Generating Vocabulary Sets for Implicit Language Learning

using Masked Language Modeling

by

Vatricia Edgar

A Thesis Presented in Partial Fulfillment
of the Requirement for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Ajay Bansal, Chair
Ruben Acuña
Alexandra Mehlhase

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Globalization is driving a rapid increase in motivation for learning new languages, with online and mobile language learning applications being an extremely popular method of doing so. Many language learning applications focus almost exclusively on aiding students in acquiring vocabulary, one of the most important elements in achieving fluency in a language. A well-balanced language curriculum must include both explicit vocabulary instruction and implicit vocabulary learning through interaction with authentic language materials. However, most language learning applications focus only on explicit instruction, providing little support for implicit learning. Students require support with implicit vocabulary learning because they need enough context to guess and acquire new words. Traditional techniques aim to teach students enough vocabulary to comprehend the text, thus enabling them to acquire new words. Despite the wide variety of support for vocabulary learning offered by learning applications today, few offer guidance on how to select an optimal vocabulary study set.

This thesis proposes a novel method of student modeling which uses pre-trained masked language models to model a student's reading comprehension abilities and detect words which are required for comprehension of a text. It explores the efficacy of using pre-trained masked language models to model human reading comprehension and presents a vocabulary study set generation pipeline using this method. This pipeline creates vocabulary study sets for explicit language learning that enable comprehension while still leaving some words to be acquired implicitly. Promising results show that masked language modeling can be used to model human comprehension and that the pipeline produces reasonably sized vocabulary study sets.

ACKNOWLEDGEMENTS

I would like to thank Zachary Monroe for the hours of support he gave, conversations we had and ideas we shared. I would also like to thank Ruben Acuña for the many discussions we had and suggestions he made. Finally, I would like to thank Ajay Bansal for our meetings, brainstorming sessions and email chains, and all the hard work he put in to make this possible.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

Today, digital language learning platforms serve a 9.16 billion dollar market (“Online language learning market size, trends, opportunities forecast”, 2019). With globalization increasing interest in language learning, that market is expected to grow to 20.21 billion dollars by 2026, according to Verified Market Research (“Online language learning market size, trends, opportunities forecast”, 2019). Despite the success of digital language learning platforms, language researchers have found the methods of many to be lagging behind modern language instruction practices (Heil *et al.*, 2016). The majority of Mobile Assisted Language Learning (MALL) applications have a focus on vocabulary learning, with many including only vocabulary study (Heil *et al.*, 2016). The emphasis on vocabulary acquisition is not surprising; foreign language learners must acquire a sizable vocabulary. Minimal understanding of authentic texts occurs with a vocabulary size of 4,000 word families (groups of words with similar meaning and form) and an optimal understanding occurring with 8,000 word families known (Laufer and Ravenhorst-Kalovski, 2010). To enable students to gain such a large vocabulary in their target language, a well-balanced language curriculum must include both explicit and implicit vocabulary learning (Nation, 2013). Unfortunately, many mobile vocabulary learning applications focus on explicit learning exclusively or have very weak forms of implicit learning (Heil *et al.*, 2016).

In the context of vocabulary learning, explicit instruction, also called intentional learning, is the deliberate study of words for the sake of vocabulary acquisition (Nation, 2013). Implicit learning, also called incidental learning, is defined by Swanborn and De Glopper (1999, p. 262) as “the incidental, as opposed to intentional, derivation and learning of new word meanings by subjects reading under reading circumstances that are familiar to them.” Often described as “guessing-in-context,” the focus of implicit vocabulary learning is on understanding the material, whether it be a film, a short story, a real conversation, or any other number authentic language contexts, although the focus of research has largely been on learning word in context while reading (Nation, 2013). Learning new words happens almost “on accident” as students guess the meaning of unknown words they encounter so they can understand the overall meaning of the material. Although implicit learning is essential to language learning, only 23 out of 50 of the most popular language apps in 2016 included implicit learning, and many of those only included a context as large as a sentence (Heilman and Eskenazi, 2006). Nation (2013) doesn’t even consider learning words in isolated sentences for the sole purpose of acquiring vocabulary as “true” implicit learning. The lack of digital learning tools for implicit learning is especially disappointing given the difficulty of implicit learning. Students must take care to choose learning material at the appropriate level, to apply appropriate guessing strategies, and to verify their guesses to prevent errors that may become difficult to unlearn Nation (2013). When unknown words do not have enough context to render them guessable, students will not gain much from a reading. The guessability of a word also depends on students’ guessing ability. Students require support to take full advantage of implicit vocabulary learning.

Several techniques have been developed to help students learn vocabulary incidentally, including traditional techniques as well as digital ones. Graded readers are texts written to a specific learning level and can either be original texts written for a particular target language level or modified versions of authentic texts (Nation, 2013).

While useful, graded readers can be problematic because their authors must make assumptions about students at different learning levels. Not all language curricula teach the same vocabulary in the same order and as Nation (2013) points out, despite overlap even the publishers of graded readers disagree on what words belong in what level. This may be especially problematic for self-taught language students who may have a very different vocabulary knowledge than most classroom-based learners.

Another technique to ease implicit language acquisition is glossing, which is the inclusion of some word definitions within the target text (Nation, 2013). Glosses include difficult or essential unknown words from the text, which means glossing may fall into the same trap as graded readers. While still useful, the author of a glossed text does not know which words will be difficult for students to understand, or even which words will be unknown. A solution to the issue of unknown glosses has been addressed in some digital glossing programs by simply providing glosses for any word that students want. Examples include WordChamp.com, a no longer available website which provided a web browser extension allowing students to hover over any word for a translation and add these words to a study bank (Loucky and Tuzi, 2010), and LingQ.com, which provides similar functionality with content from the Web ("LingQ for Schools", n.d.). Although Loucky and Tuzi (2010) found that the use of WordChamp improved reading comprehension, this type of glossing has some troublesome properties. Because this form of glossing allows for unlimited look up of words, it is nearly equivalent to dictionary use, besides being somewhat less disruptive because students do not have to leave the text. It has been shown that up to a third of language learners may use dictionaries excessively, although more research is required to discover why they do so (Nation, 2013). Healy (2018) notes that looking up every unknown word in a dictionary is disruptive and a waste of learning time. She points out that students need to be taught how to determine a word's relevance. While it is clear that dictionary use increases comprehension (Healy, 2018), it is not clear if excessive dictionary use helps or hinders students in learning how to guess in context effectively. More research is required to determine if improper dictionary use is more harmful than just being disruptive. However, given the non-personalized nature of limited glosses and the disruptive and time-wasting nature of unlimited word look ups, it seems that neither form is truly optimal.

Pre-reading activities, in contrast to glossing and graded readers, attempt to modify the student's knowledge, rather than the text, to prepare students for the text and increase comprehension during reading. Pre-reading activities include, among other activities, teaching background knowledge relevant to the text, teaching unknown grammar that is in the text, and teaching unknown vocabulary that is in the text (vocabulary pre-teaching) (Nation, 2013). Vocabulary pre-teaching is focused on vocabulary acquisition, allowing students gain the words required for comprehension. Since vocabulary pre-teaching aims to fill in vocabulary gaps so students can understand a text better, many studies on vocabulary pre-teaching use the same kinds of words as studies on glossing do. Vocabulary pre-teaching then too suffers from a lack of personalization. And unlike some automatic glossing methods discussed, it is not possible for students to choose the vocabulary that must be pre-taught because the point is to teach students the words before they read the text.

All methods discussed also have the issue of availability: Graded readers, limited glossed texts and vocabulary study sets are not available for all stories or books. This reduces the choice that students have in their learning material. A student hoping to read authentic material available in newspapers, blog posts, or more obscure books will likely not be able to find these resources. Graded readers, glossed texts, and texts with provided vocabulary study sets also have the obvious problem of monetary cost. For language instructors hoping to create new material, the cost is of time and effort to create glosses or vocabulary pre-teaching sets. The ability to personalize and automate the creation of these materials would not only decrease costs but also improve student learning. While the automatic creation of graded readers is somewhat of a daunting task given the current state of text generation technology, the creation of vocabulary sets for either glossing or vocabulary pre-teaching may provide more opportunities for automation and adaptation.

Despite the opportunity for automation and adaptation in implicit vocabulary learning, few digital learning applications have taken advantage of this (Heil *et al.*, 2016). Heil *et al.* (2016) suggests that language applications take advantage of new technologies to create more intelligent systems, but this underestimates the difficulty of creating intelligent systems. For example, the vocabulary study website Vocabulary.com adapts to students' skill level to predict what words they need to continue studying, and even includes a tool for generating study sets (Zimmer, 2015). However, the site also claims to have a database of over 100,000 questions with hundreds of millions of responses, along with proprietary technology which uses that data. Data like this is not initially available to new digital learning applications, and to our knowledge, no other digital language learning application has the ability to automatically generate vocabulary study sets for a text. Vocbaulary.com teaches vocabulary via questions, quizzes and games, giving only sentence level of contexts for words, just like many of the applications surveyed by Heil *et al.* (2016). The marked uniformity in methods used by language learning application shows a need for change: either existing applications must incorporate new features or new, innovative applications need to be made to fill this space. Unfortunately, the lack of open data that is suitable for use in machine learning and data science for this task is a major barrier for new applications. Even with large amounts of data, the creators of new applications may not initially have the computational power to process that data.

1.2 Research Goals

In this thesis, we propose an open source pipeline for generating student-specific and text-specific vocabulary study set for use either with vocabulary pre-teaching or with limited glossing. This tool aims to create personalized vocabulary study sets specific to the structure of the text being read and to the vocabulary knowledge of the student studying the text, without requiring large amounts of data about the student or similar students. The pipeline consists of two main parts. First, students are presented with a vocabulary test that estimates not the size of their vocabulary, but the specific words that they know. Second, students enter the target text, and a vocabulary set is produced by modeling the students' ability guess the unknown words in the text. The set of all words that are impossible or very difficult to guess make up the study set. The student is modeled by a language model, which performs the task of predicting words or tokens that are missing from the text.

As either portion of this pipeline represents a significant amount of research and knowledge, the focus of this work is on the latter half of the pipeline. A rough version of the vocabulary test has been implemented to support this work. Both portions of the pipeline are individual tools that could feasibly be integrated into existing software or used to build new software. The tools will be released under an open source license.

This thesis investigates the assumption that impossible or very difficult to guess words are useful for students for use in vocabulary pre-teaching or glossing, and tests efficacy of using language modeling to model the guessing ability of students of foreign languages. It also investigates the use of other language features to support this goal, including the use of lemmas, synonyms, and word similarity. The research questions are as follows:

- Is guessability a good measure of a word's usefulness in vocabulary pre-teaching and glossing vocabulary sets?
- Can language modeling be used to model the predicting ability of language students?
- What types of students are modeled best by existing language models?
- What factors can be used to improve the model for the purpose of modeling student predicting ability?
- What does a vocabulary study set generator that uses language modeling look like?

The first question is answered with an analysis of manual vocabulary selection methods used in vocabulary pre-teaching and glossing research followed by an analysis of automatic methods for performing tasks similar to vocabulary selection. The following three questions are answered by a research study which selects candidate models whose attributes indicate they might reliably model humans students, followed by an analysis which compares cloze tests responses made by the model to cloze (fill-in-the-blank) tests responses made by language students. The final question is answered in the form of a software artifact which uses the selected language model for the production of vocabulary study sets.

1.3 Organization

This work is organized as such: The following chapter outlines both linguistic and computational background knowledge relevant to this research. The next chapter discusses works related to vocabulary estimation and the automatic selection of vocabulary for study. After, we discuss the requirements for vocabulary selection and discuss how language modeling may be useful in fulfilling those requirements. We then discuss the design and implementation of the vocabulary set generation pipeline. This is followed by a description of the methodology for testing the model used by the pipeline and for creating example study sets to examine. The results of both the model testing and of the sample sets created are presented, and finally, we discuss limitations, impact, and future work.

Chapter 2

BACKGROUND

2.1 Linguistic Background

Here, we discuss linguistic research pertinent to this thesis. We explore the benefits of implicit language learning and the conditions under which those benefits can be fully realized. In particular, we investigate the benefits of two support techniques for implicit learning: vocabulary pre-teaching and glossing.

2.1.1 Implicit Vocabulary Acquisition

Implicit learning is an extremely important part of language learning in general and vocabulary acquisition in particular. The vocabulary knowledge gained via implicit learning builds up over time, so that the first time a student encounters an unknown word the some knowledge is gained, and the next time more is learned about the word, until eventually the word is learned (Nation, 2013). For example, Horst *et al.* (1998) found that about 20% of words were learned to some degree. Since these cumulative gains have been relatively neglected in implicit vocabulary research, we do not know for sure how much knowledge is truly gained (Nation, 2013). Although it has been shown that multiple exposures help students learn words (Mart, 2012), we have little idea of exactly how many exposures are required for a word to be learned (Ko, 2012). However, we do know that extensive reading is needed for large gains, with one estimate suggesting that students read one million words a year, which amounts to 10 to 12 novels (Nation, 2013).

Several factors affect whether unknown words in a text can be guessed and acquired. The number of words in the text that are already known to the student is one such factor. A meta-analysis of incidental vocabulary research by Swanborn and De Glopper (1999) showed that 15% of unknown words are learned when 97% of words or more are already known. This research is in accordance with research by Laufer and Ravenhorst-Kalovski (2010) which shows that language learners must know between 95% and 98% of the vocabulary in a target text to have a comprehension that is comparable to that of native speakers. If 95% to 98% of words in a text must be known to provide adequate context for guessing, 20 to 50 running words must be known to guess a single unknown word. This means if unknown words are clustered together, student comprehension of that part of the text may be reduced enough to render the words unguessble. However, the number of running words that are known is not enough to determine if a word can be predicted by students. Research shows that students typically only use context clues that are nearby a word to guess that word; less than 10% of clues for a word come from surrounding sentences (Nation, 2013). If the local context is not context rich, or if the context provides misleading meaning, the students will have trouble correctly guessing the meaning of the word (Çetinavcı, 2014). The student's personal guessing ability is also a major factor in determining whether a word is predictable (Nation, 2013). Students receiving the same text with the same unknown words may guess differently depending on the guessing strategies used and the students' proficiency with those strategies.

2.1.2 Vocabulary Pre-teaching and Vocabulary Glossing

Vocabulary pre-teaching and glossing both attempt to fill in learners' vocabulary knowledge gaps so they can understand the text and guess new words. Vocabulary glossing provides definitions and other information about unknown words in a text, either printed in the margins or at the bottom of the page (Nation, 2013), or digitally, where students can click on or hover over words for a gloss (Loucky and Tuzi, 2010). Glossing has been shown to have positive results on students' comprehension and vocabulary acquisition (Nation, 2013). Since students required 95% to 98% of words to be known, Nation (2013) suggests that some glossing studies may show poor results if the glosses do not supply enough words. Several studies conducted in the last 10 years have shown that vocabulary glossing has a positive effect on reading comprehension, including Duan (2018), Ertürk (2016) Zandieh and Jafarigohar (2012), and Ko (2012). Another study even showed that glossing was more effective than pre-teaching Alessi and Dwyer (2008).

Vocabulary pre-teaching provides students with knowledge of unknown words by teaching some unknown vocabulary before reading. This method can be used to give students the 95% to 98% of words in a text that are required to be known to have native-level comprehension. While intuitively sound, research on vocabulary pre-teaching is inconclusive. Some studies, such as those by Park (2004) and Mousavian and Siahpoosh (2018) have shown clearly positive results of pre-teaching over no preparation. Others show no significant improvement, such as (Jahangard *et al.*, 2012) which showed only a non-statistically significant improvement with vocabulary pre-teaching, and (Alessi and Dwyer, 2008) which showed vocabulary pre-teaching results in an increase in reading speed but not in comprehension.

