When is *Wall* a *Pared* and when a *Muro*?: Extracting Rules Governing Lexical Selection

Aditi Chaudhary[†], **Kayo Yin**[†], **Antonios Anastasopoulos**[‡], **Graham Neubig**[†]
[†]Carnegie Mellon University, [‡]George Mason University
{aschaudh,kayoy,gneubig}@cs.cmu.edu antonis@gmu.edu

Abstract

Learning fine-grained distinctions between vocabulary items is a key challenge in learning a new language. For example, the noun "wall" has different lexical manifestations in Spanish – "pared" refers to an indoor wall while "muro" refers to an outside wall. However, this variety of lexical distinction may not be obvious to non-native learners unless the distinction is explained in such a way. In this work, we present a method for *automatically* identifying fine-grained lexical distinctions, and extracting concise descriptions explaining these distinctions in a human- and machine-readable format. We confirm the quality of these extracted descriptions in a language learning setup for two languages, Spanish and Greek, where we use them to teach non-native speakers when to translate a given ambiguous word into its different possible translations. Code and data are publicly released here.[1]

1 Introduction

With increasing globalization there is a widespread prevalence and need for good materials and tools to help people learn languages. Curating such content manually requires a large time and cost investment which poses a challenge particularly for languages where protection and revival efforts are ongoing (Moline, 2020). One of the most important and challenging processes in learning a new language (L2) is vocabulary acquisition (Ellis, 1996; Moore, 1996), which is generally made easy by associating L2 words with words from the first language (L1) (Hulstijn et al., 1996; Watanabe, 1997). In many cases, L1 words or word senses can be unambiguously associated with L2 words. For example "linguistics" and "lingüística" essentially form a one-to-one mapping between English and Spanish. However, different languages carve up the semantic space of the world in different ways leading to *semantic subdivisions*, distinctions made in one language not made in another. For example, "wall" in English is manifested differently in Spanish as "pared" or "muro", as shown in Figure 1, and for an L1 English speaker it may not be immediately obvious when one should be used over the other. A skilled teacher or comprehensive language learning resource may be able to provide explanations that resolve this ambiguity. For example, Robertson (2020) present word definitions in-context for Finnish learners, while CAVOCA (Groot, 2000) takes a learner through various stages of the word acquisition process including word usage, syntax.

Figure 1: Semantic subdivision for the concept 'wall' results in different lexical manifestations in Spanish: 'muro' for *outside wall* and 'pared' for *inside wall* whereas in English both are referred as 'wall'.

In this work, we propose a method to *automatically* discover rules regarding fine-grained lexical distinctions and present L2 learners with concise descriptions derived from them in an interactive framework. Research in L2 vocabulary acquisition (Groot, 2000) has shown that it is effective to combine strategies using explicit definitions and examples in context. However, our work contrasts to most prior work in this field such as CAVOCA (Groot, 2000) or Duolingo,[2] which use learning

[1] https://github.com/Aditi138/LexSelection

[2] https://www.duolingo.com/

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6911–6929
November 7–11, 2021.

content manually created by subject matter experts. This necessity for curation makes it difficult to comprehensively scale this approach to many languages.

Specifically, our framework consists of two steps: i) use a parallel corpus to identify words in L1 which have different lexical manifestations owing to a semantic subdivision in L2, and ii) create human- and machine-readable concise descriptions that allow for easier interpretation of each lexical distinction. First, we extract source (L1) and target (L2) parallel sentences for each shortlisted L1 word. We then extract lexical and semantic features, as well as a label encoding the lexical choice in the target language from these parallel sentences for each L1 word. Finally, we train a prediction model that distinguishes between the lexical choices, and extract human-understandable descriptions from this model. These descriptions could either be used as-is, or could be used as a starting point for further curation by educators.

To confirm the quality of the extracted descriptions, we conduct a study where we use them to teach English native speakers lexical distinctions arising from semantic subdivisions in Spanish and Greek. We make our study interactive by presenting the learning content in the form of *cloze* tests (Taylor, 1953) where the English word to be taught is presented to the learner in context along with extracted concise description. The learner is then required to select the most appropriate lexical choice from the given set. The main methodological contributions therefore are *automated* methods to:

- Identify fine-grained lexical distinctions arising due to semantic subdivisions. To evaluate this and future work, we also create a lexical selection dataset for two language pairs, English-Spanish and English-Greek.
- Extract rules to help humans understand the usage of lexical distinctions in context. Studies with 7 Spanish and 9 Greek learners show that they learn faster when given access to our extracted descriptions; for example they achieve an (avg.) accuracy of 81% within roughly 20 questions, as opposed to more than 40 questions required otherwise.

