# Assessing receptive vocabulary using state-of-the-art natural language processing techniques

Scott Crossley and Langdon Holmes
Vanderbilt University, United States



Semantic embedding approaches commonly used in natural language processing such as transformer models have rarely been used to examine L2 lexical knowledge. Importantly, their performance has not been contrasted with more traditional annotation approaches to lexical knowledge. This study used NLP techniques related to lexical annotations and semantic embedding approaches to model the receptive vocabulary of L2 learners based on their lexical production during a writing task. The goal of the study is to examine the strengths and weaknesses of both approaches in understanding L2 lexical knowledge. Findings indicate that transformer approaches based on semantic embeddings outperform linguistic annotations and Word2vec models in predicting L2 learners' vocabulary scores. The findings help to support the strength and accuracy of semantic-embedding approaches as well as their generalizability across tasks when compared to linguistic feature models. Limitations to semantic-embedding approaches, especially interpretability, are discussed.

**Keywords:** natural language processing, corpus linguistics, lexical knowledge, Doc2Vec, BERT, word-embeddings, lexical annotations

## 1. Introduction

Exploring lexical knowledge on the part of second language (L2) learners can provide researchers and practitioners with information about learners' cognitive development, learners' lexical processing, and assessment principles. L2 lexical knowledge has been investigated through a variety of methods including survey items, vocabulary assessments, behavioral studies, and corpus analysis (Berger et al. 2019; Kyle et al., 2018; Lemhöfer et al., 2008; Milton, 2009). Over time, corpus approaches to understanding L2 lexical knowledge, especially when combined

https://doi.org/10.1075/jsls.22006.cro | Published online: 20 September 2022
*Journal of Second Language Studies* ISSN 2542-3835 | E-ISSN 2542-3843

with natural language processing (NLP) techniques, have become common place (Crossley, Skalicky, et al., 2019; Kyle & Crossley, 2016). Such approaches rely on using NLP techniques to automatically annotate learner corpora for specific lexical features such as word length, word frequency, or word concreteness and then make associations between these annotations and variables related to lexical knowledge include vocabulary test scores (Hashimoto & Egbert, 2019), human ratings of vocabulary knowledge (Crossley, Salsbury, et al., 2011a, 2011b), or student grade level (Kerz et al., 2021). Additionally, lexical annotations can be used to track lexical development over time (Crossley & Skalicky, 2019).

The majority of NLP annotation techniques measure lexical knowledge at the word or phrasal level. Early studies focused on word length and word frequency (Conrad, 2005; Grant & Ginther, 2000) while later studies began to annotate words to include features such as phonological neighbors, lexical response time, number of word associations, age of acquisition, and word concreteness (Kyle & Crossley, 2015; Kyle et al., 2018). Phrasal annotations were also introduced that measured associational strength, frequency, and range (Garner & Crossley, 2018; Garner et al., 2018). Annotations of lexical items has remained state of the art for measuring L2 lexical items since the late 1990s with studies demonstrating the strength of these features to predict vocabulary knowledge and development (Crossley & Kyle, 2022; Laufer & Nation, 1995; Meurers, 2012, 2021). Such studies provided researchers and practitioners with a wealth of knowledge about how the lexicon develops in L2 learners, how words are processed and stored, and how assessments can be improved and validated. However, the lexical annotations described above generally only examine word properties and not word meaning (i.e., semantics).

Outside of L2 research, computational linguists have continued to refine natural language processing techniques, and research has advanced from simply annotating linguistic features found in language samples to modeling language semantics using continuous vector representations for words derived from large data sets. At a practical level, such approaches examine the distributional representations of words in texts with the understanding that words with similar meanings tend to occur in similar contexts. These approaches embed words in a vector space to compute semantic relationships among words. Seminal work on semantic vector representations date back to the late 1990s with the development of latent semantic analysis (LSA, Landauer et al., 2007), which uses dimensionality reduction techniques to condense a large word by document co-occurrence matrix derived from a corpus of texts into a lower dimensional space. LSA transforms the words of a document-term matrix into a vector, often of length 300. It allows for the semantic similarity between two words within a corpus to be measured by calculating the cosine of the angle of the two words' vectors. LSA

and related approaches have been widely applied in information retrieval, sense disambiguation, and topic modelling. The simple word-embedding approaches found in LSA have been augmented in Word2vec models that generate static representations of words that weigh distributions based on surrounding words (as compared to an entire text). Word2vec uses a shallow neural network to develop these representations (Mikolov et al., 2013). Even newer embedding approaches based on transformer models use much larger neural networks with an architecture called attention (Vaswani et al., 2017). These transformer models develop contextual representations of words, such that the same orthographic form will have a different embedding depending on its context. They also require magnitudes more training data and processing power. Nonetheless, transformer models (also known as large language models) have become state of the art in NLP due to their ability to outperform other methods in a variety of tasks.

However, semantic embedding approaches, including Word2vec models and especially transformer models, have rarely been used to examine L2 lexical knowledge. One would expect that embedding approaches, which model the underlying semantics of a language, would perform well at predicting or classifying L2 lexical knowledge, especially when compared to NLP techniques that annotate lexical features of texts (e.g., frequency, phonological neighbors, lexical response times, and word concreteness). Thus, this study investigates the predictive strength of lexical annotations and embeddings derived from L2 student writing to model receptive vocabulary scores. The goal of the study is to examine the strength of using semantic models of lexical production compared to more traditional lexical annotations of production to better understand L2 lexical knowledge. Our hypothesis is that models of lexical production that are based on language semantics will outperform models based on lexical feature.

## 2. Literature review

### 2.1 Lexical knowledge

Lexical knowledge is generally understood through global trait models that have traditionally examined two dimensions: (1) breadth of lexical knowledge or lexical size and (2) depth of lexical knowledge which measures the manner and degree to which known words are organized (Meara, 1996, 2005a; Read, 1998). Breadth is generally operationalized through lexical diversity (i.e., the variety of words produced) or word frequency (i.e., how frequent a word is within a language). Depth is operationalized to include any measurements that examine the strength of networks and/or interactions of links among words (Moghadam et al., 2012) includ-

ing semantic associations and the semantic representations of the word (Nagy & Scott, 2000).

The two dimensions bifurcate over the notion of knowledge and whether it is related to knowledge of the entire lexicon (breadth) or is related to the strength of knowledge for individual words (depth). There are a number of problems with this binary approach. First, breadth and depth dimensions ignore properties related to core lexical knowledge like word concreteness, familiarity, and imageability (Crossley & Skalicky, 2019) which allow for quicker lexical processing or retrieval (Crossley, Salsbury, et al., 2011a, 2011b; Meara, 2005b). Second, it is also not always clear which lexical features should be assigned to which of the two dimensions. For instance, word frequency has historically been considered a measure of breadth of knowledge because learners that produce more infrequent words should have a larger vocabulary. However, the distributional properties of words based on frequency also strengthen connection between words and meanings (Ellis, 2002). These connections overlap strongly with depth of lexical knowledge (i.e., the organization of words in the lexicon).



