Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words?

Kevin P. Yancey, Yves Lepage
Waseda University
2-7 Hibikino, Wakamatu-ku, Kitakyushu city, Fukuoka, Japan
kpyancey@fuji.waseda.jp, yves.lepage@waseda.jp

Abstract

Vocabulary knowledge prediction is an important task in lexical text simplification for foreign language learners (L2 learners). However, previously studied methods that use hand-crafted rules based on one or two word features have had limited success. A recent study hypothesized that a supervised learning classifier trained on a large annotated corpus of words unknown by L2 learners may yield better results. Our study crowdsourced the production of such a corpus for Korean, now consisting of 2,385 annotated passages contributed by 357 distinct L2 learners. Our preliminary evaluation of models trained on this corpus show favorable results, thus confirming the hypothesis. In this paper, we describe our methodology for building this resource in detail and analyze its results so that it can be duplicated for other languages. We also present our preliminary evaluation of models trained on this annotated corpus, the best of which recalls 80 % of unknown words with 71 % precision. We make our annotation data available.

Keywords: lexical simplification, vocabulary knowledge prediction, crowdsourcing

1. Introduction

Our goal is to build a text simplification system for L2 learners of Korean to aid reading comprehension and expedite language acquisition through reading. Reading, and extensive reading in particular (i.e., the practice of reading large amounts of easy, entertaining text), has been shown to have many benefits to language acquisition, including improving grammar (Krashen, 2003) , writing (Mason and Krashen, 1997), and listening skills (Elley, 1991), as well as reading proficiency and motivation (Crawford Camiciottoli, 2001).

A task critical to simplifying texts for L2 learners is estimating the proficiency level required to understand a word (i.e., the word's "complexity"), so that unknown words can be predicted and simpler replacements can be selected. Many past proposals focused on rules that used word frequency, length, or some combination thereof as a proxy for word complexity (Devlin and Tait, 1998; Bott et al., 2012; Shardlow, 2013), but these features do not correlate perfectly with word complexity for L2 learners. More recently, Tack et al. (2016) proposed a model that measured a word's complexity as its first level of occurrence within a corpus of graded textbooks for L2 learners, evaluating it against a corpus annotated by four L2 learners of French. However, their model was only able to correctly classify 40 % of the unknown words.

Tack et al.'s paper hypothesized that better results might be obtained by compiling a much larger annotated corpus and casting the problem as a supervised learning problem. In this paper, we will explore this alternative. We will describe our methodology for constructing a large annotated corpus by crowdsourcing via the Internet, and will evaluate the predictive capability of several models built from this corpus via supervised learning, comparing our results to previous approaches. Finally, we will describe the contents of this annotated corpus and make it available to other researchers to build their own models and compare with our results.

2. Building an L2 Annotated Corpus of Unknown Korean Words

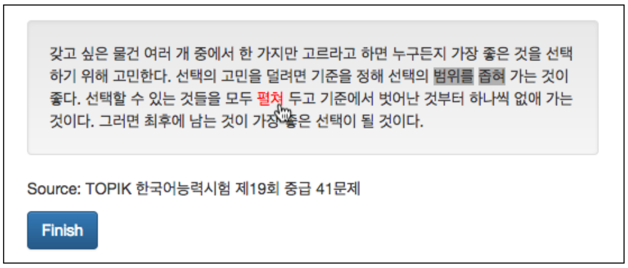
To collect a training corpus, we crowdsourced the annotation of Korean text via an Internet survey, where L2 learners annotated the words they did not understand.

2.1. Corpus Selection

We used graded synthetic texts extracted from the reading sections of past Test of Proficiency in Korean (TOPIK) exams (National Institute for International Education Development in Korea, 2017), which is the standard Korean language proficiency test administered by the South Korean government. This test rates L2 learners' Korean proficiency on a scale of 6 levels, which correspond roughly to the 6 levels in the Common European Framework of Reference (CEFR) (Won, 2016). Prior to July 2014, TOPIK exams came in three difficulty levels, or "grades": beginner (for levels 1-2), intermediate (for levels 3-4), and advanced (for levels 5-6). We automatically downloaded, parsed, and extracted reading passages from 25 different TOPIK exams for each of these three grades. Among these extracted texts, we excluded passages containing blanks or special characters that are used as markers for the exam questions. We created the corpus using the remaining 263 short passages, consisting of 31 beginner passages, 143 intermediate passages, and 89 advanced passages, most having 3–6 sentences each.