2 Problem Formulation

For the purpose of this paper, we define the task of *lexical selection* as choosing contextually correct translations from a set of target translations for an ambiguous word in the source language (Lefever and Hoste, 2010). We first define some variables: $\boldsymbol{x} = x_1, x_2, \ldots, x_{|\boldsymbol{x}|}$ denotes a sentence in the source language (L1), $\boldsymbol{y} = y_1, y_2, \ldots, y_{|\boldsymbol{y}|}$ is its translation in the target language (L2) and V_x and V_y are the source and target vocabulary respectively. Given a source sentence $\boldsymbol{x}$ containing an ambiguous word x_i, then trans$(x_i) \subseteq V_y$ denotes the set of its "possible" target translations i.e. words in the target language to which the ambiguous word x_i might be translated (concrete methods to define this set are explained later). The task of lexical selection involves choosing the most appropriate translation $y_i \in$ trans(x_i), and can be performed either by machines or humans.[3] In this work, we particularly focus on *machine-learned methods to help humans learn lexical selection*, extracting lexical selection models that are not only usable by machines, but also interpretable by humans in order to aid the process of learning a new language. We thus plan to extract the rule set $\mathcal{R}_{v_x}$ which governs this lexical selection process in a human- and machine-readable format.

3 Identifying Semantic Subdivisions

In this section, we describe in detail the procedure for identifying L1 words that have different lexical manifestations in L2 owing to semantic subdivisions. For the purpose of this work, we refer to these different lexical manifestations in L2 as lexical choices and the corresponding L1 words as focus words. Our work is "loosely inspired" by ContraWSD (Rios et al., 2018) and SemEval-2013 (Lefever and Hoste, 2013) which construct a dataset for cross-lingual word sense disambiguation, using a semi-automatic approach combining frequency-based heuristics with human supervision. These datasets are restricted to a subset of manually selected nouns (20 for SemEval-2013 and 70-80 for ContraWSD). In contrast, our approach is fully automated going beyond using just frequency-based filters. Furthermore, we do not restrict to any one word class leading to words being identified across different word classes (nouns, verbs, adjectives, adverbs) for both Spanish and Greek.[4]

We start with a parallel corpus $D = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \cdots, (\boldsymbol{x}_{|D|}, \boldsymbol{y}_{|D|})\}$ where $(\boldsymbol{x}_m, \boldsymbol{y}_m)$ denote the source and target sentence pair. Next, we

[3] The notation here refers to single-word translations which are the focus of this work.

[4] More details in Section §5.

extract word alignments automatically using a word aligner that finds sets of pairs of source and target words $A_m = \{\langle x_i, y_j \rangle : x_i \in \boldsymbol{x}_m, y_j \in \boldsymbol{y}_m\}$, where for each word pair $\langle x_i, y_j \rangle$, x_i and y_j are semantically similar to each other within this context.

To focus on translations of the underlying content, as opposed to morphological variations, we then lemmatize all words in both the source and target sentence pairs. Thus, V_x and V_y refer to the lemmatized vocabulary of the source and target language. Going forward, all words refer to their respective lemmatized forms. We perform automatic part-of-speech (POS) tagging, dependency parsing and word sense disambiguation (WSD) on the source side data, resulting in a POS tag and word sense associated with each source word, $\text{tag}(x_i) \in T_x$ and $\text{sense}(x_i) \in S_x$ where T_x is the set of POS tags and S_x is the word sense vocabulary in the source language.

In order to identify the focus words, we extract a list of lemmatized L1 word types v_x filtered by their part-of-speech (POS) tags t_x giving us tuples of the form $\langle v_x, t_x \rangle$. This ensures that we don't conflate meanings across POS tags, because in many languages the semantics of a word can vary widely across its different POS tags.[5] We refer to the extracted tuples $\langle v_x, t_x \rangle$ as focus words for simplicity. We then extract the focus words with their respective lexical choices as follows:

1. Extract translations : For each aligned word pair $\langle x_i, y_j \rangle$ compute the number of times $c(v_x, t_x, v_y)$ the lemmatized source word type ($v_x = \text{lemma}(x_i)$) along with its POS tag ($t_x = \text{tag}(x_i)$) is aligned to the lemmatized target word type ($v_y = \text{lemma}(y_j)$) across the whole corpus. Also, store the number of times the word sense of x_i ($s_x = \text{sense}(x_i)$) appears with the source word type, source POS tag and the translation word type in $g(v_x, t_x, s_x, v_y)$.