### 2.2 Measuring L2 lexical knowledge

There are a number of ways to explore lexical knowledge in L2 learners. Traditional approaches have depended on lexical assessment such as vocabulary size tests, translation or elicitation, and word association tasks (Milton, 2009). Behavioral methods that measure L2 learners response times to linguistic stimuli are also commonly used to assess L2 knowledge (Berger et al., 2019; Crossley & Skalicky, 2019; Lemhöfer et al., 2008). More recently, the use of NLP driven annotations based on L1 norms[1] to examine lexical knowledge have become common (Crossley, Salsbury, & McNamara, 2009, 2010; Morris & Cobb, 2004).

NLP annotations of lexical features are able to adequately measure both breadth and depth features of the lexicon as well as core lexical properties. Common breadth features measured using NLP annotations include lexical variety

1. Arguments have been made in favor of using NLP annotations based on L2 corpora and L2 learner judgments (Oretga, 2016). However, L2-based annotations are not available for many lexical features and properties (e.g., concreteness, word naming). For those features in which L1- and L2-based annotations are available, research does not clearly favor one annotation approach over the other. For example, Monteiro et al. (2020) reported that frequency metrics based on an L2 corpus outperformed L1-based frequency metrics in predicting L2 writing quality (although both were predictive); however, a follow up study (Monteiro, 2020) found no differences. Similarly, Crossley, Skalicky, et al. (2019) reported that L2-based frequency metrics were not stronger predictors of L2 development compared to L1-based frequency metrics (although both were equally predictive).

measures using type-token ratio counts and word frequency measures based on a variety of corpora like the British National Corpus (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA, Davies, 2010). Depth measures include features related to hypernymy and polysemy derived from WordNet (Fellbaum, 1998), word naming and lexical decision time scores taken from the English Lexicon Project (Balota et al., 2007), and measures of word association strength derived from corpora like the BNC or COCA (Garner et al., 2018). However, NLP annotations of lexical features do not generally measure semanticity in language (with WordNet derived features being the exception).

In contrast to NLP annotations of lexical features, embedding models produce semantic representations of words and texts, but they do not explicitly measure features related to depth of lexical knowledge nor do they measure breadth of lexical knowledge or core lexical properties. For instance, LSA, Word2vec, and transformer models use word vectors to measure semantic similarity between words and text segments. These techniques are quite good at uncovering semantically related words, predicting next words in sentences, and examining the overall semantic content of a text. However, embedding models do not provide analytic information about word frequency, the processing times for words, or core properties for words such as concreteness.

Much research has explored the use of lexical annotations to assess L2 lexical knowledge. In the area of lexical acquisition Crossley, Salsbury, and McNamara (2009) and Crossley, Salsbury, and McNamara (2010) used NLP annotations of lexical sophistication and diversity calculated by Coh-Metrix (Graesser et al., 2004) to longitudinally investigate various components of second language (L2) lexical development (e.g., hypernymic word relations and polysemy). More recently, Crossley, Skalicky, et al. (2019) investigated the relationship between lexical salience, lexical frequency, and language development over time using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., 2018). Other researchers have used web-based tools such as VocabProfile (Cobb, n.d.) to explore lexical acquisition in response to specific activities. For instance, Zaytseva et al. (2019) measured written and oral lexical production before and after a 3-month study abroad experience using measures related to diversity, sophistication, density and accuracy. They found that studying abroad led to increased vocabulary in written samples more so than spoken samples, especially in terms lexical diversity. Sundqvist (2019) found that extramural gaming improved measures of both productive and receptive vocabulary use.

NLP annotations of lexical features have also been used to predict speaking proficiency in learner corpora. Lu (2012), for example, used the Lexical Complexity Analyzer to successfully model the relationship between speaking proficiency and indices related to lexical density, diversity, and sophistication. Biber et al.

(2016), used the Biber Tagger (Biber, 1988) to predict the relationship between a wide range of lexicogrammatical features and speaking quality scores in a large corpus of oral standardized test responses. Studies have also modeled lexical aspects of speaking proficiency using the Tool for the Automatic Analysis of Lexical Sophistication (TAALES). For instance, Berger et al. (2019), examined the relationship between judgements of speaking quality for a large corpus of L2 speech and lexical characteristics including contextual diversity and psycholinguistic word properties. Saito (2020) recently corroborated Kyle and Crossley (2015)'s finding that L2 oral proficiency can be predicted with collocational qualities captured by n-gram indices.



In comparison to NLP lexical annotations, embedding approaches that focus on semantic relations like LSA, Word2vec, and BERT have seen relatively little attention outside of automated essay scoring (see Ke & Ng, 2019, for a review). In perhaps the earliest study, Crossley, Salsbury, & McNamara (2010) used LSA to examine the development of semantic networks in L2 speakers finding that semantic similarity scores among words increased as a function of time studying English. In a more recent study, Lu and Hu (2021) explored contextual embeddings from BERT as a means of sense disambiguation and found that augmenting existing measures of lexical sophistication with sense-aware frequency counts improved predictive power for L2 English writing quality. Sun and Lu (2021) utilized a vector space model (fastText, Bojanowski et al. 2017) to extrapolate psycholinguistic dimensions of unseen words from smaller sets of labelled lexemes (i.e., psycholinguistic databases). They then measured variation within these psycholinguistic properties in a large, longitudinal corpus (EFCAMDAT, Huang et al., 2018) and found that the tested word properties can be inferred from their positions in a vector space model. Monteiro (2020) developed L2 semantic context indices from the EFCAMDAT corpus (Huang et al., 2018) using LSA (Landauer & Dumais, 1997) and Word2vec (Mikolov et al., 2013) and reported that L2 semantic indices were significantly predictive of L2 writing and how fast L2 users judged a word to be a pseudoword or a real word. While work in this area is scant, existing research suggests that there are strong relationships between embeddings and analytic measures of lexical proficiency.

## 3. Current study

The purpose of the current study is to compare the predictive strength of NLP lexical annotations to semantic embeddings to model vocabulary knowledge in L2 learners. To demonstrate the state of the art in semantic embeddings, we compare two semantic approaches (Word2vec embeddings, and transformer models)

to NLP lexical annotations (e.g., word frequency, word associations, word concreteness). The goal is to examine how well automated approaches that incorporate lexical semanticity perform compared to non-semantic features that have been commonly used in previous NLP studies. The research questions that guide this study are:

1. Are there differences in accuracy for models predicting L2 receptive lexical knowledge between NLP lexical annotations and semantic embedding approaches?
2. What insights can lexical annotations and semantic embedding models provide about L2 lexical knowledge?



## 4. Method

### 4.1 Corpus

We used the International Corpus Network of Asian Learners of English (ICNALE, Ishikawa, 2013) for this analysis. ICNALE (Ishikawa, 2013) includes around 10,000 topic-controlled L2 writing and speech samples produced by college students and graduate students in ten countries/regions in Asia, namely China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Taiwan, and Thailand. ICNALE comprises four modules: Spoken Monologue, Spoken Dialogue, Written Essays, and Edited Essays. For this study, we used the ICNALE Written Essays which comprises 200- to 300-word essays written by each participant on two topics: part-time jobs for college students and a ban on smoking in restaurants. ICNALE includes writing samples for 2,600 English language learners, with corresponding receptive vocabulary scores as calculated using the English vocabulary size test (VST; Nation & Beglar, 2007). These scores tap into learners' receptive lexical proficiency, which is an important element of L2 acquisition (David, 2008). Receptive vocabulary is a strong predictor of speaking proficiency (Koisumi & In'nami, 2013) and potentially a more robust measure of lexical knowledge than productive vocabulary (Webb, 2009).