2.2. Designing a Survey to Annotate Unknown Words

The survey was conducted via a website custom-built for the purpose. When annotators first started the survey, they were asked to fill out a basic questionnaire about their native language, Korean proficiency level, and reasons for studying Korean. They were then presented with a series



갖고 싶은 물건 여러 개 중에서 한 가지만 고르라고 하면 누구든지 가장 좋은 것을 선택하기 위해 고민한다. 선택의 고민을 덜려면 기준을 정해 선택의 범위를 좁혀 가는 것이 좋다. 선택할 수 있는 것들을 모두 펼쳐 두고 기준에서 벗어난 것부터 하나씩 없애 가는 것이다. 그러면 최후에 남는 것이 가장 좋은 선택이 될 것이다.

Source: TOPIK 한국어능력시험 제19회 중급 41문제

Words turn red as the annotator hovers the cursor over them. Unknown words that have been clicked by the annotator are highlighted with a dark gray background.

Figure 1: Survey passage annotation

0 - Did not understand the passage at all.

1 - Understood the general topic only.

2 - Mostly understood the passage.

3 - Understood all or almost all the passage.

Figure 2: Annotator self-assessed passage comprehension scale

of passages selected semi-randomly from the corpus described in the previous section. So that passages would not be too difficult for annotators, the selection algorithm estimated the annotator's level and provided passages that were roughly one or two TOPIK levels higher so that the annotator would be expected to know roughly 80–90 % of the words in each passage. Annotators were asked to read each passage, without a dictionary, and highlight the words they did not understand by clicking on them (see Figure 1).[1] Once the annotator finished the passage, the selected words were annotated as unknown, and all the remaining words were annotated as known. After annotating each passage, annotators were asked to rate their comprehension of the passage according to the scale defined in Figure 2.

2.3. Maximizing Annotation Submissions

Each annotator was asked to finish at least 2 passages so that we would have enough data to estimate each annotator's level in addition to learning which words they did or did not know. However, to maximize the amount of data we could gather from each participant, we added the following gamification features to encourage users to annotate additional passages:

Points Annotators were awarded points for each passage completed.

Titles Annotators earned titles for earning certain numbers of points. They were regularly presented with a status screen showing them their points earned and their progress towards earning the next title.

Language[2]	# of Ann. Passages	% of Ann. Passages	# of Annotators	% of Annotators
English (en)	1,450	60.8 %	279	78.2 %
Korean (ko)[3]	262	11.0 %	1	0.3 %
Danish (da)	258	10.8 %	1	0.3 %
Chinese (zh)	87	3.6 %	16	4.5 %
Portuguese (pt)	79	3.3 %	8	2.2 %
Spanish (es)	49	2.1 %	6	1.7 %
German (de)	39	1.6 %	11	3.1 %
Vietnamese (vi)	17	0.7 %	3	0.8 %
Other (but known)	56	2.3 %	18	5.0 %
Unknown	88	3.7 %	14	3.9 %

Table 1: Annotated passages by native language

Rankings Annotators were ranked by points earned and could view their current ranking (and that of other contributors) on a rankings page of the website.

To make it easy for annotators of various language backgrounds to contribute, the website's UI was available in 5 languages: English, Chinese, Japanese, French, and (for those whose native language wasn't supported) Korean.
We promoted the survey on forums and social media where L2 learners of Korean were likely to find it: mainly the websites waygook.org (an online forum for foreigners living in Korea), reddit.com/r/Korean (an online forum for L2 learners of Korean), iTalki (an online language learning website), and Facebook. Additionally, having seen the survey's announcement, a person with a very large number of followers helped promote the survey via Twitter.

3. Analysis of Annotation Results

In 3 months time, we collected a total of 2,385 annotated passages from 357 distinct annotators of varying levels, countries, and language backgrounds. Extracting the annotations and removing duplicates (see Section 4.1.) resulted in 48,622 distinct annotator/baseword pairs, each annotated as either known or unknown.

3.1. Demographics of Annotators

The annotated corpus includes native speakers of 17 languages (see Table 1) living in a variety of countries (see Table 2).

3.2. Factors Impacting the Volume of Annotated Submissions

In this section, we discuss some of the factors that affected the quantity of annotations received. Here we refer to those who visited the survey website as "visitors", a superset of those who contributed annotations (i.e., the "annotators").

[1] Written Korean text is spaced with 어절s (pronounced "eo-jeol"), which are comprised of an inflected word form followed by some number of particles. To keep the annotation process simple, we used 어절s as the unit of annotation.

[2] When not supplied by the annotator in the initial questionnaire, native language was inferred from the preferred language indicated in the annotator's HTTP headers.

[3] A native Korean speaker also submitted annotations while proofreading the essays for us, as she also encountered some words she had not known in some of the advanced passages, so we included her annotations in the dataset to help identify very complex words.