A meta-analysis of vocabulary instruction and reading comprehension research conducted by Stahl and Fairbanks (1986) suggests that differences in the method of vocabulary instruction (such as the type and qualities of instruction), the setting that instruction occurs in, the structure of texts testing and the vocabulary that is taught may also contribute to the inconsistency. This study considers 149 papers which explore the relationship between vocabulary instruction and reading comprehension, and concludes that vocabulary instruction has significant impact on reading comprehension for texts where the words taught appear. Nation (2013) suggests that vocabulary pre-teaching may require rich instruction and a significant time commitment to be effective. In the study by Mousavian and Siahpoosh (2018), students were taught immediately before reading in 50-minute sessions and spent only ten minutes reading. However, Park (2004) achieved successful results with only five minutes of vocabulary teaching, with students reading four texts of around 200-250 words in length.

Despite inconsistent findings, vocabulary pre-teaching is still commonly found in language curricula. For example, Nisbet (2010) describes guidelines for selecting vocabulary for text comprehension and usefulness in other texts. The study by Shei (2001) notes that textbooks typically include vocabulary and grammar lessons before introducing readings and use this observation as the inspiration to create an application that automatically generates vocabulary lessons for students to study before reading a text. The popularity of vocabulary pre-teaching may be due to the preferences of students. Several studies have shown that students enjoy using this method. In a study comparing vocabulary pre-teaching to another pre-teaching method, students felt that they understood the text better with pre-teaching despite test results showing the opposite (Mihara, 2011). Similarly, research by Chang (2007) shows that pre-teaching for listening comprehension tests did not improve testing results but did improve student confidence. The preference for vocabulary pre-teaching and the evidence that does suggest a positive influence of pre-teaching shows that, like glossing, it is a popular and valuable method of increasing vocabulary knowledge.

2.2 Computational Background

The computational background pertinent to this thesis includes concepts from computer assisted language learning and computational linguistics. Specifically, we discuss students modeling, a technique used to personalize language learning, and language modeling, which creates a statistical model of human language.

2.2.1 Student Modeling

Student models (also called learner or user models) are used in adaptive learning systems and e-tutoring systems to personalize learning materials. Student models consist of a database of user information and a model that acts as an interface between this information and the other models that make up the system (Gamper and Knapp, 2002). Student models may include information about students preferences, learning styles, knowledge level, progress and behaviors (Ennouamani and Mahani, 2019). Students typically self-identify preferences and either self-identify their own knowledge level or are presented with a pre-test to discover their knowledge level (Ennouamani and Mahani, 2019). Applications incorporating student models also test students over time as they use the application to discover weaknesses and strengths, thus allowing the model to be updated (Ennouamani and Mahani, 2019). Information from the model is used to match the users' preferences in terms of content or display as well as adapting those preferences to student knowledge and skill level (Gamper and Knapp, 2002). In a tutoring system, the student model interacts with the domain model (which contains learning materials from the domain) and tutor model (which determines appropriate learning activities) to produce the best tutoring method (Gamper and Knapp, 2002). In adaptive learning systems, the student model, the domain model and the context model (which include information of the user's environmental context, such as location) all inform the adaptive model which chooses the best learning material from the domain model based on information from the student and context models (Ennouamani and Mahani, 2019). An illustration of a basic e-tutoring system is shown in Figure 2.1.

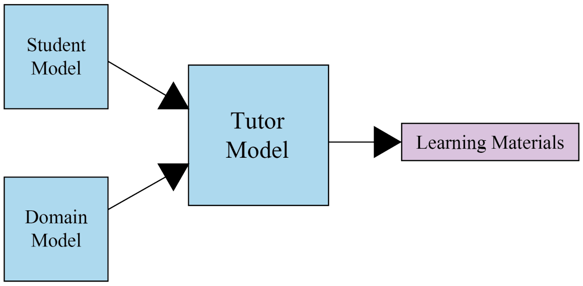


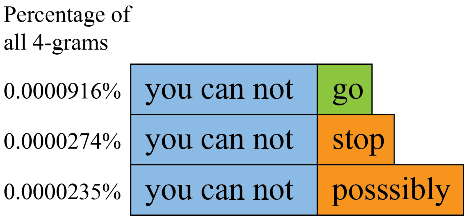
Figure 2.1: E-Tutoring Basic Workflow

Student models are used in a variety of language learning applications. Zaidi *et al.* (2017) applied reinforcement learning to curriculum development in an e-tutoring system. The model guides difficulty level of materials to present to students, so that if the student does not have difficulty with some vocabulary, the model continuously gives the student harder and harder vocabulary until reaching the point where the vocabulary is just barely above the student's level. Websites such as Doulingo and Memrise incorporate spaced repetition that adapts to the student's vocabulary knowledge and present students with a model of their learning progress (Heil *et al.*, 2016). Vocabulary.com collects data on hundreds of millions of students and uses this data to determine what materials students should learn and also uses this data to determine what vocabulary should be studied from a text (Zimmer, 2015). Shei (2001) is an older tutoring application with a much simpler user model: The model consists only of vocabulary that the user knows, as gathered from an initial vocabulary test and subsequent tests as the user interacts with the application. The vocabulary is sorted by frequency and the model is then used to determine what words students should study in a given text based on the frequency value of words in the text.

2.2.2 Language Modeling

Language models are used in a wide variety of Natural Language Processing tasks (Russel *et al.*, 2013). They predict the probability distribution of expressions in natural languages (Russel *et al.*, 2013). Language models encapsulate both semantic and syntactic information about a text (De Mulder *et al.*, 2015). They are developed using corpora (large bodies of natural language text) and calculate the probability distribution over the vocabulary of the corpora they are trained on (Russel *et al.*, 2013). In some language models, linguistic features are manually encoded, but machine learning based models learn linguistic features such as word frequency, grammatical structure, text length, and semantics (De Mulder *et al.*, 2015). Language models are often applied to the task of symbol or expression prediction, that is, given a sequence of language symbols or expressions (i.e. a history of sequences), predict next symbol or expression in the sequence (De Mulder *et al.*, 2015). It is common for these tasks to specifically predict words given a sequence of words (De Mulder *et al.*, 2015). They do this by determining the highest probability word given the target sequence of words (De Mulder *et al.*, 2015). One of the earliest techniques is the n-gram model (De Mulder *et al.*, 2015). N-grams are N length sequences of symbols or expressions, and in language modeling the symbols or expressions in an n-gram may be characters, words, sentences, etc (Russel *et al.*, 2013). N-gram models predict expressions by determining the the number of times that the n-gram consisting of the target sequence and an expression appear in the corpus and dividing that by the number of time the (n-1)-gram, consisting of the target sequence alone, appears (De Mulder *et al.*, 2015). Figure 2.2 shows an example using Google N-gram viewer. Here, we see the percentage of 4-grams in the corpus which match each of the example 4-grams.

Considering only this set of 4-grams, the word “go” would be predicted as the most likely successor of “you can not” since it appears more than the others.



Percentage of all 4-grams		
0.0000916%	you can not	go
0.0000274%	you can not	stop
0.0000235%	you can not	posssibly

Figure 2.2: 4-gram Example Using Google N-gram

More advanced techniques including the use of neural networks and their derivatives, such recurrent neural networks (RNNs) and long short-term memory (LSTM) RNNs. These machine learning techniques learn the probability distribution (De Mulder *et al.*, 2015). Neural network language models have been shown to encode important linguistic features including syntax and semantics, removing the need for linguistic features to be manually programmed (De Mulder *et al.*, 2015). The basic workflow of neural network based language models is presented in figure 2.3

Natural language involves long-distance dependencies; knowing only the end of a sentence, paragraph or other unit of language is not sufficient for understanding the meaning of the sentence. Recurrent neural network models allow for a theoretically unlimited history for predicting the next word in a sequence, but in practice long-distance context is not taken into account because of the vanishing gradient problem, which causes the influence from earlier inputs to become to small to matter (De Mulder *et al.*, 2015). LSTM models partially solve this problem by allowing unaltered information from previous inputs to flow to the current input (De Mulder *et al.*, 2015). Attention mechanisms have also been applied to this problem. Attention refers to the model’s focus on portions of the input (Bahdanau *et al.*, 2014). Models can be trained to only “pay attention” to the most relevant parts of an input sequence, which results in models that perform better for longer sentence lengths, as is the case for the machine translation models discussed in Bahdanau *et al.* (2014). While attention mechanisms were initially applied as an addition to RNN and LSTM models, a new type of model, called the transformer, using only attention was introduced by (Vaswani *et al.*, 2017). Although (Vaswani *et al.*, 2017) uses transformers for machine translation, transformers have also been applied to language modeling specifically (Radford *et al.*, 2018), (Devlin *et al.*, 2018). Radford *et al.* (2018) and (Devlin *et al.*, 2018) both developed language models for pre-training. The language models are not intended to be used as language models, but rather intended to be fine-tuned for specific natural language understanding tasks such as question and answer and classification. Despite excellent performance on NLU tasks after fine-tuning, these models have sub-par performance as language models alone, likely because important sequential information is lost when using attention alone (Wang *et al.*, 2019).

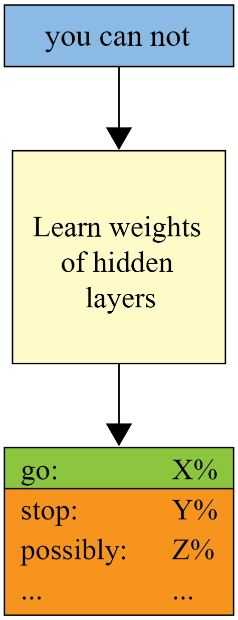


Figure 2.3: RNN Basic Workflow

While the traditional language modeling task is to predict the next expression when given a history of expression, bi-directional language models predict an expression when given a context (De Mulder *et al.*, 2015). Unlike a history, which we will define as the sequence of expressions preceding the target expression (the left-context in languages read left to right), a complete context includes both past and future sequences, the left and right contexts. Bi-directional language models more accurately predict the probability distribution of a language since they receive more information about the vocabulary (De Mulder *et al.*, 2015). An early bi-directional language modeling technique used two neural networks, one using the left context and the other user the right, and combined their results to create the final output (De Mulder *et al.*, 2015). Truly bi-directional RNNs and LSTMs use additional hidden layers which receive input from the output layer (De Mulder *et al.*, 2015). Essentially, once the right side of a word begins to be processed, that information is fed back into the processing of that word. Bi-directional transformers were introduced by Devlin *et al.* (2018) and are shown to outperform uni-directional transformers on fine-tuned tasks. Bi-directional transformers are distinct from bi-directional RNNs and LSTMMs because the right and left side of a word is processed at the same time, instead of using the hidden output layer technique Devlin *et al.* (2018).

Chapter 3

RELATED WORK

3.1 Vocabulary Estimation

Here, we discuss works related to vocabulary estimation, the first step in the vocabulary set generation pipeline. We also discuss issues related to the estimation of vocabulary knowledge and the difficulties observed by CALL researchers on the topic of vocabulary estimation for student modelling.

3.1.1 Traditional Vocabulary Estimation Tests

There are several features of student vocabulary knowledge that may be tested, such as students vocabulary breadth (how many words are known) or depth (how well words are known) as well as measures of receptive (recognizing language) or productive (producing language) knowledge (Pignot-Shahov, 2012). For the purpose of understanding students' ability to read, productive measures are not necessary. However, the measure of vocabulary depth may be important, as depth includes important information about words including use in idioms and collocation (Pignot-Shahov, 2012). Unfortunately, studies on measuring vocabulary depth knowledge are inconclusive and use varying definitions, with some focusing on "layers" of knowledge, building on the fact that vocabulary knowledge is cumulative, while others break word knowledge into components such as spelling and collocation and test on the individual components (Pignot-Shahov, 2012). Vocabulary breadth tests aim to measure the amount of words rather than the amount of knowledge for each word. According to Pignot-Shahov (2012), vocabulary breadth tests are usually either dictionary based, where words on the test are selected from a dictionary, or frequency based, where words are selected based on their frequency in a language corpus. Breadth knowledge is also useful for for reading as 95% to 98% of word in a text are required for adequate understand.

For the purpose of this study, we will only consider receptive breadth tests. Depth tests require a certain level of attention to the presentation and test type, such as having different types of questions to test different layers or components. Since we aim to make a flexible tool that can be integrated into any application, regardless of presentation, depth tests cannot be considered. In a breadth test, the only thing that must be tested is if the student ”knows” a word, so developers using the tool are free to choose how to test that knowledge. In addition to only considering receptive breadth vocabulary tests, we will only consider frequency-based breadth testing approaches. This is because frequency-based approaches are more common for language students Pignot-Shahov (2012). Another useful distinction is that between vocabulary diagnostic tests, vocabulary size estimation and estimation of the specific words in a learner's vocabulary. While vocabulary diagnostics aims to approximate students' vocabulary levels and identify gaps, vocabulary size estimation attempts to estimate the exact number of words known by students. The estimation of specific words takes estimation another step further by attempting to pin down the exact words that students know. Many tests only aim to test general vocabulary knowledge or size, but some CALL systems require a specific list to model student knowledge. We discuss traditional vocabulary tests which estimate at the first two levels.

The Vocabulary Levels Test is a diagnostic vocabulary test that aims to place students in vocabulary levels that represent their vocabulary knowledge (Kremmel and Schmitt, 2018). The test uses a word family frequency list to bucket students into levels (Kremmel and Schmitt, 2018). Each level, levels 1 through 4, is represented by the word family frequency bands of 2,000, 3,000, 5,000 and 10,000 (Kremmel and Schmitt, 2018). Students are tested on 30 items within each level to determine the level they are at. This test has been widely used to diagnose vocabulary weaknesses at certain vocabulary bands as well as to quickly measure students' abilities for language level placement (Kremmel and Schmitt, 2018). The frequency levels represent broad vocabulary knowledge levels and highlight gaps in student knowledge. Research on the test has shown that it has high reliability for this purpose (Kremmel and Schmitt, 2018). Studies using the test have also provided evidence for a "fact" about vocabulary knowledge that was previously only backed by intuition: If students know words at a lower frequency, they likely know many words in higher frequency bands (Kremmel and Schmitt, 2018). While the Vocabulary Levels Test does not aim to measure vocabulary size, nor specific vocabulary knowledge, this fact has implications for tests that do attempt to test either.

Meara (1992) presents the so-called yes/no test which is another diagnostic test which provides instructors with a general profile of student knowledge and can be used to monitor program over time. The question format heavily simplifies the process of testing by simply asking students whether they know the words on the test. As Meara (1992) states, this is off course flawed for two reasons: Students may not be honest about words they know, and students, like researchers, may have different definitions of "knowing" a word. Meara (1992) attempts to mitigate these issues by including fake words. The number of fake words that the student indicates as being known is used to deduce the actual number of words known. A table included in (Meara, 1992) shows how to convert the number of "known" responses and the number of false "known" responses to a vocabulary score. Despite this obvious disadvantage, the test has the advantages of allowing students to finish it fast and being able to be automatically scored.

The Vocabulary Size Test is an estimation of student vocabulary size introduced by Beglar and Nation (2007). Like the Vocabulary Levels Test, it uses a word family frequency list for its testing material. However, a key difference between the two tests is that the Vocabulary Size Test uses frequency bands of size 1000, so that the first frequency band represents the first 1000 words, the second the next 1000 words, and so on. This allows for a higher level of granularity, which is required for the purpose of testing vocabulary size as opposed to knowledge level. Each level tests a 1000-word frequency band with ten multiple choice questions, each assessing students' knowledge of the meaning of a single word. Since ten questions are used for each band, every question represents 100 word families. To estimate a student's vocabulary level, simply multiply the number of correct answers by 100. This method is admittedly somewhat flawed, as Beglar and Nation (2007) states, a student who tested well for the first 3000 words only may have some gaps in that knowledge and may know some words above that frequency. However, the test has been validated and shown to be a useful tool for researchers and instructors (Beglar, 2010).

3.1.2 Digital Vocabulary Estimation Tests

The first step in some CALL systems is a vocabulary test which estimates how much vocabulary a student has and what words are in that vocabulary. This differs from the requirements of the tests discussed above which only aim to test either knowledge levels or vocabulary size. Instead, some CALL systems such as adaptive learning systems and tutoring systems must know the exact words that are in a student's vocabulary so they can accurately adapt to the student's needs. Here, we explore two automatic vocabulary tests, one of which is an online vocabulary test which is not part of a tutoring system and the other of which is part of a language tutoring system. We also discuss several issues that have been observed in vocabulary testing.