We used data from one prompt (ban on smoking, SMK) to evaluate models receptive vocabulary knowledge. One essay was removed for technical reasons and, thus, our final corpus consisted of 2,599 essays written by 2,599 L2 English learners from ten countries (see Table 1 for breakdown of essays by country). The remaining ICNALE prompt (part time job, PTJ) comprises 2,600 essays written by the same learners as the SMK prompt. These essays were used to augment the data for the embedding models during training, but they were not used during the development of the models to predict receptive vocabulary knowledge. The PTJ

essays were used in training the embedding models to ensure the models had sufficient data for successful training.

**Table 1.** Country information

| Country | Number essays | Linguistic distance |
|---|---|---|
| China | 400 | 1.5 |
| Hong Kong | 100 | 1.25 |
| Indonesia | 200 | 2 |
| Japan | 400 | 1 |
| Korea | 300 | 1 |
| Pakistan | 200 | 1.75 |
| Philippines | 200 | 2 |
| Singapore | 199 | 1.5 |
| Thailand | 400 | 2 |
| Taiwan | 200 | 1.5 |



### 4.2 Receptive vocabulary knowledge

Participants' receptive vocabulary levels in ICNALE were assessed using an English vocabulary size test (VST) prior to writing their essay submissions. The VST included fifty test items in the 1000–5000 word levels (ten items per 1000 word band) from the monolingual version of VST (14,000 words; Nation & Beglar, 2007), which was delivered in a spreadsheet format (Ishikawa, 2013). When multiplied by 100, test scores represent the approximate number of word families known by an individual. For example, a VST score of 33 suggests receptive knowledge of approximately 3,300 words.

### 4.3 Lexical annotations

The Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle et al., version 2.8) was used to measure the lexical features of each text.[2] Each of the following features were entered into a model to predict the participant's VST scores. Lexical features to examine depth and breadth of lexical knowledge and core lexical items were selected based on previous studies that indicated their strength in explaining L2 lexical development (Berger, Crossley, & Kyle, 2019; Berger, Crossley, & Skalicky, 2019; Crossley & Skalicky, 2019; Mostafa et al.,

2. TAALES is freely available at linguisticanalysistools.org

2021). Indices were computed for both content and function words, and all measures selected were derived from either experimental studies, survey responses, or corpus-based statistics.

*Age of acquisition*

Age of acquisition (AoA) indices approximate the average age that native English speakers learn a word. Words that are acquired later in life can be considered more sophisticated (e.g., *repudiate*) than words learned earlier in life (e.g., *dog*). These indices are based on norms reported by Kuperman et al. (2012) and are computed with lemmatized word forms.

*Concreteness*

Concreteness measures the tangibility of a word's referent. More concrete words, such as *tree* and *table,* refer to physical, perceptible objects. Less concrete words, such as *thought* and *ethical,* refer to abstract concepts. Words with lower concreteness are considered more sophisticated. Scores were calculated using the concreteness norms reported by Brysbaert et al. (2014) and are based on word lemmas.

*Word familiarity*

Word familiarity measures how likely it is that a person would know the word. Well known words that are more commonly used, such as *breakfast, television,* and *book,* would have higher familiarity. Less commonly used words that may not be known, such as *egress* and *encephalon,* would have lower familiarity. Scores were calculated using the 4,943 lemmas of the MRC Psycholinguistic Database (Wilson, 1988).

*Word meaningfulness*

Word meaningfulness measures the extent to which a word is related to other words. It is based on human judgements of how related a target word is to other words. Words that are less broadly meaningful like *chagrin* and *astuteness* will activate fewer words. On the other hand, a word like *cup* will be more broadly meaningful, activating related words such as *soup, saucer,* and *coffee*. Words with lower meaningfulness are considered more sophisticated. Scores were calculated using the 2,644 lemmas of the MRC psycholinguistic database (Wilson, 1988).

*Lexical response times*

Lexical response times measure the response time in milliseconds it takes for a human participant to respond to a lexical stimulus. A single norm was included from Balota et al. (2004), who reported participant's response time when deciding

whether a stimulus was a real word or a non-word. Longer lexical response times indicate more sophisticated words, such as *tangential*. Shorter lexical response times indicate less sophisticated words such as *happy*. Scores were calculated using raw, unlemmatized word forms.

*Word associations*

Word associations measure the number of stimuli words that elicit the target word in a word association task. Words with more associations, such as *love* (elicited by 181 different stimuli), are more readily accessible then words with fewer associations, such as *bride* (elicited by 6 stimuli). Words with fewer associations are considered more sophisticated. Scores were calculated using the associations norms for 5,019 stimulus words and 10,470 response words reported in Nelson et al. (2004) and found in the University of South Florida (USF) database.

*Phonological distance*

Phonological distance measures how similar in sound a word is to other words. This is operationalized as the Levenshtein distances between a word and its 20 nearest phonological neighbors, where Levenshtein distance is the smallest number of insertions, deletions, and/or replacements that transform the target word into one of its neighbors. Words that are more distant from their phonological neighbors, such as *cardiovascular, conspicuous,* and *calisthenic*, are considered more sophisticated than words with more phonologically similar neighbors, such as *fairies*, *wedded*, and *banter*. Scores were calculated using the phonological distance norms reported by Balota et al. (2004) and are based on raw, unlemmatized word forms.

*Word frequency*

Word frequency measures the frequency of words in a reference corpus. For this study, the SUBTLEXus corpus (Brysbaert & New, 2009) was selected as the reference. SUBTLEXus is a 51-million-word corpus of American film and television subtitles. Frequencies extracted from corpora that reflect spoken language tend to align more closely with psycholinguistic norms developed in clinical settings (Paetzold & Specia, 2016).

*Collocation strength*

Collocation strength measures assess the degree of association between two words. The specific measure of association strength selected was Delta-P, which is defined as the adjusted probability of a second word occurring, given the preceding word. The Delta-P measure was calculated for adjacent words (bigrams) and utilizes the spoken section of the Corpus of Contemporary American English

(Davies, 2010). Bigrams that exhibit weak association, such as *interested for* and *discovered around,* can be considered less sophisticated than more strongly associated bigrams, such as *interested in* and *discovered that.*

*Contextual distinctiveness*

Contextual distinctiveness measures the amount of information a word provides about its context. The specific measure selected was McD (McDonald & Shillcock, 2001), which is based on relative entropy or Kullback-Leibler divergence. It measures the distance between *Q,* the probability distribution of all possible word contexts in a corpus, and *P*, the probability distribution of word contexts for the target word in the same corpus. If *P* provides little information about its context (less distinctive), it will be less distant from *Q*. If *P* provides more information about its context (more distinctive), it will be more distant from *Q*. The more greatly these two probability distributions differ, the more contextually distinct the word. Less distinctive words such as *today* are used in a variety of contexts. As a result of their contextual flexibility, they are considered less sophisticated. More distinctive words such as *lone* provide more information about their context and are considered more sophisticated. Scores were calculated using the 8,000 lexemes reported by McDonald and Shillcock, whose work was based on the spoken BNC (2007).