2. Filter on frequency : Extract tuples of source types and POS tags $\langle v_x, t_x \rangle$ that have been aligned to at least two target words at least 50 times ($|\{v_y : |c(v_x, t_x, v_y) \geq 50\}| \geq 2$), to account for alignment errors. To avoid ambiguity on the target side, translations aligned to words other than the word v_x in question (at least 3 times) are excluded.

3. Filter on entropy : Remove source tuples that have an entropy $H(v_x, t_x)$ less than a pre-selected threshold. The entropy is computed using the conditional probability of a target translation given the source type and POS tag:

$$p := p(v_y | v_x, t_x) = \frac{c(v_x, t_x, v_y)}{c(v_x, t_x)}$$

$$H(v_x, t_x) = \sum_{v_y \in \text{trans}(v_x, t_x)} -p \log_e p$$

where $\text{trans}(v_x, t_x)$ is the set of target translations for the source tuple $\langle v_x, t_x \rangle$ and $p(v_y | v_x, t_x)$ is the conditional probability of the target translation for this source type v_x and its POS tag t_x. High entropy suggests that a word is ambiguous, with fine-grained distinctions that likely require context to be resolved, and thus is a word we should focus on.

4. Filter on word sense : Remove source tuples whose target translations have distinct source-word senses. For some words, the differences between target translations can be straightforwardly explained by the different source word senses. For example, *banco* in Spanish refers to the financial institution, given by the WordNet (Miller, 1995) sense 'bank.n.02' while *orilla* refers to the edge of a river, outright matched to 'bank.n.01'. For such words, the word sense definitions would be an easy-to-provide rule for learners, but we want to go beyond that. We are interested in finding those words where the word sense information alone is insufficient to distinguish between the lexical choices and are hence likely to be hard for human learners. For a source tuple, use the highest occurring word sense for a given target translation v_y computed as:

$$Q(v_y) = \arg\max_{s_x \in S_x} g(v_x, t_x, s_x, v_y)$$

Finally, retain the source tuples whose target translations all have the same sense, giving us L lexical choices $\text{trans}(v_x, t_x) = \{v_{y_0}, \ldots, v_{y_{|L|}}\}$ for a source tuple (v_x, t_x).

4 Lexical Selection Model

After identifying a set of focus words in the source language, we train a lexical selection model parameterized by $\theta_{\langle v_x, t_x \rangle}$ for each focus word $\langle v_x, t_x \rangle$. We extract the parallel sentences from D that include the focus word and its corresponding lexical choices, denoting them with $D_{\langle v_x, t_x \rangle}$. The model takes as input the source sentences $\boldsymbol{x}_{\langle v_x, t_x \rangle} \in D_{\langle v_x, t_x \rangle}$ and predicts the contextually correct target

[5] "Brown" as a verb (as in "brown the meat") is treated differently from the adjective sense (as in "brown hair").

translation v_y from a set of possible translations $\text{trans}(v_x, t_x) = v_{y_1}, v_{y_2}, \cdots, v_{y_k}$

Since we aim to induce concise, human-understandable explanations of semantic distinctions that can be presented to learners to help them better understand the lexical selection process, we train a prediction model which allows us to easily extract such descriptions for each lexical choice $v_y \in \text{trans}(v_x, t_x)$. In this paper, we use human-readable descriptions of the features learned by a linear model, where these features are defined over a set of lexical and semantic features extracted from the source sentences in $D_{\langle v_x, t_x \rangle}$. For designing features, we take inspiration from prior work which uses extracted contextual information to improve cross-lingual sense disambiguation in machine translation systems (Garcia-Varea et al., 2001; Carpuat and Wu, 2007b,a).

4.1 Model Features

For training a lexical selection model $\theta_{\langle v_x, t_x \rangle}$ for the focus word $\langle v_x, t_x \rangle$, we construct training data from the source-target sentence pairs $D_{\langle v_x, t_x \rangle}$. We focus on features extracted only from the current source sentence, although the framework can be easily extended to include features from the target sentence as well. We represent each source sentence $\boldsymbol{x}_{\langle v_x, t_x \rangle} \in D_{\langle v_x, t_x \rangle}$ with a set of features extracted from the *neighborhood* of the focus word context relevant to the lexical selection process. This neighborhood includes (1) words from the source sentence that occur within a fixed window of the given ambiguous word, and (2) the head and dependents of the focus word as given by the dependency parse of the sentence. For each word in this relevant context, we extract the following lexical features:

- **Lemma** Lemma of the token.
- **WSD** Word sense of the token as extracted from a state-of-the-art word sense disambiguation (WSD) model.
- **Bigram** Bigrams constructed from lemmas of the words present within a fixed window around the focus word. We exclude punctuation and stop words within the window.[6]

4.2 Model Training

To enable extraction of human-understandable descriptions, we use a model that is conducive to interpretation: the linear SVM (LinearSVM; Cortes and Vapnik, 1995), which gives us feature weights $\theta_{\langle v_x, t_x \rangle}$ that can be easily interpreted as the importance of each feature in making the decision. Since there can be n-ary lexical choices for a given focus word, we train using the one-vs-rest (OvR) method which trains one model per each lexical choice v_{y_k}, where data from v_{y_k} are treated as positive examples and data from all other choices as negative, allowing us to extract feature weights for each decision.