TestYourVocab.com presents a two-part vocabulary test to users along with a survey to collect data for linguistic research (“Test your vocab - how many words do you know?”, n.d.). The test is a frequency-based test like the others discussed here, except the words are distributed on a logarithmic scale based on frequency, as opposed to the linear scale used in the Vocabulary Size Test. The words are also randomly selected from the frequency range being tested. In the first part of the test, the user is presented with a test that represents a wide variety of frequency bands. The user must simply respond if they know a word or not, like the yes/no test but without fake words. The first test narrows down the vocabulary list to the specific area of frequency where the user is estimated to be at. The next test narrows down the frequency level of the student even further to give the final estimated vocabulary size. The test aims to find the midpoint of the user’s vocabulary by finding the midpoint of the incorrect answers on the test. Given a word the user does know, if the number of words at lower frequencies that were marked as unknown is equal to the number of words at higher frequencies marked as unknown, the given word is the user’s vocabulary midpoint. The test has a margin of error of 10%.

In FollowYou!: An Automatic Language Lesson Generation System (Shei, 2001), before students can use the system for study, they must complete a vocabulary assessment. The results of this assessment are then used to generate lessons for the student, as described later in this paper. The test begins by asking students if they identify as a beginning, intermediate, or advanced student. The student is then presented with a cloze (fill in the blank) question test consisting of 50 cloze questions embedded in paragraphs. The target vocabulary in these questions ranges from easy (high frequency) to hard (low frequency) words. If the student consistently gets words below a certain frequency correct, but consistently gets words above that frequency wrong, then that frequency is marked as the student's vocabulary level. Words below that level are considered known to the student and those above as unknown. The system then uses this knowledge to generate vocabulary lessons when the student later provides a target text.

3.2 Automatic Vocabulary Selection

In this section, we discuss the automatic vocabulary selection methods of two tools: the modern web application Vocabulary.com and an older project, which is still similar to that presented in this thesis, FollowYou!.

3.2.1 Vocabulary.com

Vocabulary.com offers many services, such as a comprehensive dictionary, premade vocabulary study sets for a variety of texts and topics, the ability for teachers and students to create their own vocabulary study sets, and gamified study methods which include sentence level context for words (Zimmer, 2015). The website also offers automatic vocabulary set generation, which creates sets for use within the website. Vocabulary.com attributes the source of their success to the millions of responses they have collected for over 100,000 questions. They plot the number of correct and incorrect answers against student levels and determine if a question is a good way of determining student level by observing the steepness of the curve. These questions are used to determine the level students are at. Vocabulary.com is a popular tool and instructors have reported that it increases vocabulary knowledge, reading comprehension, and student engagement. However, as previously suggested, the use of large data sets of student questions and answers prohibits new applications from using this method initially. Since the website judges vocabulary difficulty based on words presented in isolation or in isolated sentences, there may also be a missed opportunity to consider the factors that affect reading comprehension over the course of a paragraph or an entire document. This missed opportunity is the exact kind of problem that could be solved if more accessible methods for generating vocabulary were available

3.2.2 FollowYou!

FollowYou!:An Automatic Language Lesson Generation System is a project whose goals and methodology are perhaps the closest to the project presented in this thesis compared to any other project discussed here (Shei, 2001). The goal of FollowYou! is to turn any authentic text given by the user into a textbook like vocabulary lesson. The intention is to bridge the gap between the high support, low personalization world of a textbook, and the low support, high personalization world of authentic text. FollowYou! begins with a vocabulary test that estimates what words students know and stores these words. This forms the basis of the student model. When given a text, FollowYou! identifies content words in the text that are unknown to the user and filters them by their frequency rate. Words that are at or slightly above the student's vocabulary level (as measured by the vocabulary test) are selected for presentation in the textbook lesson. The choice to use only words that are at or slightly above the student's level was inspired by comprehensible input theory, which suggests that texts students read should be slightly above their level. The paper gives no insight into what happens when most of the words in the given text are far above the student's level. After the vocabulary is selected, students receive a textbook like vocabulary lesson which includes each word's definition, collocations, synonyms, and an example sentence. After the lesson, students read the text. While reading the text, students are free to refer to the vocabulary lesson and can select more words to be included in the lesson. Students then take a post-test in the form of a cloze text to verify knowledge of the words. All words that have been successfully learned are added to the student's stored vocabulary and over time the student's vocabulary level is raised as more and more words are learned.

The most relevant part of the project is in the selection of vocabulary. The application begins with a list of words in the text which are unknown to the user and filters them based on the comprehensible input theory. The researchers developing FollowYou! took comprehensible input theory to mean that words studied in the text should be at or above the student's level, but the project did not discuss pre-teaching lesson support for words far above the student's level. If a student encounters a word that is far above their level in the text, they are free to add it to the list of words to study. The selection of vocabulary is not on what words will make the text more understandable but rather on what order the words should be learned, if it is assumed that words should be learned in order of frequency.

Chapter 4

REQUIREMENTS FOR VOCABULARY SETS

4.1 Manual Vocabulary Selection

This section discusses vocabulary selection methods used by researchers studying glossing and vocabulary pre-teaching. Manual selection methods may provide useful insights into the qualities words should possess to be included on lists for these methods.

4.1.1 Selection Methods of Pre-teaching and Glossing Research

The words in vocabulary lists made for vocabulary pre-teaching and glossing usually follow some sort of criteria. The criteria may include requiring that the words be unknown to the students, requiring that the words be difficult for the students, requiring that the words be essential to understanding the text. However, what is actually meant be "difficult" or "essential," is not always clearly defined by the researchers. Additionally, the requirements used to create vocabulary sets vary by the goals of the research, which may measure overall text comprehension, vocabulary inference ability or the successful acquisition and retention of new vocabulary. Despite these differing goals, it is still useful to survey the criteria used to create vocabulary lists for pre-teaching and glossing to guide the development of the vocabulary set generator. Table 4.1 summarizes some of the methods used to manually select vocabulary lists for pre-teaching and glossing.

Methods used in studies include word lists, frequency lists, expert opinions, and pilot studies. Word lists are created for general studying purposes, such as an aca-

Table 4.1: Selection Methods Used in Vocabulary Studies

	Unknown Words				Difficult Words			Essential Words	
	WL*	FL**	EO***	PS****	WL*	EO***	PS****	EO***	PS****
(Park, 2004)					✓	✓		✓	
(Chang, 2007)						✓	✓		
(Alessi and Dwyer, 2008)						✓		✓	
(Mihara, 2011)								✓	
(Ko, 2012)				✓			✓		
(Zandieh and Jafarigohar, 2012)				✓					
(Gan, 2014)				✓					
(Ertürk, 2016)				✓			✓		
(Kim and Cha, 2017)			✓			✓		✓	
(Duan, 2018)	✓	✓	✓	✓					✓

*WL - Word List, **FL - Frequency List, ***EO - Expert Opinion, ****PS - Pilot Study

demic word list containing vocabulary that is common in an academic setting, or for particular topics or uses, like a medical word list containing common medical terminology. Word lists are used in glossing or pre-teaching studies to discover words that students might already know so they can be avoided. This is the case in (Park, 2004), where words commonly found in the word lists of Japanese high school English language textbooks were avoided for pre-teaching because these words were likely to be known by the Japanese students. It is also the case in (Duan, 2018) which avoids words used in Lexicon for English Majors - A Supplement to the English Curriculum by the Shanghai Foreign Language Education Press (2001).

Frequency lists are used in a similar fashion– high frequency words are considered to be easy or are assumed to be known and are therefore avoided. Duan (2018) also used this approach to ensure that the target words selected for glossing are not known to students in the study. Nisbet (2010) provides guidelines for instructors which runs somewhat counter to the approach of teaching lower frequency words. Instead, it is suggested that pre-teaching include only low frequency words that are essential for comprehending the text. However, teaching medium frequency words is encourage because they are likely to be unfamiliar to students, but also more likely to be useful outside of the text .

The use of expert opinions for selecting vocabulary sets has the potential to be a bit more personalized than word lists or frequency lists. Here, we define "expert opinions" based pre-teaching or glossing lists to include any hand selected vocabulary list created for a specific text or specific set of students. In several studies, the researchers asked language instructors to create study sets or created the study sets themselves (Mihara, 2011), (Alessi and Dwyer, 2008), (Kim and Cha, 2017), (Duan, 2018), (Park, 2004), (Chang, 2007). One study included words which were included with the text being read and thus were chosen by the creator of the learning material (Kim and Cha, 2017). The instructors chosen to create the vocabulary sets in some of the studies were either instructors for the students participating in the study or instructors at the same university or school as the students (Kim and Cha, 2017), (Alessi and Dwyer, 2008). These instructors were therefore familiar with the student's knowledge either directly through instruction or indirectly through familiarity with the curriculum the students were being taught. This leads to a more accurate representation of unknown or difficult words. Experts were also sometimes asked to select words that were essential to the text. The studies using the essential word selection criteria were therefore also adapted to the text itself.

The most personalized and accurate method may be the pilot study, in which students in a group statistically similar to the participants of the main study are tested on vocabulary knowledge. Participants in a pilot may be exposed to the text used in the main study and asked to identify unknown words. This method gives researchers an exact idea of which words are unknown (Gan, 2014) Zandieh and

Jafarigohar (2012), (Ko, 2012), (Duan, 2018), Ertürk (2016). Pilot studies are also conducted to discover which words will be difficult for students, such as in the studies by Ko (2012) and Ertürk (2016) which asked students to attempt to guess unknown words as part of their respective studies, as well as (Chang, 2007) which conducted a pilot study to measure difficulty of words for the vocabulary list. Duan (2018) also used a pilot study with language students to determine if the words in the text were essential (Alessi and Dwyer, 2008), (Kim and Cha, 2017), (Park, 2004).

4.1.2 Extracting Requirements for Vocabulary Sets

We have seen that the criteria for words in vocabulary lists include being unknown, difficult, essential, or some combination of these. However, several of the methods for determining if vocabulary meets these requirements are problematic for automatic and adaptable vocabulary set generation. We will now analyze the methods and requirements discussed to discover those with the most potential for automation and adaptability.

Studies which used expert opinions to select vocabulary lists do not detail what methods or detailed criteria the experts used in their selection. At the highest level of detail, the studies explored in this research simply stated that the experts chose words they believed were likely to be unknown to or difficult for the students or were essential to the text. The criteria used to determine if words fall into any of these categories is undefined.

Word lists offer a more feasible option for automation, as word lists for language study are available and could feasibly be used to automatically select vocabulary for study. However, the word lists in the works surveyed were selected because the participants in the study were likely to be familiar with the word list itself. This raises questions about the adaptability of the word list method. A word list may not indicate easy words if it is not known whether the students are familiar with the word list.

Frequency lists offer another method that has automation potential. It is important to note that, as stated by Beglar and Nation (2007), frequency is not the only indicator of vocabulary knowledge. However, the validation of the Vocabulary Levels Test and Vocabulary Size Test, which both use word frequency, shows that word frequency is a strong indicator of word knowledge, even if it is not the only indicator (Beglar, 2010). Therefore, frequency may be a useful tool for discovering unknown words.

The final method to consider is the pilot study. Pilot studies were used to discover unknown, difficult, and essential words. At first glance, this method of measuring the criteria may not appear to have much use for automatic vocabulary set generation. Conducting true ”pilot studies” on users to determine good study words for similar users would likely not be accepted. Gathering user data to predict unknown, difficult, or essential words for similar users requires building a large user base, which is not helpful for new applications. Conducting a pilot study on the user reading the text to gather study words would defeat the purpose of the pipeline, as we want students to study vocabulary before reading or use glossing to aid reading. However, the method used by Ko (2012) and Ertürk (2016), who cites Ko (2012) as the source of their method, provides a useful insight for determining difficulty. Interestingly, these studies ask participants to not only identify unknown words but also to guess unknown words. Words that are not known or correctly guessed by the participants in the studied are identified as likely to cause difficulty in reading and are added to the vocabulary list. This method of vocabulary selection supports the intuitive idea, based on studies in implicit vocabulary acquisition, that guessable words do not need to be pre-taught and should be left for implicit acquisition during reading.

From this analysis of criteria for vocabulary sets and methods for determining if words meet those criteria, we have discovered two metrics for measuring if a word is suitable for a vocabulary set. These metrics include frequency, which can be used to determine if a word is unknown, and guessability, which can be used to determine if a word is difficult. This thesis uses frequency in the vocabulary test portion of the vocabulary test set pipeline to determine if words are unknown. Guessability is related to both the context provided by the text and student guessing ability. These do have the ability to be modelled automatically, as is discussed in the next section. While automatically determining if a word is essential may be possible, it was not investigated. The criteria of "unknown" and "difficult" require modeling of student knowledge and ability, while "essential" are mostly dependent on the text itself. This work focuses on student modeling, and thus did not consider purely text based metrics.

4.2 Automatic Vocabulary Selection

This section discusses automatic methods of detecting the guessability of a word in a text. This includes a discussion on cloze (fill in the blank) tests, which must have carefully chosen, guessable blanks, and language modeling, which has been used to determine guessability in cloze test studies.

4.2.1 Automatic Cloze Test Generation

An extremely common area of interest in CALL research is the automatic generation of cloze questions. Tests made of these questions are useful for testing students' vocabulary knowledge, reading comprehension skills and word prediction skills. While the task of cloze test generation may not seem directly related to the task of vocabulary selection for pre-teaching and glossing, the tasks have several similarities. These similarities stem from the selection of "target words." In a cloze test, a target word is the word in a cloze question which will be replaced by a blank for the student to fill in or select from a list of distractors. In vocabulary selection, a target word is a word in a text which is chosen to be included in a list for glossing or pre-teaching. The target words for either task are selected inversely. In a cloze test, the word selected must have enough context in the surrounding text to allow the student to correctly predict the answer. In vocabulary pre-teaching and glossing, the goal is to find unknown words that are not possible or extremely difficult to guess because they lack context. Thus, research in automatic cloze test generation, specifically concerning the choice of target words, provides information on what makes a word predictable and how this can be automatically determined. This information can then be used to inform the process of filtering out predictable words from the set of unknown words in a text, thus producing a set of unpredictable words for study. Unfortunately, much of the focus of cloze test research is on the choice of distractors for use in multiple choice cloze tests, as opposed to research on target word selection. Here, we explore three papers which focus on the selection or evaluation of target words in multiple choice or open cloze tests. A summary of the methods used in these works is presented in table 4.2.

(Hill and Simha, 2016) proposed a context-based method of generating cloze tests based on input text given by the user. The method aims at producing closed (multiple choice) cloze questions which test reading comprehension more than vocabulary or factual knowledge, which is why there is a heavy focus on context. The paper makes a distinction between the full context of a text and the narrow context of a smaller group of words, two to five words in length. According to the paper, tests that truly assess a reader's comprehension should chose target words which are obviously correct given the full context of the text, and distractors which are equally plausible given a narrow context. Named entities and function words (words which serve a grammatical purpose but do not provide content information) were not considered for selection as target words. Locally frequent words are removed from the pool as well under the assumption that they are likely easy to guess, but words that are globally frequent across the language are kept because they may still be difficult to guess in the specific context. Finally, word co-occurrence is used to determine which words have the most context available in the full scope of the text. The paper assumes that words that occur frequently together provide context for each other, since words that co-occur frequently are frequently found in each other's contexts. To find suitable distractors, similar words are compared with Google N-gram. Using n-grams of size two through five, the distractor with the most common n-gram is chosen, as this word appears in the local context most frequently. This method resulted in an average of more than 90% of target words fitting the full and narrow contexts, meaning most target words should be easily guessable in both contexts by students with good reading comprehension skills. The method was good at removing distractors that are plausible in the full context (only 11.6% plausible distractors were missed). The method had the best results in choosing words that fit the narrow context when using larger n-grams, with an accuracy of about 74%.