Country	# of Ann. Passages	% of Ann. Passages	# of Annotators	% of Annotators
Japan (JP)	399	16.7 %	6	1.7 %
United States (US)	383	16.1 %	88	24.6 %
Denmark (DK)	258	10.8 %	1	0.3 %
Philippines (PH)	239	10.0 %	50	14.0 %
Korea, Republic of (KR)	186	7.8 %	32	9.0 %
Australia (AU)	102	4.3 %	14	3.9 %
Indonesia (ID)	95	4.0 %	18	5.0 %
Other/Unknown	723	30.3 %	148	41.5 %

Table 2: Annotated passages by country

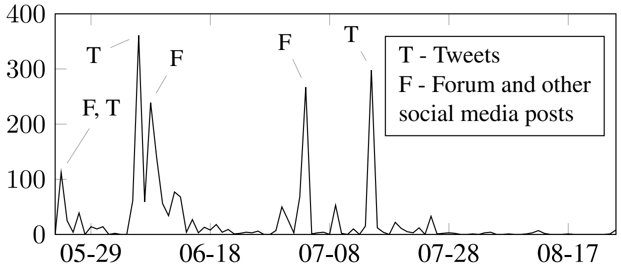


Figure 3: # of annotated passages by date

3.2.1. Promotion through Forums and Social Media

We tended to receive a large volume of annotated passages immediately following announcements on online forums and social media, followed by a much smaller volume of annotated passages in the following days or weeks. This is seen in Figure 3, where the letter T marks the dates promotional tweets were sent out and the letter F marks the dates posts were made to forums and social media websites.

We use the timing of annotated passages received and the sources of web traffic (see Table 3) to discern which websites worked best for recruiting annotators. In our case, promotion by the aforementioned Twitter user, who has over 3,500 followers with an interest in Korean popular culture, resulted in the most annotated passages. This user independently sent tweets about our survey on 3 separate occasions, resulting roughly 300 new annotated passages each time. The second largest group of annotators came from Reddit, whose contributors also showed a great diversity of language levels.

3.2.2. Gamification

Annotators gave positive feedback about the gamification features. Figure 4 also shows its positive effect on the number of annotated passages received, where we find that 29 % of annotators completed enough passages to earn the first “title”, but that the submission rate of annotated passages drops off sharply after that point.

Source	# of Visitors	% of Visitors
Twitter	968	41.96 %
Facebook[4]	222	9.62 %
Reddit	155	6.72 %
Waygook.org	99	4.29 %
Other/Unknown	863	37.40 %

Table 3: Sources of annotation survey website visitors

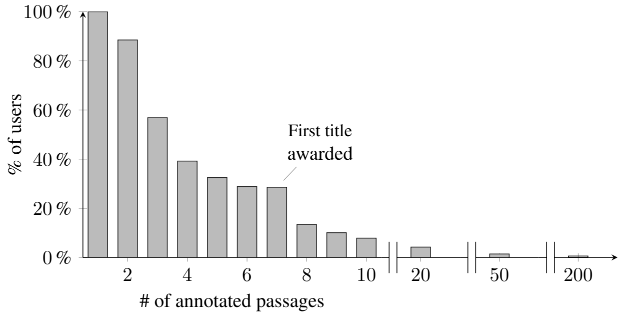


Figure 4: # of annotated passages per annotator

3.2.3. Login Accounts and Privacy Concerns

Feedback from visitors and website usage statistics show that requiring annotators to register accounts negatively impacted the number of visitors who chose to annotate passages. Initially, a number of would-be annotators raised privacy concerns over using social logins, fearing they might be used to collect personally identifying information about them. After adding an option to register “anonymously” using only an arbitrary username and password that is not linked to an email or social login, website usage statistics still showed that as many as 50 % of visitors left the website once they reached the registration page.

3.3. Quality & Inter-annotator Agreement

The literature indicates that annotation tasks of this kind tend to have low inter-annotator agreement by conventional measures. Paetzold and Specia (2016) reported a Krippendorff’s Alpha agreement coefficient (Hayes and Krippendorff, 2007) of 0.244, which they hypothesized was due to differences in language backgrounds and proficiency levels of the annotators. Tack et al. (2016) also noted that there can be high variation in the annotation of content words, even among annotators of the same level and language background.

