *Teaching and Learning Vocabulary: Bringing Research to Practice*, 2005, pp. 211–225.

[32] R. Shams, M. Z. Haider, G. Roy, S. R. Majumder, M. A. Razzaque, and M. S. H. Naina, *English for Today (for class IX and X)*. Bangladesh: National Curriculum & Textbook Board (NCTB) Text Book, 2013.

[33] A. Biemiller, "Vocabulary: Needed if more children are to read well," *Reading Psychology*, vol. 24, no. 3–4. pp. 323–335, 2003.

[34] A. Biemiller, "Words for English-Language Learners," *Tesl Canada Journal*, vol. 29, no. 6, pp. 198–203, 2012.

[35] A. Biemiller, "Which Words Are Worth Teaching?," *Perspectives on Language and Literacy*, vol. 41, no. 3, pp. 9–13, 2015 [Online]. Available: http://search.proquest.com/docview/1717466820?accountid=13552%0Ahttp://primo-direct-apac.hosted.exlibrisgroup.com/openurl/RMITU/RMIT_SERVICES_PAGE??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ%3Aeducation&atitle

[36] A. Biemiller, "Size and sequence in vocabulary development: Implications for choosing words for primary grade vocabulary instruction," in *Teaching and Learning Vocabulary: Bringing Research to Practice*, 2005, pp. 227–248.

[37] A. Coxhead, "A new academic word list," *TESOL Quarterly*, vol. 34, no. 2, pp. 213–238, 2000 [Online]. Available: http://www.jstor.org/stable/3587951

[38] I. S. P. Nation, "How Large a Vocabulary Is Needed for Reading and Listening?," *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, vol. 63, no. 1, pp. 59–81, 2006 [Online]. Available: http://muse.jhu.edu/content/crossref/journals/canadian_modern_language_review/v063/63.1nation.html

[39] I. Nation and D. Beglar, "A vocabulary size test," *The Language Teacher*, vol. 31, no. July, pp. 9–13, 2007.

[40] P. Nation and R. Waring, "Vocabulary Size, Text Coverage and Word Lists," *Vocabulary: Description, Acquisition and Pedagogy*. pp. 6–19, 1997.

[41] P. Nation and A. Coxhead, "Vocabulary size research at Victoria University of Wellington, New Zealand," *Language Teaching*, vol. 47, no. 3, pp. 398–403, 2014.

[42] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[43] B. Santorini, “Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision),” *University of Pennsylvania 3rd Revision 2nd Printing*, vol. 53, no. MS-CIS-90-47, p. 33, 1990 [Online]. Available: http://www.personal.psu.edu/faculty/x/x/xxl13/teaching/sp07/apling597e/resources/Tagset.pdf