Felice and Buttery (2019) propose a method of evaluating open cloze tests using entropy to model the restrictiveness of the context provided around the target word. The context surrounding gap words in open cloze tests requires more careful evaluation than closed cloze tests because students do not have a limited selection of words to choose from. If there is not enough context to limit choices, it likely that students may respond with unexpected answers that are not only syntactically correct, but semantically correct as well. Felice and Buttery (2019) use the number of possible syntactically and semantically correct choices for a gap and the probability of those choices to measure the restrictiveness of the context provided by questions.

The number of choices and their probability is modeled with entropy, "which quantifies the amount of information conveyed by an event" (Felice and Buttery, 2019, p. 324). Specifically, Felice and Buttery (2019) use Shannon's Entropy, shown in equation 4.1, where P is the probability of an event (word choice) occurring. Questions with more options for responses and higher probability for those responses will have higher entropy and therefore should be more complex. Felice and Buttery (2019) measured the entropy of open cloze tests from Cambridge English examinations, which were used as the gold standard of cloze tests since they were handpicked by experts in language teaching and testing. A 5-gram bidirectional language model was used to measure entropy. The results of this test show that entropy generally correlates to difficulty level; higher level questions in the Cambridge English examinations had higher entropy while low level questions had lower entropy.

$$H = -\sum_{i=1}^{M} P_i \, log_2 \, P_i \tag{4.1}$$

The final paper on cloze test evaluation is Beinborn *et al.* (2015), which predicts the difficulty of closed cloze tests as well as C-tests and pre-fix deletion tests. All of these tests are reduced redundancy tests, where the language redundancy in a text is reduced by deleting portions of the text, thereby testing students' language proficiency. These tests choose target words which are deleted in whole or in part, where cloze tests delete the entirety of a target word, and pre-fix deletion tests and C-tests remove the first and second half of each target word, respectively. The latter two tests are open tests, while the cloze test used in this research was closed. The possible answers for the open tests were limited in length. The researchers predicted test difficulty with classification and regression models, originally introduced in (Beinborn *et al.*, 2014b) and used only for C-Tests in the initial paper. They extract features in three ways. The first set of features, a super set of which was initially used in

(Beinborn *et al.*, 2014b), include properties of the solution, properties of the text, and properties of the question. Solution properties describe the target word and its immediate context (the words before and after the gap). These include frequency, word class, inflection, spelling and others. Text properties refer to properties of the entire sentence or paragraph of the question and include readability measures such as number of clauses, average word and sentence length, and others. Test properties refer to properties of the overall test, including the number of answer candidates and the position of gap. Additional features introduced in Beinborn *et al.* (2015) are the ability for language modeling and semantic relatedness to predict the correct answer. The language modelling method uses a 5-gram statistical language model to predict answers by calculating all possible answers and selecting the most probable resulting sentence. Semantic relatedness is only used for cloze tests since it is most useful with complete content words. With this approach, the similarity of all possible candidates to all content words in the question sentence is calculated by finding the cosine similarity between each word's word vector (vector representations of words obtained by training a model). Candidates that are most similar to other words are considered the most likely fit and are selected as the answer. Beinborn *et al.* (2015) found that the language model and semantic relatedness methods to solving closed cloze questions were far worse than students' ability to answer the same questions. However, the use of language modelling and semantic relatedness as features significantly improved the regression model's ability to predict text difficulty over the features derived from other properties of the question. This suggests that, while the language model and semantic relatedness methods used by this study do not model student guessing ability perfectly, the ability for the language model or semantic relatedness methods to answer questions correctly does correlate with students' ability to answer questions correctly

Table 4.2: Automatic Target Selection and Evaluation Strategies for Cloze Tests

Evaluated Item	Evaluation Type		
Word	Local frequency	Global Frequency	
Candidates	Candidate space size	Candidate probability	
Context	Word co-occurrence	N-gram probability	Semantic relatedness
Text	Readability		
Other/Multiple	Language modeling		

4.2.2 Language Modeling as Student Modeling

The use of language modeling in these works is especially interesting. Felice and Buttery (2019) use language modeling to measure both candidate space and candidate probability for use in calculating entropy, which models the quality of the context provided by the text. Beinborn *et al.* (2015) uses language modeling directly to predict difficulty of a text: if the language could not predict a word this correlated with a higher percentage of students being unable to predict the word. This correlation shows that language models have the potential to model not only available context but also student guessing ability. Beinborn *et al.* (2015) also used semantic relatedness to account for the language model's inability to deal with important context that is larger than 5-grams. Some advanced language modeling techniques can account for more context, which makes up for this deficiency. Hill and Simha (2016) did not use actual language modeling, but did use a method similar to n-gram statistical language modeling, as counting the number of times an n-gram appears in Google N-grams does approximate the probability of that n-gram. Although a true model with an approximate distribution of probability over a vocabulary was not created, the approach is similar in principle.

It is not surprising that language models are used to model context and complexity in these work as language models encapsulate both semantic and syntactic information about a text. Linguistic features such as word frequency, grammatical structure, text length, and semantics can be captured by a language model alone. The language models used by these cloze test studies have the advantage of removing the need for manual encoding of linguistic features, which makes it much easier to adapt applications using language models to many languages. For these reasons, we chose language modeling as the tool to determine the difficulty of unknown words in a text.

Previous discussion of language modeling has shown that language models are created in many different ways, from the earliest n-gram models, to RNNs and LSTMs, to the modern Transformer architecture. Now that we have decided to use language modeling to model student vocabulary acquisition, the next question is "what type of language model?" To choose the proper type of model, we examine what we know about student reading ability and about successful models for cloze test creation to discover a model which may predict words in a similar fashion as humans.

The first issue we have seen was the issue of long distance relationships, as noted by (Beinborn *et al.*, 2015), who had to use semantic relatedness to model long distance connections and found that the method works well. The second issue is of bidirectionality. As discussed in the Background section, the traditional language modeling task is to predict a word given its history. However, students must use words that come after the unknown word. For example, the blank in "the ___ used the litter box" is impossible to guess with only the left context ("the"), but given the right context, the words "litter box" show with high probability that the blank is "cat." The final issue pertains to the development of future digital learning applications. These language models should be easy to train without an inaccessible amount of data that explodes

Table 4.3: Features of Language Models

	N-gram	RNN	LSTM	Transformer
Variable length input		✓	✓	✓
Long distance context			✓*	✓
Bidirectional (Sequential)		✓	✓	
Bidirectional (Parallel)				✓

*LSTMs typically can consider more context than RNNs, but less than transformers

training time to unreasonable lengths. The earliest type of language model was the n-gram model, however, the n-gram is of fixed length and cannot expand to fit the length of whole sentences (De Mulder *et al.*, 2015). RNNs allow variable size input but suffer from the vanishing gradient problem and in practice do not consider much context (De Mulder *et al.*, 2015). This leaves us with LSTMs and transformer, which are both more capable of handling long distance context (De Mulder *et al.*, 2015), although transformers perform better (Vaswani *et al.*, 2017). Both model types have bidirectional variants, however, bidirectional LSTMs consider the left context and the right context sequentially (Wang *et al.*, 2019), while bidirectional transformers consider both in parallel (Devlin *et al.*, 2018). While students only read in one direction, getting the left context first, then the right context, they also have the ability to jump around in the text after reading it. Students can make observations of context clues that occur before and after the word at any time and make connections between the left and right context. Parallel bidirectionality may model this better than sequential bidirectionality. These important features of each model are summarized in table 4.2.2. It is clear from the table that LSTMs and Transformers have the most features that we are looking for

Both LSTM-based and transformer-based models are available under open licenses as pre-trained models, meaning that one only needs to download the models and the libraries to use them (Devlin *et al.*, 2018), (Howard and Ruder, 2018). Both models were intended to be fine-tuned on specific language tasks, however, since they are language models and can be used for word prediction as is, this is not strictly necessary. We will investigate the use of bidirectional transformers since we believe that the ability to process information in parallel is an important factor. However, the sequential nature of language may make the LSTM based model better suited. This model will be discussed under Future Work.

Chapter 5

THE VOCABULARY SET GENERATION PIPELINE (VSGP)

5.1 Vocabulary Test

Here, we discuss the algorithm used to estimate students' vocabulary. We first compare our method to existing tests, then present the algorithm, then discuss its limitations.

5.1.1 Comparison to Other Tests

The vocabulary test is the first step in the Vocabulary Set Generation Pipeline. It draws inspiration from the frequency-based vocabulary tests created by "Test your vocab - how many words do you know?" (n.d.), Beglar and Nation (2007) and Shei (2001). Recall that TestYourVocab.com by "Test your vocab - how many words do you know?" (n.d.) uses a logarithmically scaled frequency list to create two tests, one which tests vocabulary at a broad level and the other at a narrow level. The Vocabulary Size Test by Beglar and Nation (2007) tests students at word family frequency bands of size 1000 and give 10 pre-selected questions per frequency band. The percentage of correctly answered questions corresponds to the percentage of words in that level. In FollowYou!, students are first asked to identify their language level as advanced, intermediate, or beginner, and are given questions that test 50 words at various frequencies to determine the lowest frequency words at the edge of their vocabulary knowledge Shei (2001).

The first difference between the vocabulary tests presented above and the vocabulary test included in the VSGP is that the VSGP was built for flexibility. The test does not create or present questions but rather chooses which words should be tested and automatically evaluates test scores to produce an estimate of which words the student knows. This method allows the pipeline to create tests from any list of words sorted by frequency, allowing it to be adapted to any language. The choice of question type and presentation is up to the user of the library.

The next difference from other tests is a consequence of the flexible design choice: the words selected are partially randomized as opposed to being handpicked for testing as they are in (Beglar and Nation, 2007). However, since frequency is an important factor in knowing a word (Beglar and Nation, 2007), we believe that to a certain extent, any words within a frequency band are equal good indicators of word knowledge at that frequency band as any other. “Test your vocab - how many words do you know?” (n.d.) points out some several issues with using a frequency list “straight” with no modifications, including cognates, words which may be unknown but are easy to guess if the user knows a related word, and words that users might think they know but actually don’t. These issues are specific to the language being tested, the first language of the user, and to the testing format. To keep the test flexible, pruning the list of unhelpful words is left to those using the tester. However, there is a more general issue that the tester does address. If the words tested are not evenly distributed across the frequency band it will not represent the frequency band well. The test algorithm, which is discussed in the next section, addresses this.

The next difference is that of granularity. In the Vocabulary Size Test, each question represents 100 words. We believe that estimating exact vocabulary knowledge requires a finer level of granularity than 100 words per a question. However, granularity means that the test will be longer. FollowYou! addresses this by asking for the student’s level and selecting a test specific to that level. Since the vocabulary list that the test is created from can be of any size and any level, we did not feel this approach fit well. The library can be used with a beginner-only vocabulary, for example, and it would not make sense to ask for the student's level in that case. TestYourVocab.com uses a two-layer approach to first discover the user's approximate frequency level and then narrow it down more specifically, but the test assesses only around 160 words in total, which we believe is insufficient for exact word estimation. We suggest a similar approach that repeatedly narrows down the user's estimated vocabulary until a point specified by the developer.

5.1.2 Testing Algorithm

The vocabulary test comes in two pieces. The first creates the vocabulary test and the second analyzes the results. This allows users of the library to present the vocabulary in any way they like: They may present a questions, matching questions, or simple yes/no questions. The analysis half of the tester can take the results of the test and produce the student's frequency level.

The vocabulary test creator takes a vocabulary list, sorted by frequency, as input. The vocabulary list is split into four parts, each still in order of frequency so that the first quartile contains the top 25% most common words and the last quartile contains the bottom 25%. Each quartile has a test containing twenty words. To ensure an even distribution, the quartiles are further split into twenty lists and one word is chosen from each of those, so that every 5% of the list has one word. The quartile tests are returned, along with the number of words in the quartile, which is used to calculate results. An illustration of the method for creating each quartile test is presented in Figure 5.1

The vocabulary test analyzer first analyzes the test results of each quartile. For each quartile, it calculates the percent of words known. It does that by a calculating the running average of percent of words known in the quartile. Words are only

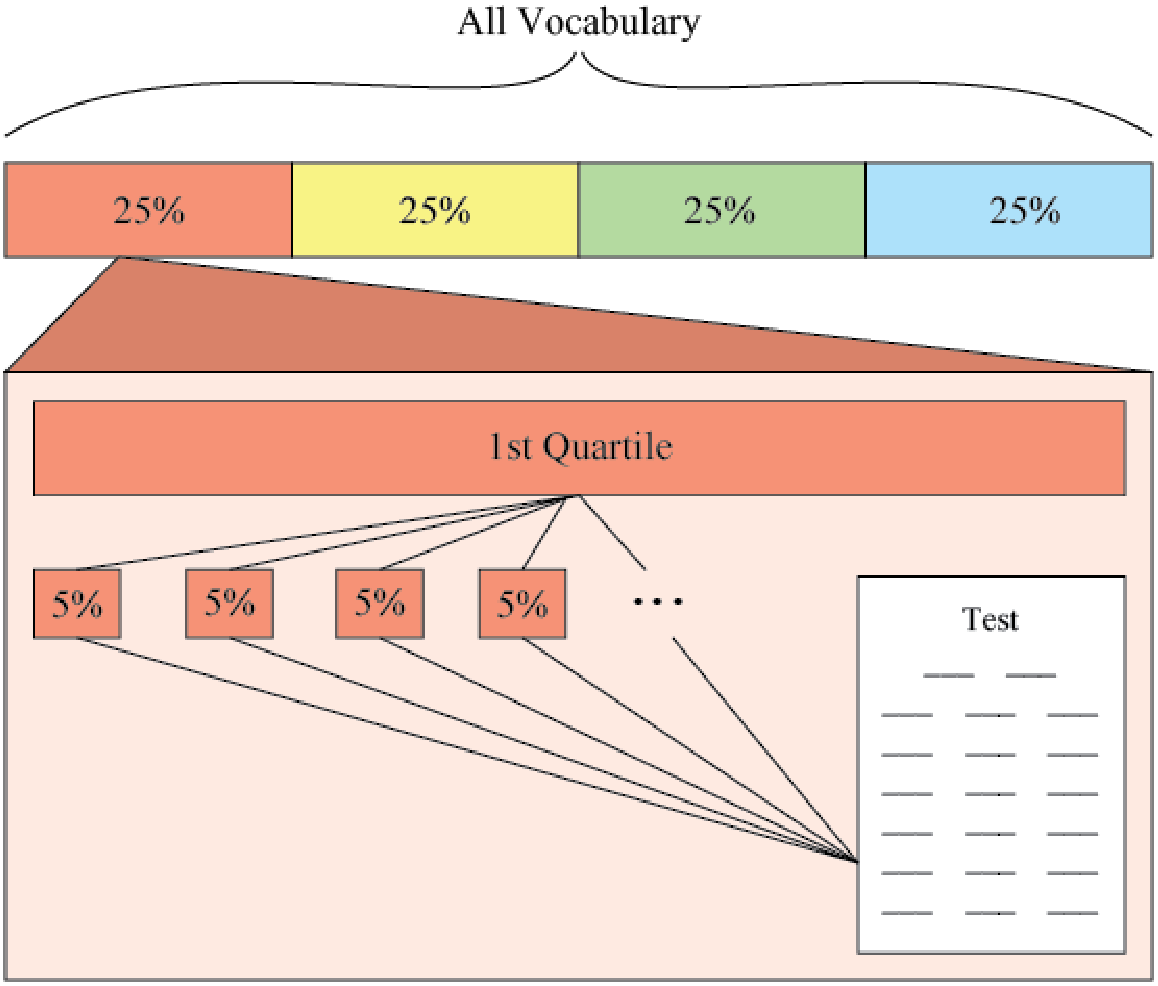


Figure 5.1: Illustration of Quartile Vocabulary Test Creation

considered to be valuable for indicating possible other known words if they appear in sequence. If, for example, a user consistently gets the first 50% of words in a quartile mostly correct, and then sporadically gets some words correct in the next 50%, this indicates that the user likely knows many of the words below the 50% line and that the other words were likely outliers. If the running average falls, and stays, below 90%, all words after that point are considered outliers thus do not contribute to the quartile's percentage of known words. If the final percent was below 10%, the entire quartile is discarded because so few known words indicate likely outliers.