The In-Corpus Model described in Section 4.1. predicts unknown words for each annotator based on how other annotators annotated those same words. So, we use this model as a tool to measure agreement between each annotator in a way that accounts for level differences and identify annotators whose annotations may be unreliable. We do this by computing the cross entropy between each annotator’s annotations and the model’s predictions. We then compare this to the distribution of cross entropies that result if the same words are labeled as known/unknown at random in the same proportion. This analysis showed that 74 % of annotators (which corresponds to 90 % of all annotations) submitted data that agrees with the model significantly better than randomly annotated data (p=0.01). [5]

[4] While a sizable number of visitors came from Facebook, for the most part the timings of these visits did not correlate with survey receipt of annotated passages.

[5] Another 8 % of annotators either did not select any words as unknown, or selected all words as unknown. A manual review of the remaining annotator’s annotations suggests that some performed the annotation task backwards: selecting the words they did know instead of the words they did not know. Many others did

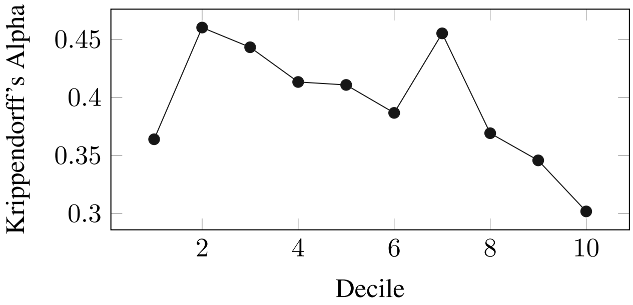


Figure 5: Inter-nnotator Agreement for Annotators Grouped into Deciles by Estimated Proficiency Level

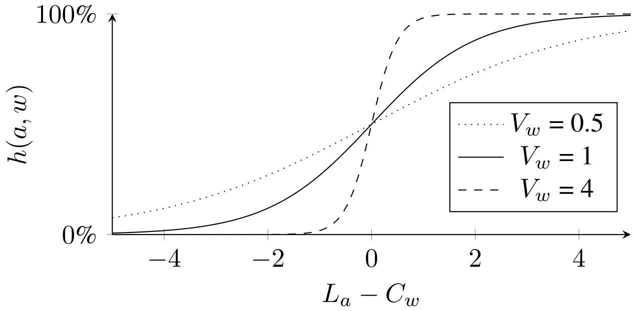


Figure 6: Probability of a word being known as predicted by h with different word complexity variances V_w

Within this 74 % of annotators, comparing annotators of similar estimated proficiency level by grouping them into deciles, we get an average Krippendorff's Alpha among the deciles of 0.394 (see Figure 5). By computing the alpha pairwise, we see the effect that the difference in proficiency level has on inter-annotator agreement: pairs of annotators with less than one percentile difference have an alpha of 0.312 on average, but those with at least a 40 percentile difference in proficiency level have negative alphas (indicating systemic disagreement) on average.

4. Unknown Word Prediction Models

In this section we demonstrate that by using a large annotated corpus as a labeled training dataset, better unknown word prediction models can be built. We evaluate three prediction models built from this annotated corpus: an In-Corpus Model that is limited to predicting words found within the training corpus, and two general models based on Support Vector Machines (SVMs). We compare the performance of these models to three baseline models. Finally, we investigate how large of an annotated corpus is needed to maximize the performance of these models.

4.1. In-Corpus Model

We start by preprocessing the annotated corpus, turning it into a labeled training dataset: each annotation's word is normalized to its baseword by removing particles, inflections, and morphological suffixes,[6] and duplicate annotations are removed so that there is at most one annotation per annotator/baseword pair (see "Labeled Dataset" in Appendix for details).

A manual review of this data, however, reveals that many of the self-reported Korean proficiency levels of the annotators are unreliable: frequently being 1–3 levels higher than what the annotator's actual level appears to be. This leaves us with no option but to estimate annotators' proficiency levels based on their annotations. But, we cannot estimate the proficiency levels of each annotator without considering the complexity of words they have annotated, which is what we're trying to learn, so this becomes a chicken-and-egg problem.

We solve this problem by learning the annotator proficiency levels and word complexities simultaneously using a method that is similar to logistic regression. We do this by first defining an equation, h, that models how the probability of an annotator knowing a word is related to the annotator's level and to the word's complexity, and then applying gradient descent to estimate the annotator proficiency levels and word complexities simultaneously.

For this model to make sense, h needs to be an S-shaped curve so that the estimated probability of a reader knowing a word increases monotonically as the difference between the annotator's level (denoted L_a) and the word's complexity (denoted C_w) increases (see Figure 6). Thus, we choose to define h as follows:

$$h(a, w) = \phi\left(V_w^{-1}(L_a - C_w)\right) \tag{1}$$

where a is the annotator, w is the word, and ϕ is the sigmoid function $\phi(x) = 1/(1 + e^{-x})$. The word complexity variance, V_w, is added to control the steepness of the curve for each word.[7] With equation h defined, we define a loss function over our training dataset using the cross entropy formula, and apply gradient descent to find the annotator levels, L, and word complexities, C, that maximize the likelihood of the observed annotations.