4.3 Rule Extraction

As mentioned above, we use human-readable descriptions of the features learned by a linear model to be presented to the human learners. More broadly, we refer to these descriptions as "rules", however these rules could take other forms as well, and we hope that future work by us or others could find other creative ways to induce or define these rules.

For each focus word $\langle v_x, t_x \rangle$, we extract the rule set $\mathcal{R}_{\langle v_x, t_x, v_{y_k} \rangle}$, which is the set of rules for selecting a given lexical choice v_{y_k} from the set of possible choices $\text{trans}(v_x, t_x)$. For this, we extract salient features from the trained model $\theta_{\langle v_x, t_x \rangle}$ for each lexical choice. As mentioned above, using the OvR classification method we get one model per choice v_{y_k}, from which we can then extract the top-N features having the highest weight coefficients for each choice. In order to present this rules in a human-readable form, we create concise rule templates as shown in Appendix B.1.

5 Automated Validation

Since our main research goal is to aid human learners in their learning, we focus on two approaches of evaluation: (a) automated validation, a preliminary evaluation where we validate to what extent our interpretable model can perform cross-lingual lexical selection, and (b) human evaluation (§6) which answers our main question of whether it can teach human learners the usage of L2 words.

For the automated evaluation in particular, we verify several things. First, we check whether our interpretable lexical selection model is able to learn cross-lingual lexical selection at all by measuring its performance compared to selecting the most frequently occurring translation in the corpus for a given focus word ("Frequency"). We also compare with another alternative interpretable model, decision trees (DTree) trained using the same features

[6] Stop words as provided by NLTK(Bird and Klein, 2009)

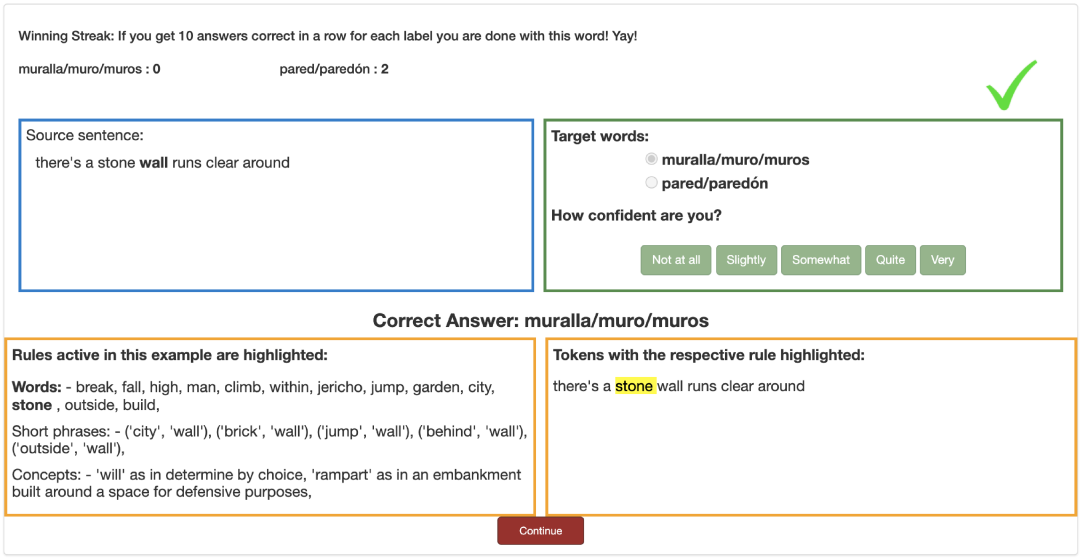


Figure 2: Learning Interface. Rules for the correct answer are displayed to the learner after each question. Individual rules that apply to the given example are highlighted for the convenience of the learner. "wall" here refers to an outside wall and the adjective *stone* serves as a hint in arriving at the correct answer.

as LinearSVM, to validate the choice of SVMs as an interpretable model over other alternatives. Further, we check how our interpretable linear SVM model compares with a "performance skyline"; a less interpretable BERT-based neural model (Devlin et al., 2019) that extracts representations of the source sentence from BERT and trains a classifier to predict the correct lexical choice.