After each quartile's percent of words known is calculated, the results are used to calculate the frequency range the student belongs in. First, we determine if any of the quartiles can be completely removed from the range. Consecutive early quartiles with 100% vocabulary knowledge are considered completely known and consecutive later quartiles with 0% vocabulary knowledge are considered completely unknown. This determines which quartiles the student's vocabulary knowledge does lie in, what we will call knowledge-quartiles for the sake of brevity. Then, the percentage of consecutive known words (as defined by our process of using a running average of percent known) of each knowledge-quartile is used determine which part of the knowledge-quartiles the student's vocabulary knowledge lies in. In the earlier knowledge-quartile, the percentage (X%) of consecutive known words is used to remove the first X% of words from the knowledge-quartile, because these words are considered completely known. The words in the last known knowledge-quartile which were not percentage of consecutive known words are removed because the words that are not known consecutively are considered outliers. Figure 5.5 illustrates the process of narrowing down the student's believed vocabulary region from the entire given vocabulary to the region indicated by the test results. The process leaves us with a list of words, where the student's vocabulary frequency level is thought to be on one of the words within that list. The frequency levels at the start and end of this list are returned.

After the region where the student's vocabulary knowledge is estimated to be in is calculated, the test may be run again within only that region to narrow it down further. The end of the range represents the very edge of where we believe the student's vocabulary knowledge might lie and the start of the range indicates the point below which most vocabulary should be known. The developer may choose to use either value to estimate vocabulary; using the start value is stricter. The complete pipeline is presented in figure 5.3.

5.1.3 Discussion

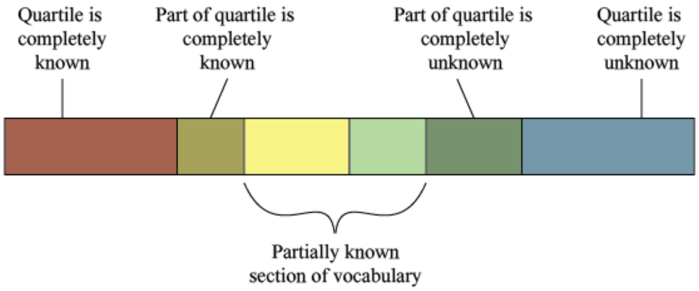

Quartile is completely known
Part of quartile is completely known
Part of quartile is completely unknown
Quartile is completely unknown
Partially known section of vocabulary

Figure 5.2: Narrowing the Region within which the Student's Vocabulary Level Liesw

Being a frequency-based vocabulary estimation tool, the test does have several issues which have been noted by previous researchers, such as (Heilman and Eskenazi, 2006). The project REAP, presented by Heilman and Eskenazi (2006), is a vocabulary focused language tutoring system that includes vocabulary assessments. REAP takes a list of target words as input and uses those to find texts through which the students can learn the target words. The texts are not chosen based on what words the student already knows, but instead are chosen based on their grade level in the American school system. The results of the first vocabulary assessment are used to filter out known words from the student's target word list. Further testing is done to measure if students have learned target words. Although REAP discussed the importance of having a student model containing words known to the student for selection of appropriate reading material, testing is not used in this manner due to several issues with estimating student vocabulary knowledge which are discussed in the paper.

Vocabulary knowledge cannot be tested by testing every word a student might know. Therefore, testing methods must estimate students' knowledge of all words with their knowledge of only a subset of words. As the researchers developing REAP point out, this estimation is error prone especially because of gaps in student knowledge. Many vocabulary tests rely on frequency lists to determine the "frequency

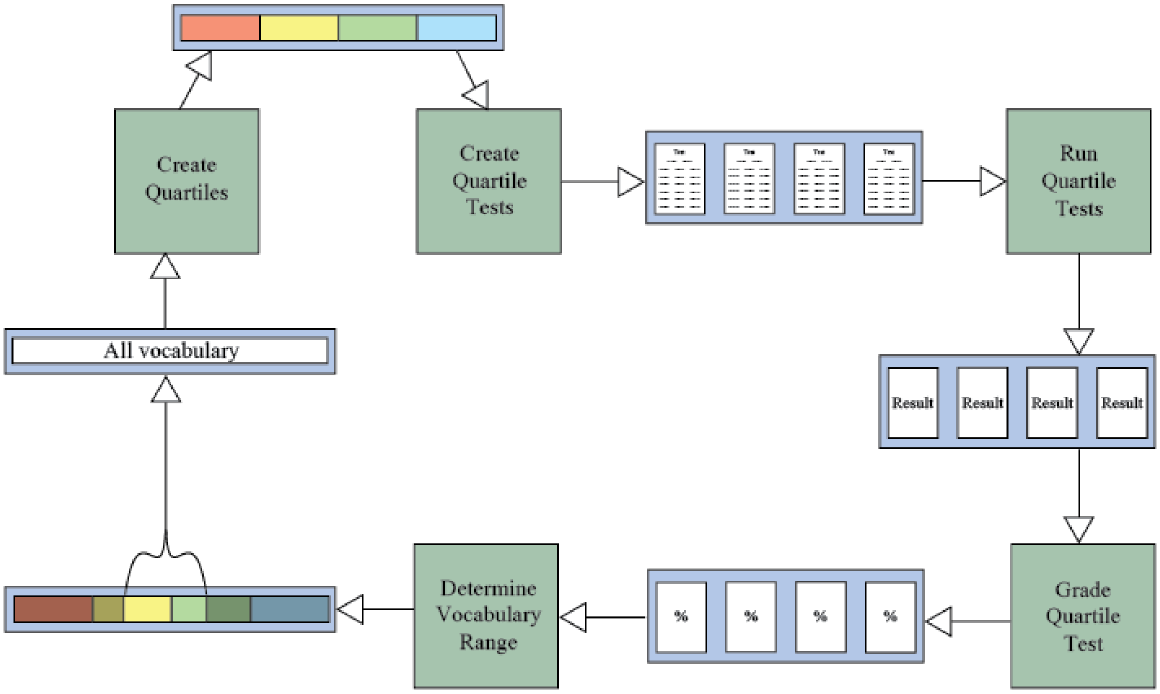


Figure 5.3: Illustration of Complete Vocabulary Test Pipeline

band" a student is at (Pignot-Shahov, 2012). However, this does not account for gaps in student knowledge. A student may know many of the first 2000 most common words in English, but perhaps not all Beglar and Nation (2007). A student may also know plenty of less common words in a wide variety of lower frequency bands and these words are missed with a frequency-based vocabulary test Beglar and Nation (2007). Although the Vocabulary Size Test has been validated, its creators also noted that word frequency is only one factor in vocabulary acquisition order for language students, albeit they note that it is an important factor Beglar and Nation (2007).

Being a frequency-based vocabulary estimation tool, our test suffers from these issues. Rather than producing a perfect list of known words, this method finds a point at which many words are no longer known in order of frequency. However, no vocabulary estimation test can be completely perfect without testing every word. Since frequency has been noted to be an important factor in vocabulary acquisition order, we believe that this estimation is still useful. Further research is required to determine other factors that may be used to estimate word knowledge. Being that the focus of this project was work on prediction of difficult words in a text, we believe that this test is adequate for exemplifying the process of the vocabulary set generation pipeline.

5.2 Vocabulary Set Generator

Here, we discuss the generation of vocabulary sets, given a text and the student's vocabulary. First, we describe the model used to represent the document, then the model used to represent the student, and finally, the complete pipeline together.

5.2.1 Document Model

The first component in the generator is the Document Model. The Document Model consists of Sentences and Words which represent the text that the student is trying to read. The Document model also contains the Student Model to simulate reading of the text. The most fundamental part of the Document model is the Word class. The Word class includes all the information needed to determine if the word is known or unknown. The attributes of the word class are presented in Figure 5.4. The two of the most important attributes are the actualTokens and modelTokens. BERT models use WordPiece embeddings, rather than whole word embeddings, so tokenization (and prediction) is done on a sub word level (Devlin *et al.*, 2018). This means we must keep track of which word each token belongs to. The actualTokens attribute is a list of the original tokens of the word, as generated by the BERT Tokenizer stored in the Student class. The modelTokens attribute is a list of the tokens that will be passed to the model for prediction. BERT uses a special token, "[MASK]" to represent tokens that are masked and must be predicted. When a Word object is created, it checks if it is a known word by asking the Student Model. If it is a known word, the modelTokens attribute is the same as the actualTokens attribute. Otherwise, the modelTokens attribute is a list of "[MASK]" tokens as long as the actualTokens. Other important attributes include the isMasked attribute, which is set to true if the Student Model reports that the word is not known, and the predicted attribute, which is set to true if the student model reports that the word can be predicted.

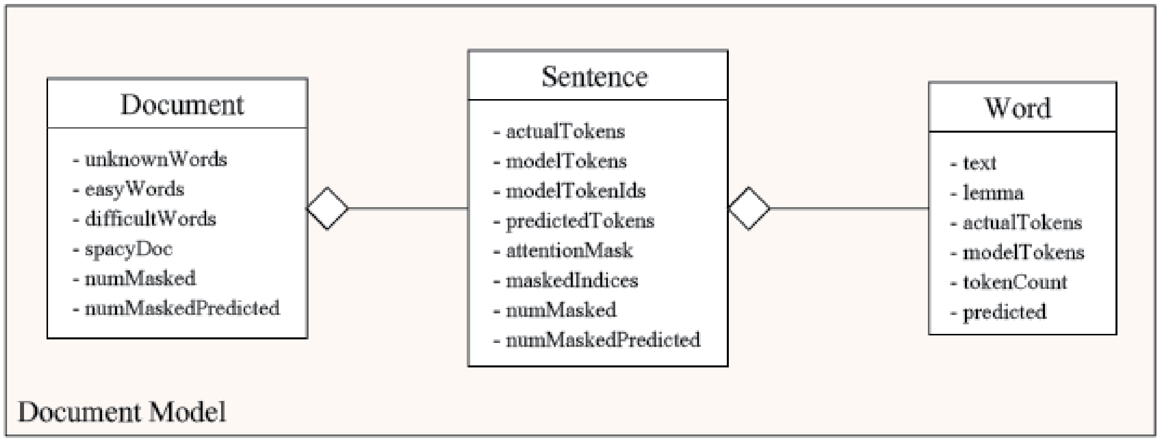


Figure 5.4: Class Diagram of Document Model Classes

Words are the basic building blocks of the Sentence class. The attributes of the Sentence class are shown in 5.4. Like the Word class, the Sentence class contains the actual and model tokens of the sentence. For each word in the sentence, a Word object is created and the word's actual and model tokens are added to the actual and model token lists of the sentence. To be processed by the Student Model, the modelTokens list also includes special start and end tokens, as well as padding tokens that fill the list to its maximum size. The Sentence class mostly consists of information required for predictions, including an attention mask, which informs the Student Model of which part of the input is padded and which is not, and a list of token IDs, which are generated by the BERT Tokenizer and used by the BERT model to make predictions. The Sentence class also contain the indices where masked words are at so they can be compared when predictions are made.

The Document class is made of Sentences. To create the Sentences, the Document uses spaCy, an industry strength natural language parsing library in Python, to extract each sentence from the text ("spacy · industrial-strength natural language processing in python", n.d.). Each sentence generated by spaCy is parsed to create the attributes of the Sentence class as discussed above. Since all sentences passed to BERT must be of a specified length, the average of length of each sentence, in words, is calculated and the max length of a sentence, in tokens, is set at 2.5 times the length in words. Should any sentence go beyond that length in tokens, the sentence is simply cut in half.

Once a text is parsed into a Document, it is ready to be predicted. Since prediction is performed by the Student Model, we will discuss the Student Model first before discussing the complete pipeline.

5.2.2 Student Model

The Student Model contains all information needed to discover unknown words and unguessable words. The most important of these include the list of vocabulary known by the student and configuration parameters specific to the student being modelled. The configuration parameters determine what BERT model to use and how to use it. As discussed in Methodology, there are several BERT models to choose from and different models may model different students better than others. Other factors used to personalize the behavior of the pre-trained model include: number of guesses available to the model, whether to use the spaCy word similarity functionality in determining if a word is predicted correctly, and whether to use synonyms if a word is predicted correctly. Since the source of synonyms used in testing limits API requests for non-commercial purposes, the synonym functionality is included in the Student Model but not used currently.

With the known vocabulary list, the configuration information and the BERT model created from the configuration information, the Student Model performs three major functions. First, it determines if the student knows a word by determining if the "word" is actually a number, punctuation, or other symbol; if the word is a proper noun (named entity); and finally if the word is in the known word list. Words are also considered known if they are function words (words that serve a grammatical purpose) because we are aiming to create vocabulary lessons, not grammar lessons. The second major function of the student Model is running predictions on sentences to determine which words will be difficult for the student to guess. We use the huggingface transformers Python library that contains many transformer language models encapsulated in easy to use code ("Bert", 2020). The Student Model uses the library to predict words when given a Sentence object. The Student Model then updates the Sentence to record the predicted words. The final major function the Student Model performs is determining if a word has been predicted. Since BERT predicts sub word tokens instead of whole word tokens, we must reconstruct each predicted word from the word's predicted tokens, taking into account that BERT is given multiple chances to predict a word and a word must be constructed from each prediction. Given a list of the guesses for each of the word's predicted tokens, the tokens for each guess are combined to form a word. The predicted tokens are then compared to the lemma of the original word. The lemma of a word is the base form of a word that can be transformed for grammatical purposes (inflection). This includes grammatical transformations such as pluralization and verb conjugation. By comparing the lemma of the prediction and target word instead of the word in its original form, we get results that better mimic student comprehension (see Methodology and Results). Then, if the configuration is set to include similarity or synonyms, we check if the prediction is similar using spaCy's similarity score and if the prediction is a synonym with the

Merriam-Webster Thesaurus API. The three main functions of the Student Model are used by the Document Model to model the student reading the text, which is described in the next section.

5.2.3 Final Generation Pipeline

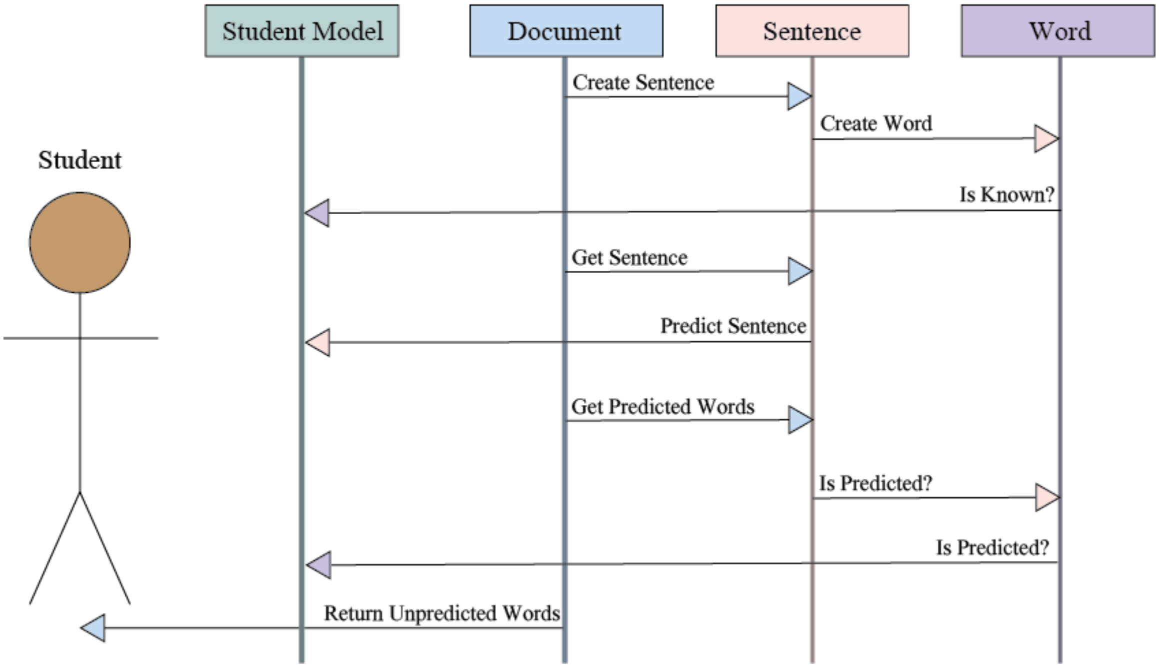


Figure 5.5: Illustrations of the Vocabulary Test Set Pipeline

The first step in the final generation pipeline is to create a Student Model. The student model expects a vocabulary list, which can be gathered with the vocabulary test and stored in any form that the developer chooses, and a model configuration object containing the parameters for the model to perform under. During the analysis of models described in Methodology and Results we select the best configurations for different levels of students (beginner, intermediate, and advanced). However, while these configurations provide baseline information about a student, there is a possibility for the configuration to be adjusted as a student uses the application. Thus, the suggested baseline configurations are provided but are not built-in, allowing for the developer using the pipeline to make adjustments as needed.