A limitation of the In-Corpus Model is that it can only estimate the complexity of words whose baseword occurs in the training set (i.e., "seen" words). Because our annotated corpus consists of only 263 short passages, this limits our In-Corpus Model to making predictions for the words included in one of the 3,612 word families that correspond to the basewords found in our corpus. Authentic texts usually contain many words outside these word families. So, in the next section, we introduce a general model based on Support Vector Machines (SVMs) that works for both seen and unseen words.

4.2. General Models that Handle Unseen Words

We build two general models using SVMs. For input features, we use the annotator level learned by the In-Corpus Model and a variety of word features. Platt scaling (Platt and others, 1999) is used to convert the SVM's output classification scores into probabilities. We evaluate two SVM-

not annotate enough examples of unknown words for this analysis to conclusively show if their annotations were genuine.

[6] So that, for example, 경제 "economy" and경제적으로 "economically" are treated as the same word

[7] So that $h(r, w)$ is continuous for all learned variables, we replaced V with $e^{V'_w}$ where $V'_w = \ln V_w$, and learn V' instead of V directly. Thus, the final equation becomes $h(r, w) = \phi\left(e^{-V'_w}(C_w - L_a)\right)$.

Model	Cross Entropy	Accuracy at $t = 0.5$	Precision at 80 % Recall
Seen Words Only			
In-Corpus Model	0.328	85.5 %	73.1 %
General Models			
SVM (RBF)	0.341	84.3 %	70.8 %
SVM (Linear)	0.361	83.0 %	67.4 %
Baseline Models			
TOPIK Word Level	0.393	81.0 %	62.1 %
Word Frequency Only	0.399	80.9 %	61.0 %
Word Frequency & Length	0.398	81.0 %	61.2 %

Table 4: Evaluation metrics of prediction models

based models: one using a linear kernel, and the other using the RBF (Radial Basis Function) kernel.

To select word features useful to the classification problem, we use Pearson's correlation to compare the word complexities estimated by the In-Corpus Model to various word features. This yields 23 candidate word features, which are listed in Table 5. We use Recursive Feature Elimination with Cross Validation (RFECV) (Guyon et al., 2002) to further reduce the list of features and avoid over-fitting, but find that only the Noun feature harms the models.

4.3. Experiments

We use 10-fold cross validation repeated 10 times to evaluate these three models and compare their performance to three baseline models. We also run experiments training the best model on subsets of the training data to investigate how much annotation data is needed to achieve good results.

4.3.1. Baseline Models

The General SVM Models described in the previous section incorporate a comprehensive set of word features, which is practical only because we have a sufficiently large labeled dataset for training. Since such a resource is often lacking, many previous approaches have often used only one or two features as proxies for word complexity. Most commonly, word frequency, word length, or some combination thereof has been used, such as was done in Bott et al. (2012). For L2 learners, however, defining word complexity by the word's level of first occurance within a graded corpus may produce more accurate results, as was investigated by Tack et al. (2016).

To determine if similar results could be achieved with these approaches, we build three baseline models in the same manner described in the previous section, but using only the word features proposed by these approaches: one using only log inverse word frequency, one using both log inverse word frequency and word length, and one using word level within a graded corpus. For the last of these, we use the TOPIK exams as our graded corpus.

4.3.2. Model Evaluation Metrics

Each of these models predict the probability that a given word is unknown by a given reader. Since the cost of misclassifying an unknown word will, in practice, often be different than that of misclassifying a known word, it is useful to consider discrimination thresholds other than 50 %. For brevity, we will call this discrimination threshold t.

Using the collected annotations as the gold standard, we evaluate these models using 10-fold cross validation repeated 10 times, comparing the predictions to the actual classes given by the annotations. We measure three metrics:

Cross Entropy - The cross entropy between the predicted probabilities and the actual classes. This metric measures the prediction accuracy of the model across all possible values of t, but can be difficult to interpret.

Accuracy at $t = 0.5$ - The percentage of annotations where the predicted class matched the actual class when the classification threshold, t, is 50 %.

Precision at 80 % Recall - The precision[8] of the model if t is set such that recall[9] is 80 %.

The last of these is the most relevant metric for our purposes because our analysis[10] indicates that identifying 80 % of unknown words should be sufficient for intermediate readers to understand 95 % of the words when reading advanced texts, which is normally sufficient for comprehension (Laufer and Ravenhorst-Kalovski, 2010).