## 4.4 Semantic embedding

We used both static and contextualized embedding approaches to extract semantic information from each text. This information was then used to model participants' VST scores. We implemented both approaches in the Python programming language.[3] Each approach is discussed below.

*Doc2vec*

Doc2vec is based on Word2vec (Mikolov et al., 2013), which is a method to represent semantic information as a vector of numbers that represent the distributional probabilities of words. Word2vec uses a shallow neural network with a single hidden layer to learn the probability distributions of words in a corpus. Using a continuous bag of words (CBOW) implementation, the neural network learns to predict an unknown center word given the vector representations of the surrounding words. At each pass, the values of the vectors are slightly adjusted so that they perform better. Since a single vector is learned for all occurrences

3. The Python scripts used to develop these models are available at github.com/langdonholmes/lexical_analysis

of the same word type, regardless of its syntactic function or the semantic sense in which the word is used, the embeddings are considered static. Doc2vec (Le & Mikolov, 2014) is an extension of Word2vec in which a single vector is trained to represent a whole paragraph or an entire document. The Doc2vec implementation used in this study learns word representations and document representations in parallel and is conceptually similar to the Word2vec CBOW method with an additional paragraph vector included.[4] At each pass, the shallow neural network attempts to predict an unknown center word given the vector representations of the surrounding words and a vector representation of the paragraph. Because document vectors make available information about the types of words in a document, they should be predictive in determining a writer's lexical proficiency.

One problem with Word2vec and Doc2vec is that the semantic representations that result from the neural network model are difficult to explain and interpret. The complexity of the neural network and opaqueness of the hidden layers that inform the network mean that the decision processes that lead to the vector representations are unavailable. Thus, if the model is biased or outdated, is based on unjust decisions, or is based on incorrect assumptions, that data is not available for human interpretation.

For this study, we trained document embeddings with the Gensim (Rehurek & Sojka, 2010) library in Python. We used NLTK's (Bird et al., 2009) word space tokenizer, and included all tokens produced by the tokenizer. We optimized two important hyperparameters: vector size and epochs. Vector size determines the dimensionality of the vector space, and in more practical terms, how many values will constitute a document's vector representation. While the original paper (Le & Mikolov, 2014) used a vector length of 400, different vector lengths have been shown to work better in different contexts (Lau & Baldwin, 2016). The epochs setting determines how many times the network is trained on the entire dataset. The original (Le & Mikolov, 2014) paper used 10–20 epochs, which means that the model saw each text 10–20 times during training. In our instantiation, we searched across 50, 100, 200, 400, 600, 800, 1000, and 1200 epochs. Epoch settings within this range have been shown to be appropriate for smaller datasets (Lau & Baldwin, 2016).

In order to provide an accurate evaluation of each method's performance, we partitioned our data into training, validation, and test sets. The SMK prompt was divided into development and test sets following an 80/20 split. The development

4. Lau and Baldwin (2016) recommend 'seeding' Doc2vec algorithm with pre-trained word vectors. However, we could locate no pre-trained vectors specific to L2 production. Thus, we used the Doc2vec algorithm as implemented in the original paper (Le & Mikolov, 2014). In this implementation, word vectors are learned alongside document vectors during training.

set was augmented with 2,600 additional essays written in response to the PTJ prompt. The augmented development set was then divided into training and validation sets, also following an 80/20 split. The training and validation sets were used during hyperparameter optimization and training. The test set was reserved only for evaluating the model's performance. We found that length 100 vectors trained over 400 epochs produced the best results in our task with our training data (see Figure 1). We trained document vectors on our training set using these hyperparameters.

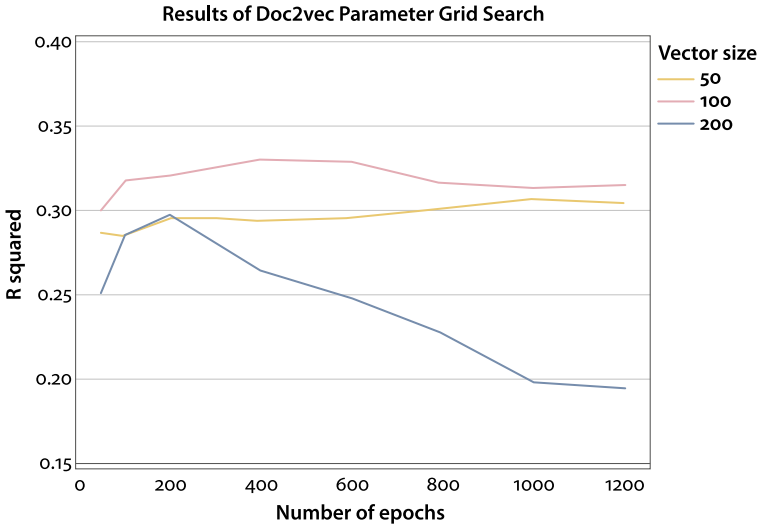


**Figure 1.** Results of grid search for optimal Doc2vec hyper parameters

### *Transformers*

Transformer models differ from Word2vec models because they use neural networks with multiple hidden layers and include an attention mechanism. Transformer models like those found in Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) take into consideration the order in which words appear (i.e., *love and hate* would be represented differently than *hate and love*) and have attention mechanisms which allow input weights to be based on importance in a task. Whereas Word2vec takes a small, predefined window of context words into consideration, BERT's self-attention mechanism allows it to dynamically choose which words are important for its calculations from a wider context window (the full length of the input). Contextual representations allow BERT to extrapolate differences between the uses of the word *bank* in the sentences *The man robbed the bank* and *The man sat on the river bank*. In Word2vec,

*bank* would have a single vector representation based on these two sentences while BERT would have different representations for each use of the word.

Like the Doc2Vec model, BERT models are based on neural networks. However, BERT neural network models include millions of parameters that interact in complex ways, making it even more difficult to fully explain or interpret what the model is doing in each pass when compared to Doc2vec models. Thus, like Doc2vec, the semantic representations that result from BERT do not lend themselves to interpretation, and it is difficult to assess whether the decision process made by the model is appropriate and unbiased. Another problem that arises from using transformer models like BERT is the immense cost associated with pretraining a language model because of the size of the data, the layers of the neural network, the bidirectional nature, and the attention mechanisms. Thus, unlike Doc2vec, it is common practice to use pre-trained language models and finetune them for different tasks. BERT, which we use in this study, was pre-trained with a masked language modelling task on a corpus comprising 2.5 billion words from Wikipedia and 800 million words from the BooksCorpus (Devlin et al., 2019). Finetuning works by influencing the pretrained BERT's weights and biases a small amount to leverage knowledge about diverse language-related tasks. In comparison to training a model from scratch, finetuning can be performed with significantly less data and processing power. In practice, finetuning involves providing the pre-trained transformer model with labelled training data that are specific to the downstream task. The weights and biases are influenced through a procedure called backpropagation. One caveat with finetuning is that it is not feasible to alter the tokenization scheme. As a result, we used the same WordPiece tokenizer that was utilized during BERT's pre-training.