After the Student Model is created, the Document model is created. To create a Document Model, the text of the document to be read and the Student Model are required. When the Document creates the Sentences and Words that it is made of, the Student Model is used to determine which words should be masked. When the Document is predicted, the Document Model cycles through the Sentences and uses the Student Model to the predict the sentence's tokens. Then each Word in each Sentence is updated, using the Student Model to determine if the word has been correctly predicted. When all words have been marked as predicted or not predicted, the Document Model gathers all the words that were not predicted. These are returned as the vocabulary study set.

Chapter 6

METHODOLOGY

6.1 Comparing Model Performance to Human Performance

Here, we describe the method for comparing the model's guessing ability to human's guessing ability. We chose model configurations we believe will perform well based on a configuration data set, and test the chosen configurations on a testing set.

6.1.1 Description of Data

Cloze tests are a popular method for testing student vocabulary knowledge, reading comprehension skills and word prediction skills. For this reason, cloze tests were chosen to test the guessing ability of the selected masked language models and to compare the guessing ability of the masked language models to student guessing ability. To compare the model's performance to human performance, we need two related data sets: A set of cloze questions with correct answers and a set of human responses to those cloze questions. With this data, we can have the model "answer" every question, then compare the model's ability to predict words to human's ability to predict words, as described below under "Procedures." Existing data sets were used. The original cloze questions are by Zweig and Burges (2012). The human responses to the cloze questions are made available by Beinborn *et al.* (2014a) and the first 100 of these responses were used by Beinborn *et al.* (2015), whose study was already described in Related Work under Student Modelling in CALL. The following sections provide information on the original cloze question data set, the human response data set, and how the data was processed before it was used for analysis.

Original Cloze Questions

The cloze questions used are a subset of the data set generated by Zweig and Burges (2012). This subset was selected by Beinborn (2016) for their research on question difficulty prediction which includes Beinborn *et al.* (2014b) and Beinborn *et al.* (2015). The data set generated by Zweig and Burges (2012) was originally intended for use in language modelling. The goal was to create tests that were challenging for the N-gram models that existed at the time by creating answers which can only be discovered by considering the larger context of the sentence. The questions are modelled on questions used for human students and therefore are suitable for research on language learners, which is what Beinborn (2016) used the questions for. The sentences were selected from 5 Sherlock Holmes books. Each sentence contains a low frequency word which was selected as that left blank in the cloze question. Four distractors are chosen in a two-step process: First, a list of 30 possible distractors is automatically selected, then the list is groomed by humans to remove synonyms and grammatically incorrect words. The distractors are unimportant for the purpose of testing the models because the ultimate application of the model is not to select the correct choice from a list of distractors but instead to generate a correct choice. Therefore, to better emulate the environment that the model will be used in, distractors are not considered for the purpose of scoring the model. The data set generated by Zweig and Burges (2012) consists of 1040 questions, 200 of which are used in this thesis. The 200 questions were selected by Beinborn (2016) and used to collect data on the difficulty language learners have with cloze tests. The results of their study showed that there is a variety of both easy and difficult questions because there is a high standard deviation of the error rates (number of incorrect responses/total number of responses) for the questions in this set.

Human Responses

Beinborn (2016) (also (Beinborn *et al.*, 2015)) conducted a series of surveys to collect data about the difficulty for humans of the 200 cloze questions. The surveys conducted were in the following form: Each survey contained 10 questions and asked students for their first language, their approximate number of years of English study (in buckets of "¡ 1", "1-3", "3-5", "5-8", "8-10", or "> 10"), and their self-identified CEFR (Common European Framework of Reference for Languages) level (A1, A2, B1, B2, C1, or C2). The students are then asked each of the ten multiple choice questions. Because of the test format, the number of responses per a question is not consistent. The study also targeted advanced English learners, so the data is skewed towards high number of years studied and the C levels of CEFR. According to Beinborn (2016) the students were allowed to take multiple surveys, but data was not collected to connect responses from each survey to specific students. Thus, we will approximate the number of students that the dataset contains by assuming that each student took one ten-question survey, so every ten responses represents a single student. The total number of responses was 4310, for a total number of students of 431. Table 6.1 describes the data in detail. Because the level of granularity indicated by the CEFR levels was considered unnecessary for the purposes of this research, students were grouped by their CEFR level into larger levels of advanced (CEFR C1 and C2), intermediate (CEFR B1 and B2), and beginner students (CEFR A1 and A2). Beinborn (2016) considered the students' reported CEFR levels were unreliable. We believe that this makes the data more realistic given that the VSGP currently relies on self-reported levels of beginner, intermediate, and advanced. However, we will keep this in mind when evaluating the model's performance. Table 6.1 shows information on the number of questions and responses given for each level. As the data is skewed to more advanced levels, the lower levels are less representative of the general student population at that level. The combination of all students contains the most information, of course, so we will also analyze the model's performance with the complete data set.

	Beginner	Intermediate	Advanced	All
Questions Answered	170	200	200	200
Total # of Responses	380	1730	2200	4310
Total # of Students	38	173	220	431
Avg # of Responses Per Question	2.24	8.65	11	21.55
Max # of Responses Per Question	4	13	17	34
Min # of Responses Per Question	1	5	7	16

Table 6.1: Distribution of Questions and Responses Across Student Levels

Data Preparation

Two types of preparation were made before the data could be used in the validation procedures. First the student groups are split further into percentiles roughly representing better or worse performing students who are at the same level. This is similar to grouping students in the same course level by their performance; even if two students are at the same level, they might not have equal performance at that level. This grouping determines what questions in the data set are considered hard. If we are considering lower performing advanced students, than the questions where 25% or more students answered incorrectly are considered difficult, whereas with high performing advanced students the threshold is raised to 50%. Initially, 75% was also considered a threshold, but this was discarded for two reasons: one, there is not enough difficult questions at the 75th percentile for some student levels to get an accurate representation of the 75th percentile, and two, it was acknowledged that the 75th percentile is not very useful as it represents very difficult words to guess and only a small portion of language learners would benefit from a model representing that level of difficulty. An additional grouping was also used, based on the method Ko (2012) used to determine which words from the pilot study should be glossed. In (Ko, 2012), words marked as unknown and unguessable by more than 60% of the students in the pilot study were chosen to be glossed. While we believe that the 25% and 50% levels provide opportunity for more customization, we also want to compare our work to the threshold determined by an expert in the domain.

The second type of preparation splits the data into configuring and testing data sets. This is inspired by the training and testing data sets used for machine learning, which first train the model and then test the model to verify that it works on a general set of data and was not over fit to the training set (Liu and Cocea, 2017). We are not training a model but rather determining the correct set of parameters for using the model with each student group. To validate the chosen parameters, we test them with a subset of the data. In machine learning, a common split for data sets is 70% of samples for training and 30% for testing, with the samples being selected randomly for placement into either set (Liu and Cocea, 2017). This split is meant to allow for a large amount of data to be used for training while also leaving enough for testing. Since we do not need to train a model, and therefore do not need large amounts of data for any training, we opted for an even split instead.

Splitting data in this way is haunted by two issues however: class imbalance and sample representativeness (Liu and Cocea, 2017). When data is a set can be split into different classes, class imbalance refers to an uneven distribution of instances of those classes between the training and test sets. Sample representativeness refers to the similarity between instances in the testing and training sets (Liu and Cocea, 2017). The data used in this thesis may be particularly susceptible to class imbalance. The samples in our data set are split into the classes of "difficult" and "easy." If, for example, random sampling happened to cause the configuring set to contain no difficult samples, then not only would our module be configured incorrectly, but we would also likely see much worse results on the test set, which would contain an abundance of difficult questions. Liu and Cocea (2017) presents an approach which uses granular computing concepts to solve both the issues of class imbalance and sample representativeness. According to Liu and Cocea (2017), test sets can be generated with 3 levels of granularity. The first level is entirely non-granular: the data is split randomly among the test and training sets. In the second level, which attempts to address class imbalance, the data set is first split into classes and the classes are then split into the training and test data sets. The training and test data sets for each class are then combined to create the training and test data sets for the entire set. The third level splits the data set further into subclasses and again partitions the data in these subclasses into training and test sets. The resulting training and test sets are combined to create the final training and test sets. For the purposes of our study, only class imbalance is considered as the data can be most easily split into the classes of "difficult" and "easy," but it is not entirely clear what "subclasses" would exist on the continuous range of error rates within the two classes. Therefore, only level 2 splitting is used. The data is split in the difficult and easy classes and 50% of each class is placed in the configuring set while the other 50% is placed in the testing set. This ensures that the model is configured and tested with a balanced set of data. Tables 6.2, 6.3, 6.4, and 6.5 show the distribution of easy and hard words for each student group and the complete data set in the original, configuring and test data sets.

Table 6.2: Distribution of Easy and Difficult Questions for Beginner Students

	25th Percentile	50th Percentile	60th Percentile
Percent Easy Questions	0.51	0.72	0.72
Percent Difficult Questions	0.49	0.28	0.28
Total # of Easy Questions	87	123	123
Total # of Difficult Questions	83	47	47
Subset # of Easy Questions	43	73	73
Subset # of Difficult Questions	42	24	24

Table 6.3: Distribution of Easy and Difficult Questions for Intermediate Students

	25th Percentile	50th Percentile	60th Percentile
Percent Easy Questions	0.53	0.78	0.82
Percent Difficult Questions	0.48	0.23	0.18
Total # of Easy Questions	105	155	164
Total # of Difficult Questions	95	45	36
Subset # of Easy Questions	52	77	82
Subset # of Difficult Questions	48	23	18

6.1.2 Procedures

The following sections outline the procedures for not only testing the ability of the model to predict the guessability of a word, but also for adjusting parameters to configure the model to a particular group of students. First, we examine the specific

Table 6.4: Distribution of Easy and Difficult Questions for Advanced Students

	25th Percentile	50th Percentile	60th Percentile
Percent Easy Questions	0.66	0.89	0.92
Percent Difficult Questions	0.34	0.11	0.08
Total # of Easy Questions	132	178	184
Total # of Difficult Questions	68	22	16
Subset # of Easy Questions	66	89	92
Subset # of Difficult Questions	34	11	8

Table 6.5: Distribution of Easy and Difficult Questions for All Students

	25th Percentile	50th Percentile	60th Percentile
Percent Easy Questions	0.59	0.84	0.88
Percent Difficult Questions	0.42	0.17	0.12
Total # of Easy Questions	117	167	176
Total # of Difficult Questions	83	33	24
Subset # of Easy Questions	58	83	88
Subset # of Difficult Questions	42	17	12

questions that we aim to answer with these procedures, then outline the procedures used, and finally define the criteria used for determining the success of the model.

Research Questions

This section discusses research questions specific to the judging of the language models available and the choice of parameters used for each group of student being tested. As previously discussed, not only do we look at students at different levels, beginner, intermediate and advanced, but we also look at different percentiles within those levels, 25th 50th, and 60th percentile. Thus, we have 9 different user groups to study, and we must determine what values for which parameters work best for each group.

There are several parameters that must be examined to determine what works best for which level of student. The following parameters are considered in this study: the specific BERT model used, the number of predictions the model is allowed, the use of synonyms provided by the Meriam-Webster API, and the use of similarity scores provided by the Spacy NLP library. The first parameter refers to selecting one of the many available BERT models released by Google. We used the English language models and multilingual models only. These models were used as is, with no further fine-tuning for language modeling tasks as we want to evaluate the performance using the models as they are. The names of the models used are as follows:

- bert-base-multilingual-uncased
- bert-base-multilingual-cased
- bert-base-uncased
- bert-large-uncased
- bert-base-cased
- bert-large-cased
- bert-large-uncased-whole-word-masking
- bert-large-cased-whole-word-masking

The latter three parameters are used to determine if a question is answer correctly. For example, if the model is allowed 10 guesses and gets the answer correct on the sixth guess then the answer is considered correct. Similarly, if the answer is a synonym of the correct answer or is sufficiently similar to the correct answer, the answer is considered correct. For the purposes of this study "sufficiently similar" is defined as having a similarity score of more than 0.70. The similarity score itself, which is a decimal value ranging from 0 to 1, could have been a parameter but this level of granularity was considered infeasible. The similarity score chosen is admittedly somewhat arbitrary but aims to require a high level of similarity without being too restrictive. These parameters all have the potential to be used in the final language pipeline which produces the vocabulary set.

The parameters for determining if a question was correctly answered (the number of predictions, the use of synonyms, and the use of similarity scores) were chosen because it is believed that they might result in the largest performance increase for the model. The number-of-guesses parameter might increase performance if the model is consistently getting easy words wrong. For example, one sentence in the data set included the phrase "I hope to ____." The correct answer for this question was "I hope to heaven." The first guess of one BERT model was "I hope to God." "Heaven" was the model's second guess. Including scenarios like these may increase the model's performance, although it may cause the model to begin guessing words correctly that should be difficult.

Synonyms were included because in reading it is often sufficient for students to guess synonyms. For example, if a student filled in a cloze question with the word "large" when the correct answer was "huge," it would generally be accepted that the student was actually correct. By accounting for synonyms, we partially fill a deficiency that is inherent in using language modeling to of model human prediction ability: by the design of the model, the model can guess the probability of specific words, while humans only need to guess the meaning of words to understand the text. It has also been suggested that learning only partial meanings of new words encountered is not necessarily bad, since incidental vocabulary learning is a cumulative process (Nation, 2013). This brings us to the reasoning behind using similarity scores to determine the guessability of a word: if 70% of a word's meaning is understood, this may be enough for a student to understand the text and continue reading. Of course, as explained in spaCy's documentation, the similarity score provided by the library may not always be what humans expect. For example, "cat" and "dog" are 80% similar according to the example available on the Spacy website. In some contexts, this may be sufficient as a student might simply need to know that the word means "domestic animal." In other contexts, the distinction may be more important. One could argue that in a more specific context, more information would be provided, allowing both the language model and the student to determine which interpretation is correct. Detailed arguments about the efficacy of these parameters are unnecessary, however, because these parameters will be validated and tested against the data set.

The procedures of this study then aim to answer the following questions:

- Which BERT model works best?
- Will including synonyms or similarity scores improve the model's performance?
- Will looking at the first X guesses give better results? What number is X?
- What configuration of each of the above parameters works best for each the twelve student groups being considered?