4.3.3. Experimental Results

The results of our experiments are shown in Table 4. The In-Corpus Model achieves 85 % accuracy (see Table 4). By adjusting the discrimination threshold, it is able to identify 80 % of the annotators' unknown words with a precision of 73 %.

Unlike the In-Corpus Model, the General SVM Model using the RBF kernel can predict words not found int he training corpus, and achieves similar performance, having 71 % precision at 80 % recall. The General SVM Model using the linear kernel performs a little worse, getting only 67 % precision.

The best of the baseline models is the one that uses the TOPIK Word Level feature, but was substantially outperformed by our proposed general models, achieving only 62 % precision at 80 % recall. From this we conclude that this feature alone does not provide enough information to make accurate predictions, nor does the combination word frequency and word length.

To investigate how the size of the annotated corpus affects performance, we use the same cross validation procedure to evaluate the General SVM (RBF) Model while training on only a percentage of the training data available during each cross-validation fold, repeating the experiment with progressively larger percentages (see Figure 7). While the model performs better when a larger percentage of the training data is used, the gains are almost negligible after the percentage exceeds 60 % (or about 30,000 distinct annotator/baseword pairs). Measuring precision at 80% recall, there is less than 1 percentage point difference between the model trained on 60 % of the available data and the one trained on 100 % of the available data.

[8] Precision is the percentage of words predicted to be unknown that were actually unknown.

[9] Recall is the percentage of unknown words predicted as unknown.

[10] We computed this by calculating the density of advanced words (i.e., words that do not appear beginner or intermediate TOPIK exams) in advanced TOPIK tests.

Feature Name	Corpus	Pearson's $r^{\dagger}$	Definition
TOPIK Beg. Inverse Log WF	TOPIK Beginner	0.50	The log inverse word frequency (i.e., $-\ln f(w)$), of the word sampled from the corpus.
TOPIK Int. Inverse Log WF	TOPIK Intermediate	0.51	
TOPIK Adv. Inverse Log WF	TOPIK Advanced	0.38	
Inverse Log WF	Exquisite Corpus[‡]	0.50	
TOPIK Beg. DF	TOPIK Beginner	-0.53	The document frequency (DF) of the word in the corpus.
TOPIK Int. DF	TOPIK Intermediate	-0.49	
TOPIK Adv. DF	TOPIK Advanced	-0.42	
TOPIK Beg. Informativeness	TOPIK Beginner	0.02[*]	The informativeness of the word in the corpus, calculated as the Log Inverse WF divided by the phonemic length.
TOPIK Int. Informativeness	TOPIK Intermediate	0.05[*]	
TOPIK Adv. Informativeness	TOPIK Advanced	-0.01[*]	
Informativeness	Exquisite Corpus[‡]	0.08	
TOPIK DF $>$ 0 Level	TOPIK (all levels)	0.49	The lowest TOPIK level at which the word's document frequency (DF) is greater than the threshold.[b]
TOPIK DF $>$ 10 Level	TOPIK (all levels)	0.48	
TOPIK DF $>$ 20 Level	TOPIK (all levels)	0.42	
Phonemic Length	n/a	0.14	The number of 글자 (i.e., Korean vowels and consonants) in the spelling of the word.[c]
Syllabic Length	n/a	0.08	The number of syllable blocks in the spelling of the word.[d]
English Cognate	n/a	-0.12	A flag, 0 or 1, that indicates whether the word is a cognate of an English word.
Noun	n/a	0.05	The word's part of speech as determined by the `twitter-korean-text` API (Ryu, 2017) encoded using one-of-n.
Proper Noun	n/a	0.08	
Adjective	n/a	-0.01[*]	
Verb	n/a	-0.05	
Adverb	n/a	-0.04[*]	
Other POS	n/a	-0.07	

[†] The Pearson correlation between word complexity (as estimated by the In-Corpus Model) and the word feature.
[‡] Word frequencies for this general corpus are provided by the `wordfreq` Python package (Speer et al., 2016).
[*] Correlation not statistically significant ($p < 0.01$).
[a] Exquisite Corpus compiles texts from a variety of sources, including Wikipedia, Reddit, Twitter, and movie subtitles. Word frequencies for this corpus are provided by the `wordfreq` Python package (Speer et al., 2016).
[b] TOPIK Level is encoded as 0 for beginner, 1 for intermediate, 2 for advanced, or 3 for words whose document frequency (DF) is less than the threshold for all levels.
[c] For example, 앉다 (to sit) would have phonemic length of 5 because it consists of the letters (ㅏ, ㄴ, ㅈ, ㄷ, ㅏ).
[d] For example, 앉다 (to sit) has a syllabic length of 2 because it consists the blocks of 앉 and 다.