In order to assess the utility of large language models to predict lexical proficiency (as measured by VST), we finetuned the base, uncased version of BERT (available through Huggingface Transformers, Wolf et al., 2020) to predict VST score of the writer (after scaling VST scores to a floating point value between −1 and 1). Huggingface includes a standardized 'bert for sequence classification' model, which works by adding a sequence classification 'head' on top of the pretrained BERT language model. The sequence classification head adds two layers to the neural network: a dropout layer and a linear feed forward layer. The linear layer was initialized with random values. During training, we backpropagate on the entire network, including the additional linear layer.

The most important hyperparameters to set when optimizing the BERT model are the learning rate, batch size, and number of epochs (Devlin et al., 2019). Learning rate determines how dramatically the model adjusts to the new data. If the model learns too quickly, it may 'forget' some of what it has already learned. If it learns too slowly, it will fail to adapt to the data. Batch size determines how

many documents are processed in parallel. Number of epochs, as with Doc2vec, determines how many passes are made over the training set. While deep learning libraries generally provide sensible defaults for the learning rate and all other hyperparameters, it is best practice to empirically determine the correct learning rate for each task and dataset (Kohavi & John, 1995). We followed the recommendations of Devlin et al. (2019) and performed an exhaustive grid search over three learning rates (2e-5, 3e-5, and 5e-5) and three epoch settings (2, 3, 4). We selected the lower recommended batch size (16) apriori due to memory limitations on our processing unit. The hyperparameters which produced the best model were selected for finetuning (see Figure 2).



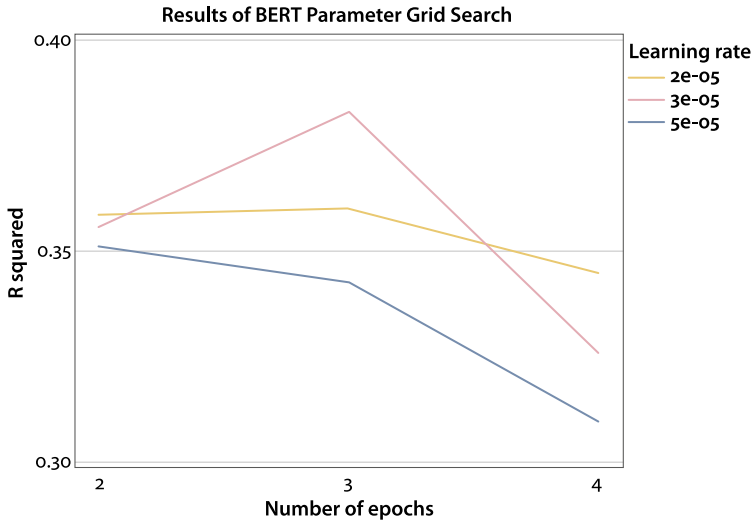


**Figure 2.** Results of grid search for optimal BERT hyper parameters

## 4.5 Statistical analysis

Linear models to predict student VST scores were developed using linguistic features, Doc2vec, and BERT. For the linguistic features and Doc2vec models, we constructed linear models in R (R Core Team, 2022) using the CARET package (Kuhn, 2008). For the linguistic features model, we used a training and test set that matched those used in the Doc2vec and BERT models to develop a linear model. The linguistic feature model developed from the training set was then applied to the held-out test set. For the Doc2vec model, we developed a linear model using only the vectors inferred from the test data. For both the linguistic features model and the Doc2vec model, estimates of model accuracy were reported using summary statistics including root mean squared error (RMSE)

and mean absolute error (MAE) between the observed and modeled holistic and VST scores. R-squared (R2) is also reported and can be used to examine the amount of variance explained by the developed model. Variable importance was explained using the varImp function in CARET. Variable importance in varImp is based on the absolute value of the *t*-statistic for each model parameter used.

Multi-collinearity between variables (i.e., variables that are highly collinear and potentially measuring the same construct) can make interpreting variable importance in linear models difficult. Thus, prior to developing our models, we calculated correlations among the derived linguistic features and Doc2vec vectors. If two or more variables correlated at $r > .799$, the variable(s) with the lowest correlation with the VST scores was removed and the variable with the higher correlation was retained.

As noted earlier, the BERT model was outfitted with a linear head as part of its finetuning process, which allowed the finetuned BERT model to output its predictions directly. Using the linear head of the BERT model, as compared to extracting BERT embeddings and training a linear model separately, is a more parsimonious use of data and the BERT model, since the linear layer is trained during model finetuning. It is also the standard method of performing regression and classification tasks with pre-trained language models. In practice, the finetuned BERT model directly outputs a predicted VST score for each input text. Before finetuning, VST scores were scaled to floating point values in the range (−1.0, 1.0). BERT was then finetuned on the augmented (SMK training + PTJ) development set. The performance of the finetuned model was evaluated on the held out test set (20% of the SMK prompt), using the same summary statistics RMSE, MAE, and R2.



## 5. Results

### 5.1 Lexical annotations model

Correlations indicated that our SUBTLEXus frequency measure was strongly collinear with the Kuperman age of acquisition measure. Because age of acquisition reported a higher correlation with VST scores, the frequency measure was removed from the linear model. A linear model for the training set using the remaining nine lexical annotations reported $RMSE = 7.993$, $MAE = 6.419$, $r = .415$, $R^2 = .172$, indicating that the linguistic features model explained 17% of the variance in the VST scores. The relative importance metrics indicate that the strongest predictors of VST scores were word meaningfulness followed by word familiarity

and age of acquisition. The weakest predictors were related to contextual diversity and word association (see model parameters summarized in Table 2).

**Table 2.** Parameters for linguistic features model

| Feature | Co-efficient | Variable importance |
|---|---|---|
| Intercept | 33.504 | |
| Word meaningfulness (MRC) | −2.292 | 4.821 |
| Word familiarity (MRC) | −0.994 | 1.950 |
| Age of acquisition (Kuperman) | 1.235 | 1.715 |
| Lexical decision response time | 0.495 | 0.956 |
| Word concreteness (Brysbaert) | −0.337 | 0.825 |
| Word associations (USF) | 0.334 | 0.694 |
| Phonological neighbors | −0.180 | 0.322 |
| Contextual distinctiveness | 0.083 | 0.196 |
| Collocation strength (COCA spoken DP) | 0.045 | 0.118 |



## 5.2 Doc2vec model

As expected, none of the length 100 Doc2vec vectors were multicollinear so all vectors were entered into the linear model. A linear model for the test data using the Doc2vec vectors reported $RMSE = 8.641$, $MAE = 6.863$, $r = .410$, $R^2 = .168$, indicating that the Doc2vec model explained 17% of the variance in the VST scores. Because the vectors in the Doc2vec model are not interpretable, we do not report their co-efficients or their variable importance.

## 5.3 BERT model

The finetuned BERT model predicted VST scores scaled to the range (−1.0, 1.0). In order to make these results comparable to our other models, the predicted values were inverse scaled back to the original VST unit scale. The finetuned model, when applied to the test set, reported $RMSE = 7.127$, $MAE = 5.438$, $r = .567$, $R^2 = .321$, indicating that the finetuned BERT model explained 32% of the variance in the VST scores.

### 5.4 Comparisons between models

We used Fisher r-to-z transformations to assess the significance of the difference between the correlation coefficients reported for the linguistic feature, Doc2vec, and BERT Models for the test sets (see Table 3). The results indicated that the BERT model outperformed the linguistic features model and the Doc2vec model. There were no differences between the linguistic features model and the Doc2vec model.