Procedures

Unlike much research on language modeling, the goal of this research is not to determine if the model can answer the questions correctly, but rather if the model can answer the questions in the same way as a human. The basic procedure is as follows: Have the model "answer" all questions in the data set, then compare the model's answers to the human's answers. If the model answers a question correctly, this question is considered to be easy according to the model, otherwise it is difficult. The definition of "correct" is different depending on the parameters being tested:

- An "exact" correct answer is when BERT guesses the exact word or a lemma of that word (if BERT predicts "like" but the word was "liked" this is still considered an exact match).
- A "similar" correct answer is when BERT guesses a word that has a 70% or higher similarity score according to the Spacey NLP library.
- A "synonym" correct answer is when BERT guesses a word that is a synonym according to the Merriam-Webster API
- An "any" correct answer is when BERT guesses a word that is correct according to any of the above measures

All of these correct answer types are tested repeatedly, allowing for each model to make 1-100 guesses. Since six models are tested, 600 guesses are ultimately made for each question. Because an extreme level of granularity is not necessary, the models are evaluated on guess-levels of size 10, so that the first 10 guesses are considered, then the first 20 guesses, etc. This results in 10 guess-levels for a total of 60 configurations to examine (guess levels to be considered X number of models to be considered). Once the parameters for exact, synonym, and similar answers are taken into account, the result is 240 model configurations must be compared.

After data is collected to determine the difficulty of words according to each configuration, the results are compared to the results generated by humans. The difficult and easy words for each configuration are compared to the difficult and easy questions for each human group. This means that ultimately 2880 tests must be run. This is done once with the configuration set to determine which parameters are the best. Once the parameters are chosen for each student group, the parameters are tested with the test set to see if they still perform well.

6.2 Examining the Output of the VSGP

In absence of a user study, we will examine the output of the VSGP with test students and texts and compare this with vocabulary sets used in vocabulary pre-teaching and glossing studies.

6.2.1 Materials

We created vocabulary lists for five sample students and two sample texts, for a total of 10 sample outputs. The sample students belong in each level, advanced, intermediate, and beginner. We use the models configured for 60% difficulty at the respective levels since this was the threshold used by Ko (2012) to chose words for glossing. The first advanced student was set to be at the 5000th frequency band of English vocabulary, the intermediate student at 3000 and the first beginner at 1000. To understand how changing the model type affects vocabulary, we have additional test students with which we controlled for vocabulary. These additional students consist of one a beginner and one advanced student, which have the same amount of vocabulary as the intermediate student. This is to compare students with different guessing abilities, grammar skills, and other non-vocabulary skills required for reading. To get a sampling of different types of texts, we chose one fictional text, *Alice's Adventures Under Ground* "Chapter 1" by Lewis Carroll and one news article from the Wall Street Journal titled "The Class of 2020 Was Headed Into a Hot Job Market. Then Coronavirus Hit." (Dill and Thomas, 2020). We have selected these texts to showcase different genres and lengths. They represent upper level, more difficult texts that we would expect students to need more support with.

The different level of students and different types of texts should give us an idea of what kind of vocabulary students using the vocabulary test set generator might encounter. We show the word count, the Flesch Reading Ease Score and the Flesch-Kincaid Level, as calculated by Microsoft word. The Flesch Reading Ease Score and the Flesch-Kincaid Level are old but often used measures of readability (Beinborn, 2016). The Flesch Reading Ease Score goes from 0 to 100, where higher scores are better. Flesch-Kincaid Level shows the approximate grade level in terms of the US education system. See table 6.6 for a summary of the texts.

Table 6.6: Features of Sample Texts

	Word Count	Flesch Reading Ease	Flesch-Kincaid Level
Fiction	3944	72.5	9.7
Non-Fiction	970	42.1	13.3

6.2.2 Procedures

We ran the Vocabulary Set Generation Pipeline once with each text and student. The model configurations for each student group matched those chosen from the model testing. The number of words in the produced vocabulary lists should give us an understanding of which students will have more words to study when using the VSGP; advanced students should need fewer words than beginning students. After generating the vocabulary lists, we will compare the number of words in the lists to the number of words in the text, specifically considering the amount of words needed for minimal comprehension and comfortable reading. We hope to show that the number of words in the produced vocabulary sets is reasonable given the number of words in the text and the number of words originally known by the student

Chapter 7

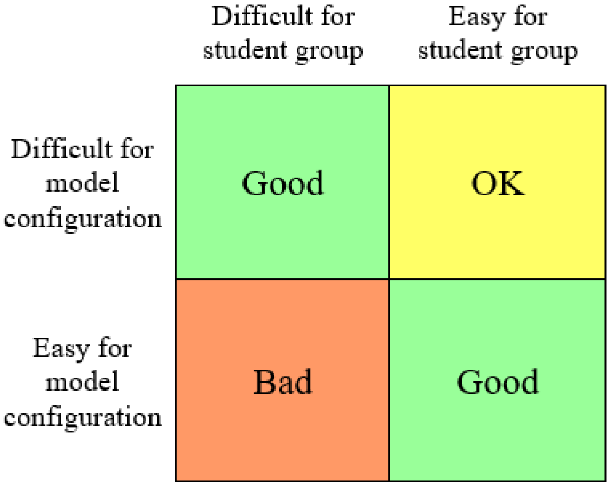
RESULTS

7.1 Results of Model Assessment and Verification

Here, we examine discuss criteria for judging the language model configurations and examine the results of the model configuration and testing procedures.

7.1.1 Criteria

In the Vocabulary Set Generation Pipeline, the model takes all unknown words in the text and attempts to determine which are difficult (true positive) and which are easy (true negative) so that difficult words can be added to the vocabulary study set. Since difficult words are considered unguessable, we know that too many false negatives may result in a reduction of student comprehension. Adding easy words to the study set (false positives) will not negatively affect comprehension, but it will reduce the number of words that students can guess in context. We would ideally like to maximize the number of words guessed in context because learning guessable words explicitly wastes time and removes a chance for practicing guessing skills. However, neither issue is as detrimental as missing a difficult word. Learning guessable words explicitly is not necessarily a waste of time as it will deepen students' knowledge of those words. Even though the words could have been guessed, the explicit learning still has value. If enough words are out of the study set, students will also still have ample opportunities to guess in context, so it is arguable if absolutely all easy words must be removed from context or if removing a decent sized subset is good enough. On the other hand, the reason for glossing and pre-teaching is to make a text comprehensible to students, so failing at this goal is worse than including many extra words. This is illustrated in figure 7.1. For this reason, we chose to examine sensitivity, the ability for the model configuration to detect difficult words, and specificity, the ability for the model to detect easy words. Sensitivity is defined as the number of true positive results over the number of actually positive cases and specificity is the number of true negatives or the number of actually negative cases. Ideally, the sensitivity should be as high as possible and the specificity should not be too low.



	Difficult for student group	Easy for student group
Difficult for model configuration	Good	OK
Easy for model configuration	Bad	Good

Figure 7.1: Confusion Matrix Depicting the Relative Importance of False Results

It is ideal to compare these results to some baseline. Since published texts containing vocabulary for studying readings or glossed texts are created by language teaching professionals, the accuracy of such professionals is a good baseline. Beinborn (2016) provides information on the ability of language instructors to predict the difficulty students will have with C-tests, a reduced redundancy test that assesses students' reading skills. This test assesses similar skills as the cloze test and is thus this data is a suitable measure of instructors' abilities to predict student difficulty in reading. Since we are trying to test the model configuration's ability to predict student difficulty in reading, we will compare the model configuration's performance to this. Comparisons should be tempered by the fact that this shows only instructor ability to detect difficulty in reading *in general* and not with the same data set used to test the model configuration. Since the instructors were not asked to predict the difficulty of questions at different student levels, we will only compare the instructors' performance to the model's performance for students at all levels.

The C-tests were completed by students as part of university entrance exams; no information was given on the level of the students due to data protection policies, which should also be considered when making comparisons. Three experienced test designers and university professors at TU Darmstadt were asked to predict the error rates of the set of C-tests by classifying the questions as follows: Fewer than 25% of students will get the question wrong, between 25% and 50% of students will get the question wrong, between 50% and 75% of students will get the question wrong, and more than 75% of students will get the question wrong. The confusion matrix of the original results are presented in tables 7.1, along with confusion matrices for only the 25% and 50% groups in tables 7.2 and 7.3, since we test the model configurations using these groups for determining if a question is difficult or easy. Because Beinborn (2016) did not survey the instructors at a 60% difficulty threshold, the 60% difficulty threshold cannot be compared to this data.

Table 7.1: Original Confusion Matrix of Instructors' Difficulty Prediction

	Instructors Predicted > 75% Incorrect	Instructors Predicted 75% - 50% Incorrect	Instructors Predicted 50% - 25% Incorrect	Instructors Predicted < 25% Incorrect
Actually > 75% Incorrect	7	26	8	2
Actually 75% - 50% Incorrect	6	29	34	7
Actually 50% - 25% Incorrect	1	23	59	24
Actually < 25% Incorrect	0	6	48	118

Table 7.2: Confusion Matrix of Instructors' Difficulty Prediction at 25%

	Instructors Predicted Difficult	Instructors Predicted Easy
Actually Difficult (> 25%)	193	33
Actually Easy (< 25%)	54	118

Table 7.3: Confusion Matrix of Instructors' Difficulty Prediction at 50%

	Instructors Predicted Difficult	Instructors Predicted Easy
Actually Difficult (> 50%)	68	51
Actually Easy (< 50%)	30	249

Since we will be assessing the sensitivity and specificity of the model configurations, the sensitivity and specificity of the instructors is provided in table 7.1.1 for comparison. We can see that the instructors had very little trouble distinguishing very difficult questions from easy questions. Low sensitivity at the 50th percentile shows that the instructors often miss identified more difficult (above 50% error rates) as being easy (below 50% error rates). Low specificity at the 25th percentile shows that the instructors often identified easy words in the grouping as more difficult. This is in line with Beinborn (2016)'s observation that the instructors were good at finding very difficult words but had more difficulty distinguishing between lower levels of difficulty. A final thing to note before making comparisons is that Beinborn (2016) observed low agreement between the instructors, despite having similar backgrounds. The implications of human inconsistency will be discussed later.

Table 7.4: Instructors' Specificity and Sensitivity at Relevant Percentiles

Percentile	Sensitivity*	Specificity**
50%	0.57	0.89
25%	0.85	0.69

*Percent of difficult words identified

**Percent of easy words identified

7.1.2 Results

Here, we examine the results of the best performing model configurations. Only the best configurations are shown here because there are many model configurations and many student groups to consider. Since we value sensitivity (the ability to find hard words) over specificity (the ability to find easy words) we attempt to choose model configurations which showed at least 0.75 sensitivity and 0.50 specificity in the configuring set, and prioritize improving sensitivity over specificity. However, should sensitivity reach at least 0.90, we prioritize specificity. Table 7.5 shows each and the performance of the selected model configuration on the test set in terms of sensitivity and specificity We also show the percent difference from the sensitivity and specificity of the instructors in the All students group in Table 7.6. The model configurations were able to reach the target sensitivity of 75% for four of the twelve student groups when tested with the test set. The target specificity of the 50% was reached for seven of the nine student groups when tested with the test set. Since 60% was the threshold used by Ko (2012) to determine if words should be glossed, we review the details of the best model configurations for each student level at the 60th percentile in table 7.7.

Table 7.5: Model Configuration's Ability to Detect Difficulty at Each Level

Level	Percentile	Sensitivity*	Specificity**
Advanced	60%	0.79	0.75
	50%	0.73	0.71
	25%	0.65	0.61
Intermediate	60%	0.67	0.45
	50%	0.7	0.62
	25%	0.69	0.46
Beginner	60%	0.67	0.36
	50%	0.88	0.3
	25%	0.64	0.65

*Percent of difficult words identified

**Percent of easy words identified

7.1.3 Discussion

We believe that the results show potential for this method of student modeling via language modeling. The best results, in terms of balancing sensitivity and specificity, were in the advanced set. This indicates that BERT models may model advanced students better than other groups. In terms of identifying the most difficult words while also identifying a decent amount (50% or more) easy words, the best performing model configuration was that selected for all students with a difficulty threshold of 60%. This may be because BERT models may represent advanced students best. Since the 60% threshold is the most difficult threshold, and since the all-students

Table 7.6: Model Configuration's Ability to Detect Difficulty vs Expert Ability to Detect Difficulty for the All Levels Data Set

Percentile	Sensitivity*	Specificity**	Sensitivity Difference	Specificity Difference
60%	0.83	0.59	N/A	N/A
50%	0.64	0.66	0.12	-0.26
25%	0.86	0.26	0.01	-0.62

*Percent of difficult words identified

**Percent of easy words identified

Table 7.7: Model Configurations at the 60th percentile

	All Students	Advanced Students	Intermediate Students	Beginner Students
Guesses	10	60	10	50
Answer Type	Exact and Synonyms	Exact and Synonyms	Exact only	Exact and Synonyms
Model	bert-large-uncased	bert-large-uncased	bert-large-cased-whole-word-masking	bert-base-multilingual-uncased

data set is skewed to the more advanced side, the all students data set at the 60% threshold represents a group that BERT models well. The final success of the model configurations comes from the comparison to the difficulty detection ability of the human instructors. Keeping in mind that the results of the human instructors are not specific to this particular data set and therefore represent only general ability to detect reading difficulty, we see that the model performs about the same as instructors in terms of sensitivity. While these are only preliminary results, since direct comparison is not possible, we believe this indicates the possibility of human level or near human level ability to detect reading difficulty. While specificity is lower, the number is not so low that too many easy words would be removed from the text. The model also has the advantage of being more consistent; as we have already observed, the human instructors did not agree well on which words would be difficult.

The results are worse for the data set including only beginning students than the intermediate, advanced or all students sets. The specificity was well below the threshold of 50% in two of the three configurations and the sensitivity was mediocre in two of the three as well. We believe this is because the questions in the beginner data set contained on average only about 2 responses. We also had only 170 questions for the beginner student group as opposed to 200 for the other groups. Thus, for the beginner set, we were unable to select the model that best fit the students' pattern of difficult and easy questions because the smaller data sets did not give any real pattern of how students perform.

We believe that the good results on the all students data set at the 60% threshold and advanced students data set shows that this method is useful for modeling student comprehension at more advanced levels. We are especially pleased with the results at the 60% threshold because this is the threshold chosen by expert opinion as a good threshold for choosing words that cause difficulty for students in reading. With more data, there may also be potential to more accurately identify good model configurations for lower levels. Considering that these models were trained only on language in general, and do not contain any data about language students and language learning, these results are interesting. These model configurations provide useful baseline information about users which can be exploited by new CALL projects without a need for large amounts of data or computing power to start with.

7.2 Output of the VSGP

Here, we examine output from the vocabulary set generation pipeline to give an idea of the types of vocabulary sets that are created for different students and different texts.

7.2.1 Results

We tested the VSGP with three sample students of differing levels and vocabulary sizes, and three sample students of differing levels and same vocabulary sizes. We used the model configurations for the students at the 60% level, except we did not include synonyms due to restrictions of the API used to get the synonyms. The model configurations with synonyms performed somewhat worse than with synonyms in terms of specificity, so we expect a to see slightly larger vocabulary lists than we would expect when using synonyms. The number of unknown words in each text and the number of words selected for study are shown in table 7.8 and table 7.10. The percent of unknown words in each text and the percent of unknown words that are *not* part of the study set are shown in tables 7.9 and 7.11. Finally, the overlap coefficient of the vocabulary sets for the students with the same vocabulary is presented in Table 7.12. The overlap coefficient shows what percentage of words in the smaller of two sets are shared in both sets. We would expect for students with the same vocabulary and different levels that the intermediate student's set is a subset of the beginner student's set (100% overlap) and that the advanced student's set is a subset of the intermediate student's set.