Table 5: Candidate word features for SVM models

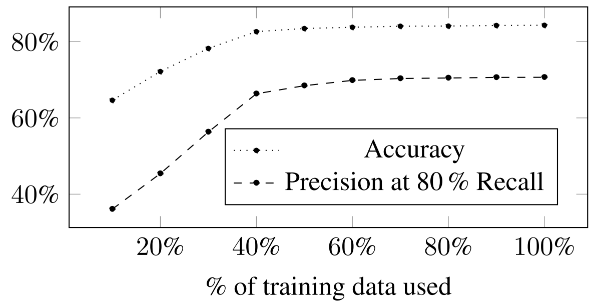


Figure 7: Evaluation metrics of SVM (RBF) Model by % of training data used

5. Conclusion and Future Work

We have demonstrated that crowdsourcing and gamification can be used to gather a large annotated corpus of words unknown by L2 learners. Furthermore, we have shown that using such a corpus as a labeled training dataset and incorporating a comprehensive set of word features, it is possible to train models that outperform previous approaches that use only a few features. The best of our models recalled 80 % of unknown words with 71 % precision, compared to the baseline's 62 %.

Our experiments also showed that similar results can be achieved with only 30,000 distinct annotator/word pairs in the training dataset. Training on larger annotated corpora showed only very small increases in performance.

There are many relevant features that our models do not yet account for: such as the word's context, attached particles, inflections, synonyms, and similarity to known words, as well as the reader's native language. In a future work, we plan to investigate how these features can be used to improve our model's accuracy.

We make our annotation data available for researchers who wish to train and/or evaluate L2 unknown word prediction models for Korean. This comes in the form of two resources: a standoff annotated corpus, and a labeled dataset that we extracted from that annotated corpus. The resources are explained in detail in the Appendix.

6. Acknowledgements

We thank the South Korea's National Institute for International Education (NIIED) for giving us permission to use passages from past TOPIK exams, and Mr. Seo Chaneweon for helping us through the approval process. We also thank Skye Payne for helping promote our survey via Twitter and the lead author's wife, Hanna Yancey, for assisting with Korean proofreading and translation. Finally, we especially thank all the annotators who took the time to take our survey.

Appendix: Description of Annotated Corpus & Labeled Dataset

Our annotation data is available for download at `http://lepage-lab.ips.waseda.ac.jp/korean-l2-unknown-words`. In this section, we describe the content and format of this resource.

The annotation data is provided in the form of 2 resources, both available in JSON and XML formats:

- **Labeled Dataset** - A preprocessed list of what words were known and unknown by each annotator, suitable for training and validation of most unknown word prediction models.
- **Standoff Annotated Corpus** - The fully detailed annotation data published as a standoff annotated corpus.

The fields available in each dataset are listed in Table 6. The following sections explain the individual resources in further detail.

Labeled Dataset

Each annotated token is normalized to its base word form with suffixes removed using the `twitter-korean-text` API (Ryu, 2017), and duplicate annotations are removed so that there are is at most one annotation per annotator/word pair. If the duplicate annotations are from different passages, the annotation from the earliest submitted passage is used (since we assume they may have learned the meaning of the word from the previous reading). If they are from the same passage, the word is considered to be unknown if any of the occurrences of it in that passage were annotated as unknown.

Standoff Annotated Corpus

The original annotations are available as a standoff annotated corpus. This is useful, for instance, for training prediction models that take the word's context into account, which is not possible with just the labeled dataset.

In order to extract the original text referenced by the standoff corpus, the TOPIK exam PDFs will have to be downloaded, converted to text, and stripped of whitespace.[11] However, most text extraction tools will not extract the text of these particular PDFs properly. To make it easier to correctly resolve the character offsets against these PDFs, we provide a python script that will correctly extract the text and resolve the character offsets.

Bibliographical References

Bott, S., Rello, L., Drndarevic, B., and Saggion, H. (2012). Can spanish be simpler? lexsis: Lexical simplification for spanish. In *CoLing*, pages 357–374.

Crawford Camiciottoli, B. (2001). Extensive reading in English: Habits and attitudes of a group of italian university efl students. *Journal of Research in Reading*, 24(2):135–153.

Devlin, S. and Tait, J. (1998). The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*, pages 163–173.

Elley, W. B. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language learning*, 41(3):375–411.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Krashen, S. D. (2003). *Explorations in language acquisition and use*. Heinemann Portsmouth, NH.

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

Mason, B. and Krashen, S. (1997). Extensive reading in English as a foreign language. *System*, 25(1):91–102.