**Table 3.** Fisher r-z transformations between models

| Models | z | p |
|---|---|---|
| Linguistic –Features – Doc2vec | 0.22 | > .050 |
| Linguistic –Features – BERT | 7.26 | < .001– |
| Doc2vec – BERT | 7.48 | < .001 |



## 6. Discussion

This study examined various NLP approaches to modeling receptive vocabulary in L2 learners including state-of-the-art semantic embedding approaches. Specifically, this study predicted the vocabulary size test scores for English language learners using lexical lexical annotations, Doc2vec semantic representations, and BERT semantic representations of the L2 learners' essays. The developed models explained between 17% and 32% of the variance in the VST scores with the lowerest variance explained by the lexical annotations and Doc2vec models and the highest variance explained by the BERT model. While lexical annotations that explore breadth, depth, and core lexical knowledge features have become commonplace in many studies of L2 performance (Grant & Ginther, 2000; Graesser et al., 2004; Koizumi & In'nami, 2013; Sundqvist, 2019), modeling lexical knowledge based on semantic features is rare (cf. Monteiro, 2020; Sun & Lu, 2021; Lu & Hu, 2021; Zhang et al., 2021). Additionally, little research has investigated links between receptive and productive vocabulary as found in this study (Meara, 2010).

The results of the study indicate moderate links between lexical annotations and semantic models based on Doc2vec and L2 receptive vocabulary knowledge and strong links between semantic models based on BERT and L2 receptive vocabulary performance. Overall, the findings help support the notion that L2 productive language features are associated with receptive vocabulary skills (Webb, 2008). The strength of the BERT model in measuring reecptive vocabu-

lary knowledge likely relates to enhanced models of semanticity based on neural network models developed on large language corpora that include features related to context and attention.

In the developed models, the lowest performance was reported for the Doc2vec models, which performed slightly lower than the lexical annotation model. A potential reason for the lower performance is that the model was trained specifically on the ICNALE corpus. In some sense, training on the same data that comprises the test set may be an advantage. However, semantic embedding models generally perform better when they are trained on larger amounts of data. While the ICNALE corpus is large by L2 standards, it may be considered small in terms of corpora from which language models are generally trained. However, it should be noted that the ICNALE training set used in our analysis had roughly 1 million tokens, nearly twice as large as Lau and Baldwin's (2016) smallest training set. Additionally, the Doc2vec model for this study performed quite well considering it is a relatively simple and shallow network with no pretraining or finetuning. This makes the Doc2vec model more nimble and less computationally heavy, which should lead to faster run times with correspondingly smaller difference in performance. However, explaining the Doc2vec results is problematic. The vectors derived from Doc2vec used to predict VST scores relate to the semanticity of the texts, but since the vectors are just numerical representations of semanticity, they are impossible to interpret, which is a major limitation of a Doc2vec approach. In practice, Doc2vec is good at predicting VST scores, but provides the researcher and practitioner with little information about what aspects of semanticity lead to a larger receptive vocabulary size.

The lexical annotation model also explained 17% of the variance in the VST scores. The linear model indicated that the strongest predictors of receptive vocabulary were related to word meaningfulness, familiarity, and age of acquisition. This was followed by features that measured lexical decision response time, word concreteness, and USF word associations. Weaker predictors included phonological neighbors, contextual diversity, and collocation strength. In brief, writers who produced words with fewer meaningful associations (as measured by both MRC word meaningfulness scores and USF word association scores) and words that were less familiar and concrete, acquired later, and took longer to recognize as scored higher on the VST. Overall, the profile of a learner that scores higher on the VST is a writer that produces more complex lexical items while, at the same time, produces phrases that adhere to expected multi-word structures. Thus, we would expect that lexical acquisition equates to the production of more sophisticated words (i.e., words that have fewer associations and are less familiar, acquired later, less concrete, and take longer to process) while, at the same time, mastering the expectations of multi-word units.

Our BERT models, which represents semantic I in texts, performed significantly better than the Doc2vec model and the lexical annotation model explaining 32% of the variance in VST scores. The BERT model likely outperformed the Doc2vec model because of the use of a pre-trained model based on over 3 billion words and the attention mechnism contained within the model. The BERT model also likely outperform Doc2vec because it includes more advanced approaches such as finetuning and alternative pooling techniques. While our model pooled the hidden state of the first token of the final layer, different combinations of hidden states (including averaging across layers, concatenating, etc.) could improve this performance as could ensemble methods that combine different pre-trained and finetuned language models together. Additionally, unlike Doc2vec, BERT has been effectively applied to tasks such as syntactic dependency parsing that are not exclusively lexical (Goldberg, 2019; Clark et al., 2019). With this consideration in mind, it is likely that BERT is essentially capturing semantic information because it was trained to predict word distributions in a corpus in a manner similar to Word2vec. However, within that process it is also learning syntatic information bringing BERT closer to understanding the nexus between lexis and syntax. However, unlike our lexical annotation model (and similar to our Doc2vec model), it is difficult to interpret the semantic embeddings in the students' texts that predicted the VST scores because of the neural network approaches used in both Doc2vec and BERT. These neural networks, based on their complexity, make explaining model decisions extremely difficult. Even the smaller BERT model used here has 340 million parameters, which is small in size compared to more recent language models (e.g., GPT-3 has over 175 billion parameters, Brown et al., 2020). Thus, while the transformer models are more predictive, they are less interpretable. Like the Doc2vec model, this is a major limitation because researchers and practitioners can glean little from the BERT model about what it means to have more or less receptive vocabulary knowledge.



## 7. Conclusion

We find that state-of-the-art BERT models based on semantic embeddings outperform linguistic annotations and Doc2vec models in predicting L2 learners' VST scores based on features found in the students' writing. This finding helps to support the strength and accuracy of semantic embedding approaches as well as their generalizability across tasks when compared to linguistic feature models. However, we also note a major drawback of semantic embedding models: interpretability. While the linguistic features model performed statistically lower than the BERT model, its output was understandable and easy to map onto existing

theories and previous studies investigating L2 lexical knowledge. The same cannot be said for the semantic embedding models, whose opaque output helps in labeling them as *black boxes*. As such, there is a trade-off between semantic embedding models and the lexical features model in terms of model performance and transparency (Došilović et al., 2018).

There are also limitations to the current study that go beyond model interpretability. Some of these issues were discussed above (e.g., larger training sets for Doc2vec models), but some issues are specific to the conducted analyses. For instance, in this study we only focused on English and not other languages. Thus, we have no real understanding if the results are generalizable beyond English. One obstacle to generalizing findings to other languages is the massive number of resources that have traditionally been necessary to develop linguistic annotation tools for a specific language. These include part of speech taggers, dependency parsers, lexical judgement databases (like ELP) and lexical synsets (like WordNet). Semantic embedding models like Doc2vec and BERT provide a partial solution to this concern because they require no hand coding, human judgments, or rule-based systems. Given a large enough corpus, the models learn the semanticity of a language unsupervised. Thus, future studies may be able to replicate the semantic embedding findings reported here in other resource-rich languages for which large enough training corpora (and compute) are available. Additionally, if large enough L2 English corpora become available, transformer models may be developed that purposefully incorporate the English production of non-native speakers (i.e., L2 normed models). This may alleviate concerns that some researchers hold (i.e., Ortega, 2016) about depending on NLP annotations based on L1 norms.