Table 7.8: Vocabulary Set Sizes for Students with Different Vocabulary Sizes

	Unknown Words: Fiction	Study Words: Fiction	Unknown Words: Non-Fiction	Study Words: Non-Fiction
Advanced	136	79	44	32
Intermediate	199	140	68	45
Beginner	342	243	129	98

Table 7.9: Words Chosen for Study as a Percent of Total Words for Students of Different Vocabulary Sizes

	% Unknown Words: Fiction	% Non-Study Words: Fiction	% Unknown Words: Non-Fiction	% Non-Study Words: Non-Fiction
Advanced	3.4%	1.5%	4.5%	1.2%
Intermediate	5.0%	1.5%	7.0%	2.4%
Beginner	8.7%	2.5%	13.3%	3.2%

Table 7.10: Vocabulary Set Sizes for Students with the Same Vocabulary Size

	Unknown Words: Fiction	Study Words: Fiction	Unknown Words: Non-Fiction	Study Words: Non-Fiction
Advanced	199	95	68	39
Intermediate	199	140	68	45
Beginner	199	169	68	58

Table 7.11: Words Chosen for Study as a Percent of Total Words for Students with the Same Vocabulary Size

	% Unknown Words: Fiction	% Non-Study Words: Fiction	% Unknown Words: Non-Fiction	% Non-Study Words: Non-Fiction
Advanced	5.0%	2.6%	7.0%	2.9%
Intermediate	5.0%	1.5%	7.0%	2.4%
Beginner	5.0%	0.8%	7.0%	1.0%

Table 7.12: Overlap Coefficient for Students with Same Vocabulary Size

	Fiction Overlap	Non-Fiction Overlap
Beginner vs Intermediate	0.91	0.98
Intermediate vs Advanced	0.93	0.92

7.2.2 Discussion

These results show exactly what we want to see for vocabulary study sets generated for students of differing levels. When controlling for vocabulary knowledge, we see that more advanced students receive fewer study words than less advanced students. The overlap coefficient is above 0.90 for all students of the same sized vocabulary and both texts, showing that, for the most part, the study sets of more advanced students are subsets of the study sets for less advanced students. At first glance, it may appear that the number of words to study is very large, however, we must take into account the student's vocabulary knowledge and language level, as well as the length of the text. Analysis of the percent of remaining words indicates that the vocabulary lists produced for most students are about the correct size. The lists can give students enough context to comprehend the text and enable them to guess the remaining words. Recall that between 95% and 98% of words are required for reading comprehension similar to that of a native speaker, with 98% indicating comfortable reading levels. Observing the percent of words that are unknown in the text versus the percent of words in the text left for implicit learning (the unknown words not chosen for study) we see great results. For the selected texts, the percent of unknown words was usually too high for good comprehension. In some cases, it was at or around the 5% threshold of unknown words which allows for good comprehension but makes reading difficult. For all students, the percent of words left to guess after studying the given vocabulary set was closer to 2% than 5%, and for most it was near 2%. For only one student, the beginner with mid-range vocabulary, remaining vocabulary might be worryingly low. However, the results from the model configuration analysis indicated that the beginner level at the 60% threshold includes too many easy words, which is likely the cause of the low remaining vocabulary. These study sets can be used for vocabulary pre-teaching or text glossing to support students in implicit language learning because they enable comprehension but do not remove so many words that the student has no words to guess in context.

Chapter 8

CONCLUSION

8.1 Limitations and Future Work

Here, we discuss known limitations of the research performed. We also suggest avenues for future work which address these limitations as well as push this project further

8.1.1 Cloze Test Data Set

The selection and testing of model configurations that represent different student groups was limited by the small amount of data available for students at lower levels. The beginner level especially was lacking in data. Conducting a new study to gather more responses would produce data that creates a better pattern of difficult and easy questions to which the model can be better configured. A set with more questions would also be help to better configure and compare the models more accurately. Additionally, more and better data on instructor ability to identify difficult words would improve comparisons to the expert baseline. Ideally, we would like to see that the instructors are responding to the same data set that the model is being tested with, and that they are asked to estimate difficulty for students at different levels using the threshold we are testing for.

8.1.2 Exploration of Other Language Models and Fine Tuning

While BERT models were chosen for study in this project, there are other models which could yield good results in modeling student reading comprehension. Of particular interest is the bi-directional LSTM ULMFiT model by (Howard and Ruder, 2018) or the modified transformer model by (Wang *et al.*, 2019) which uses an LSTM with a transformer. Since transformers have been observed to lose important sequential information (Wang *et al.*, 2019), these models might compare better with human students.

Another possible improvement might be to train models from scratch using only data that language students would be familiar with, as was done by Beinborn *et al.* (2015). Data for training a model such as this may come from existing learner corpora (corpora containing language produced by learners) or sources of language learning materials which may be gathered from the web. The learner corpora approach may not be an excellent approach, however, for several reasons. First, learner corpora may contain skewed data as often they contain information on particular topics, causing an over representation of words on those topics. Learner corpora also represent students' ability to produce language as opposed to their ability to receive language, so using learner corpora may misrepresent students' ability to read by modeling instead their ability to write. Using language learning materials to train a model may be a better approach but would require gathering large amounts of existing learning materials and sorting them by level. To relieve the burden of gathering so much data, it may be possible to use smaller pre-trained models and fine tune them with learning materials. Smaller pre-trained models may be better than the large models used in this paper because they would not be trained with huge amounts of native speaker data. This would prevent them from gaining too much native-level language knowledge before being fine-tuned for language learners. Smaller BERT models were recently released on March 11th, 2020 and may be a good avenue for future work (Google-Research, 2020).

8.1.3 GUI and User Study

This project focused specifically on the choice of vocabulary, providing a vocabulary selection pipeline that can be integrated into language learning applications. A GUI was not developed because existing applications already provide a wide variety of explicit vocabulary instruction techniques. We did not feel that developing yet another GUI for vocabulary study would provide any new insights unless it was to be the focus of the project. The lack of GUI makes conducting user studies difficult; thus, a user study was not performed. Ideally, we would like to compare student reading comprehension and implicit learning in vocabulary pre-teaching and glossing using vocabulary study sets generated by our pipeline versus existing language materials. A study such as this would take a similar form as those discussed in Chapter 4.

8.2 Achievements and Impact

Through analysis of vocabulary pre-teaching and glossing literature, it was shown that guessability is a good measure of a word's usefulness in vocabulary study sets. With this in mind, we created an automatic vocabulary study set generator which uses masked language modeling to model students' reading ability and select words from a given text that students cannot guess. We discovered model configurations that model advanced students well and those that show promise for students at other levels. Additionally, through our comparison with the data set containing students at all levels, we found that this method might be nearly as good at detecting difficult to comprehend texts for the average advanced student than human instructors and test makers. Our pipeline provides more consistency than humans do as well. With the model configurations that were chosen and tested, the VSGP produces vocabulary sets of reasonable size given the length of the text the set is created for and gives students at lower levels larger vocabulary sets than those at higher levels. We showed that the sets generated give students enough explicit vocabulary instruction to read comfortably while also leaving enough words left for implicit vocabulary acquisition. Unlike traditional study materials, which are created only for texts chosen by instructors and publishers, this method also has the potential to be applied to any text. The VSGP achieves all this without requiring much data about the student's language knowledge besides general language level and known vocabulary. This is vastly different from existing applications which require the user to begin learning before any adaption occurs and use results recorded by millions of students to determine difficulty level. Instead, the VSGP is designed to provide good results from the very first time a user uses it, whether the user is the application's first or millionth. It is open source and designed to be flexible, allowing it to be easily integrated into new systems. Ultimately, we answered our research questions, finding that masked language modeling can be used to model students' guessing ability, especially at the advanced level, and that modifying parameters such as number of guesses and the use of synonyms can improve results. This opens doors to future work on language modeling as student modeling, raising questions not on *if* language modeling can be used to model human language ability, but on how already interesting results of this method can be improved further.

REFERENCES

"spacy · industrial-strength natural language processing in python", URL `https://spacy.io/` (n.d.).

"Test your vocab - how many words do you know?", URL `http://testyourvocab.com/details` (n.d.).

"Online language learning market size, trends, opportunities forecast", URL `https://www.Verifiedmarketresearch.com/product/global-online-language-learning-market-size-and-forecast-to-2025/` (2019).

"Bert", URL `https://huggingface.co/transformers/model_doc/bert.html` (2020).

Alessi, S. and A. Dwyer, "Vocabulary assistance before and during reading.", Reading in a Foreign Language **20**, 2, 246–263 (2008).

Bahdanau, D., K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate", arXiv preprint arXiv:1409.0473 (2014).

Beglar, D., "A rasch-based validation of the vocabulary size test", Language testing **27**, 1, 101–118 (2010).

Beglar, D. and P. Nation, "A vocabulary size test", The language teacher **31**, 7, 9–13 (2007).

Beinborn, L., T. Zesch and I. Gurevych, "Difficulty prediction for language tests", URL `https://www.informatik.tu-darmstadt.de/ukp/research\_6/data/\\c\_tests/difficulty\_prediction\_for\_language\_tests/index.en.jsp` (2014a).

Beinborn, L., T. Zesch and I. Gurevych, "Predicting the difficulty of language proficiency tests", Transactions of the Association for Computational Linguistics **2**, 517–530 (2014b).

Beinborn, L., T. Zesch and I. Gurevych, "Candidate evaluation strategies for improved difficulty prediction of language tests", in "Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications", pp. 1–11 (2015).

Beinborn, L. M., *Predicting and manipulating the difficulty of text-completion exercises for language learning*, Ph.D. thesis, Technische Universität Darmstadt (2016).

Çetinavcı, B. M., "Contextual factors in guessing word meaning from context in a foreign language", Procedia-Social and Behavioral Sciences **116**, 2670–2674 (2014).

Chang, A. C.-S., "The impact of vocabulary preparation on l2 listening comprehension, confidence and strategy use", System **35**, 4, 534–550 (2007).

De Mulder, W., S. Bethard and M.-F. Moens, “A survey on the application of recurrent neural networks to statistical language modeling”, Computer Speech & Language **30**, 1, 61–98 (2015).

Devlin, J., M.-W. Chang, K. Lee and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805 (2018).

Dill, K. and P. Thomas, “The class of 2020 was headed into a hot job market. then coronavirus hit.”, URL `https://www.wsj.com/articles/the-class-of-2020-was-headed-into-a-hot-job-market-then-coronavirus-hit-11585486800?mod=hp_lead_pos7` (2020).

Duan, S., “Effects of enhancement techniques on l2 incidental vocabulary learning.”, English Language Teaching **11**, 3, 88–101 (2018).

Ennouamani, S. and Z. Mahani, “A comparative study of the learner model in adaptive mobile learning systems”, in “Proceedings of the 2nd International Conference on Networking, Information Systems & Security”, pp. 1–11 (2019).

Ertürk, Z. ÿ., “The effect of glossing on efl learners incidental vocabulary learning in reading”, Procedia-Social and Behavioral Sciences **232**, 373–381 (2016).

Felice, M. and P. Buttery, “Entropy as a proxy for gap complexity in open cloze tests”, in “Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)”, pp. 323–327 (2019).

Gamper, J. and J. Knapp, “A review of intelligent call systems”, Computer Assisted Language Learning **15**, 4, 329–342 (2002).

Gan, X., “Study on the effects of gloss type on chinese efl learners’ incidental vocabulary acquisition.”, Theory & Practice in Language Studies **4**, 6 (2014).

Google-Research, “google-research/bert”, URL `https://github.com/google-research` (2020).

Healy, H., “Dictionary use”, The TESOL Encyclopedia of English Language Teaching pp. 1–7 (2018).

Heil, C. R., J. S. Wu, J. J. Lee and T. Schmidt, “A review of mobile language learning applications: Trends, challenges, and opportunities”, The EuroCALL Review **24**, 2, 32–50 (2016).

Heilman, M. and M. Eskenazi, “Language learning: Challenges for intelligent tutoring systems”, in “Proc. Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, Proc. 8th Int. Conf. Intelligent Tutoring Systems”, (2006).

Hill, J. and R. Simha, “Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and google n-grams”, in “Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications”, pp. 23–30 (2016).

Horst, M., T. Cobb, T. Cobb and P. Meara, “Beyond a clockwork orange: Acquiring second language vocabulary through reading.”, Reading in a foreign language **11**, 2, 207–223 (1998).

Howard, J. and S. Ruder, “Universal language model fine-tuning for text classification”, arXiv preprint arXiv:1801.06146 (2018).

Jahangard, A., A. Moinzadeh and A. Karimi, “The effect of grammar vs. vocabulary pre-teaching on efl learners’ reading comprehension: A schema-theoretic view of reading”, Journal of English language teaching and learning **3**, 8, 91–113 (2012).

Kim, H.-S. and Y. Cha, “A comparative study of pre-reading activities on university students’ reading comprehension: Learning vocabulary vs. watching youtube”, Multimedia-Assisted Language Learning **20**, 1, 11–34 (2017).

Ko, M. H., “Glossing and second language vocabulary learning”, Tesol Quarterly **46**, 1, 56–79 (2012).

Kremmel, B. and N. Schmitt, “Vocabulary levels test”, The TESOL encyclopedia of English language teaching pp. 1–7 (2018).

Laufer, B. and G. C. Ravenhorst-Kalovski, “Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension.”, Reading in a foreign language **22**, 1, 15–30 (2010).

“LingQ for Schools”, “Lingq for schools”, URL `https://www.lingq.com/en/schools/` (n.d.).

Liu, H. and M. Cocea, “Semi-random partitioning of data into training and test sets in granular computing context”, Granular Computing **2**, 4, 357–386 (2017).

Loucky, J. P. and F. Tuzi, “Comparing foreign language learners’ use of online glossing programs”, International Journal of Virtual and Personal Learning Environments (IJVPLE) **1**, 4, 31–51 (2010).

Mart, Ç. T., “Guessing the meanings of words from context: Why and how”, International Journal of Applied Linguistics and English Literature **1**, 6, 177–181 (2012).

Meara, P., *EFL vocabulary tests* (ERIC Clearinghouse New York, 1992).

Mihara, K., “Effects of pre-reading strategies on efl/esl reading comprehension”, TESL Canada Journal pp. 51–51 (2011).

Mousavian, S. and H. Siahpoosh, “The effects of vocabulary pre-teaching and prequestioning on intermediate iranian efl learners’ reading comprehenstion ability”, International Journal of Applied Linguistics and English Literature **7**, 2, 58–63 (2018).

Nation, I. S. P., *Learning vocabulary in another language* (2013).

Nisbet, D. L., “Vocabulary instruction for second language readers.”, Journal of adult education **39**, 1, 10–15 (2010).

Park, G.-P., “The effects of vocabulary preteaching and providing background knowledge on l2 reading comprehension”, ENGLISH TEACHING (영어교육) **59**, 4, 193–216 (2004).

Pignot-Shahov, V., “Measuring l2 receptive and productive vocabulary knowledge”, Language Studies Working Papers **4**, 1, 37–45 (2012).

Radford, A., K. Narasimhan, T. Salimans and I. Sutskever, “Improving language understanding by generative pre-training”, URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf (2018).

Russel, S., P. Norvig *et al.*, *Artificial intelligence: a modern approach* (Pearson Education Limited, 2013).

Shei, C.-C., “Followyou!: An automatic language lesson generation system”, Computer Assisted Language Learning **14**, 2, 129–144 (2001).

Stahl, S. A. and M. M. Fairbanks, “The effects of vocabulary instruction: A model-based meta-analysis”, Review of educational research **56**, 1, 72–110 (1986).

Swanborn, M. S. and K. De Glopper, “Incidental word learning while reading: A meta-analysis”, Review of educational research **69**, 3, 261–285 (1999).

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention is all you need”, in “Advances in neural information processing systems”, pp. 5998–6008 (2017).

Wang, C., M. Li and A. J. Smola, “Language models with transformers”, arXiv preprint arXiv:1904.09408 (2019).

Zaidi, A. H., R. Moore and T. Briscoe, “Curriculum q-learning for visual vocabulary acquisition”, arXiv preprint arXiv:1711.10837 (2017).

Zandieh, Z. and M. Jafarigohar, “The effects of hypertext gloss on comprehension and vocabulary retention under incidental and intentional learning conditions.”, English Language Teaching **5**, 6, 60–71 (2012).

Zimmer, B., “Science of learning”, URL `https://www.vocabulary.com/educator-edition/Vocabulary.com-ScienceofLearning.pdf` (2015).

Zweig, G. and C. J. Burges, “A challenge set for advancing language modeling”, in “Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT”, pp. 29–36 (Association for Computational Linguistics, 2012).