Paetzold, G. and Specia, L. (2016). Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. In *ACL (Student Research Workshop)*, pages 103–109.

Tack, A., François, T., Ligozat, A.-L., and Fairon, C. (2016). Evaluating lexical simplification and vocabulary knowledge for learners of french: Possibilities of using the flelex resource. In *Proceedings of LREC 2016: Tenth International Conference on Language Resources and Evaluation*.

Won, Y. (2016). Common european framework of reference for language (cefr) and test of proficiency in korean (topik). *International Journal of Area Studies*, 11(1):39–58.

[11] In order to publish research that use these copywritten documents, one must obtain permission from the National Institute for International Education Development (`www.niied.go.kr/eng/index.do`).

Object/Field Name - Description	Datatype	Resource(s)[†]
Annotator		
`annotator_id` - A unique identifier for the annotator, assigned randomly.	integer	LD, SAC
`joined_datetime` - Indicates when the annotator first registered with the survey website.	datetime	LD, SAC
`reported_TOPIK_level` - The annotator's self-assessed TOPIK level that they reported in their questionnaire. Annotators below level 1 were instructed to chose level 1.	integer	LD, SAC
`estimated_language_level` - The annotator's relative language level as estimated by an analysis of his/her submitted annotations (see Section 4.1.). 0 is the average level among all annotators, and greater values indicate higher levels of proficiency.	float	LD, SAC
`native_language` - The ISO 639-1 code for the annotator's self-reported native language (selected from the list English, Chinese, Japanese, German, Russian, Vietnamese, French, Thai, Italian, or Spanish) or blank if "other" was selected.	string	LD, SAC
`browser_preferred_language` - The ISO 639-1 code for the annotator's preferred language according to the annotator's HTTP header.	string	LD, SAC
`country` - The ISO 3166 alpha-2 code for the country from which we received the annotator's most recent annotated passage, as inferred from the annotator's IP address.	string	LD, SAC
`studyreason_korean_popculture` - The annotator indicated that they were studying Korean because of an interest in Korean popculture.	boolean[*]	LD, SAC
`studyreason_family_or_friends` - The annotator indicated that they were studying Korean to communicate with family or friends.	boolean[*]	LD, SAC
`studyreason_live_in_korea` - The annotator indicated that they were studying Korean because they were living or planning to live in Korea.	boolean[*]	LD, SAC
`studyreason_study_in_korea` - The annotator indicated that they were studying Korean because they wished to study in Korea.	boolean[*]	LD, SAC
`studyreason_work_related` - The annotator indicated that they were studying Korean because of work of career-related reasons.	boolean[*]	LD, SAC
Passage		
`source` - The TOPIK test number, level, and question the passage was extracted from.	string	SAC
`url` - The URL of the file on the TOPIK website from which the passage was extracted.	string	SAC
`level` - The TOPIK level of the test from which the passage was extracted.	string	SAC
`offset` - The character offset of the start of the passage.	integer	SAC
`length` - The character length of the passage, not counting whitespace.	integer	SAC
`submitted_datetime` - Indicates when the annotated passage was submitted.	datetime	SAC
`annotation_duration_seconds` - The amount of time, in seconds, that the annotator spent reading and annotating the passage.	integer	SAC
`comprehension_rating` - The annotator's self-reported comprehension rating of the passage, using the scale provided in Figure 2.	integer	SAC
Annotation		
`offset` - The character offset of the start of the word, relative to the beginning of the TOPIK exam document.	integer	SAC
`length` - The length of the word in characters.	integer	SAC
`checksum` - A hash of the word to use as a checksum to verify that the correct word has been extracted from the original text.	integer	SAC
`base_word` - The base word form of the annotated word, with derivational and inflectional suffixes removed, as provided by the `twitter-korean-text` API (Ryu, 2017).	string	LD
`unkown` - True if the word was annotated as unknown, otherwise false.	boolean	LD, SAC

[†] Indicates the language resources that the field is applicable to: "LD" for Labeled Dataset and "SAC" for Standoff Annotated Corpus.

[*] This section of the questionnaire was not added to the website until July. For those annotators who never completed this section of the questionnaire, this field is blank.

Table 6: Dataset fields

Language Resource References

National Institute for International Education Development in Korea. (2017). *Past TOPIK Exams*. http://www.topik.go.kr/.

Ryu, Will Hohyon. (2017). *twitter-korean-text API*. https://github.com/twitter/twitter-korean-text, 4.0.

Speer, Robert and Chin, Joshua and Lin, Andrew and Nathan, Lance and Jewett, Sara. (2016). *wordfreq Python Package*. Zenodom, 1.6.1.