Another limitation is that transformer models like BERT do not allow for the inclusion of co-variates that might help explain linguistic knowledge because predictions are made using a linear head that is part of the finetuning process. For example, the ICNALE corpus includes a number of demographic and individual difference variables for each learner that could be included as co-variates in models. These include age and gender (for demographic information) and individual difference features such as motivation strength (both integrative and instrumental) and learner backgrounds data such as grade level, academic background, frequency of using English, experiences being taught by native speaker of English, and country of origin. Many of these features may help to explain VST scores in addition to the semantic embeddings reported by BERT, but they are impossible to include in our modeling process. Additionally, models could have been tested on country of origin to assess potential cross-linguistic influences, but the authors of ICNALE (Ishikawa, 2013) warn against using country of origin as operationalized in ICNALE because VST scores are higher in the data for ESL outer circle

countries (Hong Kong and Singapore) than for EFL expanding circle countries (Japan and Thailand). Thus, proficiency level and L1 are strongly correlated and any comparison based solely on L1 would be unprincipled.

Even considering their limitations, transformer models are state-of-the-art and commonly used in fields as diverse as health, finance, military, transportation, and security (Arrieta et al., 2020) because of their performance strengths. While uncommon in L2 studies, this study shows their strength in prediction tasks of interest to the L2 community. The likelihood of transformer models like BERT becoming more mainstream in L2 studies is strong and will become stronger considering the push towards more interpretable AI (including transformer models) in government, industry, and academia. Interpretable AI is necessary to ensure that decisions made by models are justifiable (Gunning & Aha, 2019) and the models allow for detailed explanations to increase human trust and understanding (Zhu et al., 2018). As noted by Arrieta et al. (2020), developing interpretable machine learning models can help detect and correct potential bias in training sets, highlight small changes (i.e., perturbations) that might change predictions, and help ensure a causality in model reasoning.

As transformer models become more interpretable and begin to tell us more about the underlying cognitive processes of L2 acquisition, their uptake will likely increase. Modified deep learning techniques like training neural networks to associate labelled nodes with known semantic ontologies, generate examples and/or clusters from unlabeled and/or prominent nodes to help with semantic interpretation, and identify which architectures, parameters, and training lead to the most interpretable models (Gunning et al., 2019) should make the semantic output of transformer models more actionable for L2 researchers and practitioners. Once available, interpretable semantic representations for L2 learners will help L2 researchers develop models of L2 knowledge and development related to meaning, intention, inference, and pragmatics, all areas that are difficult, if not impossible, to model computationally. These models should have immediate impacts in the language learning classroom or learning system.

From a theory-driven perspective, we do not recommend that researchers use transformer language models in place of existing NLP annotation to analyze learner language, especially if interpretation of output is critical. However, in cases where sufficient data is available, computational resources exist, and interpretability is not a concern, embedding-based approaches to NLP offer appealing utility for a wide variety of language analysis tasks, so long as researchers acknowledge and manage the potential for bias in these models. Doc2vec and other static embedding models may also prove useful to some research projects while requiring many fewer computational resources. We are optimistic that future research will develop creative methodologies that leverage embedding

models while minimizing their limitations. Meanwhile, NLP annotations of lexical features continue to provide a useful and interpretable means of studying learner language.

## References


Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Balota, D.A., Cortese, M.J., Sergent-Marshall, S.D., Spieler, D.H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology. General*, *133*(2), 283–316. https://doi.org/10.1037/0096-3445.133.2.283

Balota, D.A., Yap, M.J., Cortese, M.J., Hutchison, K.A., Kessler, B., Loftis, B., Neely, J.H., Nelson, D.L., Simpson, G.B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459. https://doi.org/10.3758/BF03193014

Berger, C., Crossley, S., & Kyle, K. (2019). Using native-speaker psycholinguistic norms to predict lexical proficiency and development in second-language production. *Applied Linguistics*, *40* (1), 22–42. https://doi.org/10.1093/applin/amx005

Berger, C., Crossley, S., & Skalicky, S. (2019). Using lexical features to investigate second language lexical decision performance. *Studies in Second Language Acquisition*, *41*(5), 911–935. https://doi.org/10.1017/S0272263119000019

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D., Gray, B., & Staples, S. (2016). Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

BNC Consortium, The British National Corpus, XML Edition, (2007), *Oxford Text Archive*, http://hdl.handle.net/20.500.12024/2554

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606* [Cs]. http://arxiv.org/abs/1607.04606. https://doi.org/10.1162/tacl_a_00051

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv:2005.14165 [Cs]*. http://arxiv.org/abs/2005.14165

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Clark, K., Khandelwal, U., Levy, O., & Manning, C.D. (2019). What Does BERT Look At? An Analysis of BERT's Attention (arXiv:1906.04341). *arXiv*. http://arxiv.org/abs/1906.04341

Cobb, T. (n.d.). *Web Vocabprofile*. https://www.lextutor.ca/

Conrad, S. (2005). Corpus Linguistics and L2 Teaching. In *Handbook of Research in Second Language Teaching and Learning*. Routledge.

Crossley, S.A., & Kyle, K. (2022). Managing Second Language Acquisition Data with Natural Language Processing Tools. In *The Open Handbook of Linguistic Data Management* (pp. 411–421). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0039

Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011a). What Is Lexical Proficiency? Some Answers from Computational Models of Speech Data. *TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect*, *45*(1), 182–193. https://doi.org/10.5054/tq.2010.244019

Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011b). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Crossley, S.A., & Skalicky, S. (2019). Examining Lexical Development in Second Language Learners: An Approximate Replication of Salsbury, Crossley & McNamara (2011). *Language Teaching*, *52*(3), 385–405. https://doi.org/10.1017/S0261444817000362

Crossley, S.A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*(4), 721–744. https://doi.org/10.1017/S0272263118000268

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334. https://doi.org/10.1111/j.1467-9922.2009.00508.x

Crossley, S., Salsbury, T., & McNamara, D. (2010). The Development of Polysemy and Frequency Use in English Second Language Speakers: Polysemy and Frequency Use in English L2 Speakers. *Language Learning*, *60*(3), 573–605. https://doi.org/10.1111/j.1467-9922.2010.00568.x

David, A. (2008). Vocabulary breadth in French L2 learners. *The Language Learning Journal*, *36*(2), 167–180. https://doi.org/10.1080/09571730802389991

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Došilović, F.K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. https://doi.org/10.23919/MIPRO.2018.8400040

Ellis, N.C. (2002). Frequency effects in language processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/7287.001.0001

Garner, J., & Crossley, S. (2018). A Latent Curve Model Approach to Studying L2 N-Gram Development. *The Modern Language Journal*, *102*(3), 494–511. https://doi.org/10.1111/modl.12494

Garner, J., Crossley, S., & Kyle, K. (2018). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, *58*(1), 51–74. https://doi.org/10.1515/iral-2017-0089

Goldberg, Y. (2019). Assessing BERT's Syntactic Abilities (arXiv:1901.05287). *arXiv*. https://doi.org/10.48550/arXiv.1901.05287

Graesser, A.C., McNamara, D.S., Louwerse, M.M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Grant, L., & Ginther, A. (2000). Using Computer-Tagged Linguistic Features to Describe L2 Writing Differences. *Journal of Second Language Writing*, *9*(2), 123–145. https://doi.org/10.1016/S1060-3743(00)00019-9

Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI-Explainable artificial intelligence. *Science Robotics*, *4*(37), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

Hashimoto, B.J., & Egbert, J. (2019). More Than Frequency? Exploring Predictors of Word Difficulty for Second Language Learners. *Language Learning*, *69*(4), 839–872. https://doi.org/10.1111/lang.12353

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, *23*(1), 28–54. https://doi.org/10.1075/ijcl.16080.hua

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118.

Ke, Z., & Ng, V. (2019). *Automated Essay Scoring: A Survey of the State of the Art*. 6300–6308.

Kerz, E., Wiechmann, D., Qiao, Y., Tseng, E., & Ströbel, M. (2021). Automated Classification of Written Proficiency Levels on the CEFR-Scale through Complexity Contours and RNNs. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, 199–209. https://aclanthology.org/2021.bea-1.21

Kohavi, R., & John, G.H. (1995). Automatic Parameter Selection by Minimizing Estimated Error. In A. Prieditis & S. Russell. (Eds.), *Machine Learning Proceedings 1995* (pp. 304–312). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-377-6.50045-1

Koizumi, R., & In'nami, Y. (2013). Vocabulary Knowledge and Speaking Proficiency among Second Language Learners from Novice to Intermediate Levels. *Journal of Language Teaching and Research*, *4*(5), 900–913. https://doi.org/10.4304/jltr.4.5.900-913

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S.A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S., & Berger, C. (2018). The Tool for the Automatic Analysis of Lexical Sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of Latent Semantic Analysis*. Psychology Press. https://doi.org/10.4324/9780203936399

Lau, J. H., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86. https://doi.org/10.18653/v1/W16-1609

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Le, Q. V., & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *ArXiv:1405.4053 [Cs]*. http://arxiv.org/abs/1405.4053

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R. H., Grainger, J., & Zwitserlood, P. (2008). Native language influences on word recognition in a second language: A megastudy. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *34*(1), 12–31. https://doi.org/10.1037/0278-7393.34.1.12

Lu, Xiaofei. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X., & Hu, R. (2021). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01675-6

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the Word Frequency Effect: The Neglected Role of Distributional Information in Lexical Processing. *Language and Speech*, *44*(3), 295–322. https://doi.org/10.1177/00238309010440030101

Meara, P. (1996). The dimensions of lexical competence. *Performance and Competence in Second Language Acquisition*, *35*, 33–55.

Meara, P. (2005a). Designing vocabulary tests for English. *The Dynamics of Language Use: Functional and Contrastive Perspectives*, *140*, 271. https://doi.org/10.1075/pbns.140.19mea

Meara, P. (2005b). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, *26*(1), 32–47. https://doi.org/10.1093/applin/amh037

Meara, P. (2010). The relationship between L2 vocabulary knowledge and L2 vocabulary use. *The Continuum Companion to Second Language Acquisition*, 179–193.

Meurers, D. (2012). Natural Language Processing and Language Learning. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405198431.wbeal0858

Meurers, D. (2021). Natural Language Processing and Language Learning. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781405198431.wbeal0858.pub2

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. https://doi.org/CitetononCRdoi:10.48550/ARXIV.1301.3781

Milton, J. (2009). Measuring Second Language Vocabulary Acquisition. In *Measuring Second Language Vocabulary Acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

Moghadam, S. H., Zainal, Z., & Ghaderpour, M. (2012). A review on the important role of vocabulary knowledge in reading comprehension performance. *Procedia-Social and Behavioral Sciences*, *66*, 555–563. https://doi.org/10.1016/j.sbspro.2012.11.300

Monteiro, K. R., Crossley, S. A., & Kyle, K. (2020). In Search of New Benchmarks: Using L2 Lexical Frequency and Contextual Diversity Indices to Assess Second Language Writing. *Applied Linguistics*, *41*(2), 280–300. https://doi.org/10.1093/applin/amy056

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System*, *32*(1), 75–87. https://doi.org/10.1016/j.system.2003.05.001

Mostafa, T., Crossley, S., & Kim, Y. (2021). Predictors of English as second language learners' oral proficiency development in a classroom context. *International Journal of Applied Linguistics*, *31* (3), 526–548. https://doi.org/10.1111/ijal.12358

Nagy, W. E., & Scott, J. A. (2000). Vocabulary processes. In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr. (Eds.), *Handbook of reading research* (Vol. *3*, pp. 269–284). Mahwah, NJ: Earlbaum.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Ortega, L. (2016). Multi-competence in second language acquisition: inroads into the mainstream? In V. Cook & L. Wei. (Eds) *The Cambridge Handbook of Linguistic Multi-competence*. Cambridge University Press. https://doi.org/10.1017/CBO9781107425965.003

Paetzold, G., & Specia, L. (2016). Collecting and Exploring Everyday Language for Predicting Psycholinguistic Properties of Words. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1669–1679. https://aclanthology.org/C16-1157

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Read, J. (1998). Validating a Test to Measure Depth of Vocabulary Knowledge. In *Validation in Language Assessment*. Routledge.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Saito, K. (2020). Multi- or Single-Word Units? The Role of Collocation Use in Comprehensible and Contextually Appropriate Second Language Speech. *Language Learning*, *70*(2), 548–588. https://doi.org/10.1111/lang.12387

Sun, K., & Lu, X. (2021). Assessing Lexical Psychological Properties in Second Language Production: A Dynamic Semantic Similarity Approach. *Frontiers in Psychology*, *12*, 672243. https://doi.org/10.3389/fpsyg.2021.672243

Sundqvist, P. (2019). Commercial-off-the-shelf games in the digital wild and L2 learner vocabulary. *Language Learning*, *23*(1), 27.

Vanderbilt, Katia, "Developing and Testing Alternative Benchmarks of Lexical Sophistication: L2 Lexical Frequency, Semantic Context, and Word Recognition Indices." Dissertation, Georgia State University, 2020. https://doi.org/CitetononCRdoi:10.57709/18616934

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

Webb, S. (2009). The Effects of Receptive and Productive Learning of Word Pairs on Vocabulary Knowledge. *RELC Journal*, *40*(3), 360–376. https://doi.org/10.1177/0033688209343854

Wilson, M. (1988). MRC psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6–10. https://doi.org/10.3758/BF03202594

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zaytseva, V., Miralpeix, I., & Pérez-Vidal, C. (2019). Because words matter: Investigating vocabulary development across contexts and modalities. *Language Teaching Research*, 136216881985297.

Zhang, H., Chen, M., & Li, X. (2021). Developmental Features of Lexical Richness in English Writings by Chinese Beginner Learners. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.665988

Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G.M. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8. https://doi.org/10.1109/CIG.2018.8490433



## Address for correspondence

Scott Crossley
Vanderbilt University
United States
sacrossley@gmail.com

## Co-author information

Langdon Holmes
Vanderbilt University
lholmes15@gsu.edu

## Publication history

Date received: 13 April 2022
Date accepted: 5 August 2022
Published online: 20 September 2022