5.1 Setup

Data: We experiment with two L2 languages: Spanish and Greek. These languages were chosen due to (1) availability of parallel corpora with which to train models, and (2) availability of linguists and annotators to verify and analyze the data used in our experimental setting. For Spanish we use 10 million English-Spanish parallel sentences from OpenSubtitles (Lison and Tiedemann, 2016), Tatoeba, TED (Tiedemann, 2012), and Europarl (Koehn, 2005). [7] For Greek, we use 31 million English-Greek parallel sentences extracted from OpenSubtitles. For word alignment we use the AWESOME aligner (Dou and Neubig, 2021), for lemmatization we use spaCy (Honnibal et al., 2020), for POS tagging and dependency parsing we use Stanza (Qi et al., 2020), and for English WSD we use EWISER (Bevilacqua and Navigli, 2020).[8]

Using our automatic pipeline (§3), we identify 157 English words which have fine-grained distinctions in Spanish and 707 English words for Greek. Among these, for Spanish there are 127 nouns, 15 verbs, 10 adjectives, 5 adverbs and, for Greek there are 452 nouns, 123 verbs, 126 adjectives and 6 adverbs. Along with nouns which do account for much of the data, we do find significant number of verbs and adjectives also exhibiting ≥ 2 lexical choices.[9] A manual inspection by a Greek-English bilingual speaker revealed that most automatically created lexical choices were correct. In just a couple of cases, lemmatizer errors lead to two choices corresponding to the same actual lemma (which were manually corrected for the user studies).

Model: We train a linear SVM lexical selection model with `sklearn` (Pedregosa et al., 2011) for each L1 focus word and divide the extracted parallel sentences into a train/test split with a 80-20 ratio per lexical choice. We perform 5-fold cross-validation to select the best model hyperparameters (detailed in Appendix A.2) from which we then extract the top-20 features for each lexical choice to form our rule set. Details on the setup of DTree and BERT are also in Appendix A.2.

5.2 Results

Table 1 shows the test accuracy averaged across all focus words for both Spanish and Greek. Regarding our underlying questions, we first find that Lin-

[7] We use only 1 million sentences from Europarl because we found sentences from Europarl to contain fewer semantic subdivisions owing to the very specific domain of the dataset.

[8] POS tagging, dependency parsing and WSD is required *only* for the source language, here English.

[9] More details in Appendix A

Lang.	Model	Test Accuracy				
		All	nouns	verbs	adj.	adv.
Spanish	Frequency (Baseline)	59.43	59.36	60.17	60.67	53.03
	DTree	62.40	62.45	61.57	65.22	54.82
	LinearSVM	**66.87**	**67.41**	**65.34**	**66.91**	**56.29**
	BERT	70.72	71.75	69.04	67.31	54.07
Greek	Baseline	58.56	59.48	53.04	60.48	61.82
	DTree	63.79	64.49	59.74	65.39	61.13
	LinearSVM	**66.46**	**67.09**	**63.30**	**67.51**	**64.98**
	BERT	71.74	70.91	78.14	68.86	62.76

Table 1: The interpretable LinearSVM lexical selection model is almost on par with the BERT skyline.

earSVM significantly outperforms both Frequency and DTree by a significant margin, indicating that it is both learning to perform lexical selection to a significant degree, and outperforming other reasonable alternatives for interpretable models.[10] This gives us confidence to proceed to use it in our following human learning experiments. Interestingly, our interpretable LinearSVM model is within 97% relative accuracy of the skyline BERT model (just 2.09 percentage points behind). The fact that the more complicated but less inherently interpretable BERT model is better overall paves the way for future work in applying model interpretation techniques (Abnar and Zuidema, 2020, *inter alia*) to extract human-interpretable rules for lexical selection, although this is beyond the scope of the current paper.[11] We find that lexical selection accuracy varies by part of speech; all models perform poorly on adverbs with (avg.) gain of only +0.97 points over the baseline (c.f. with gains of +8.04 for nouns, +5.16 for verbs, +6.24 for adjectives).

6 Evaluation with Human Learners

We move to our main evaluation where we examine *how effective our extracted rules are in aiding human learners* in understanding the distinctions in L2 words.

6.1 Evaluation Methodology

We take inspiration from existing research on second language acquisition (SLA) to design our evaluation method. For instance, Groot (2000) highlights the different learning strategies which are based on generally accepted language acquisition theories (Nation, 2005; Richards et al., 1999), which suggest that a learner is required to go through different levels of language processing for effectively learning vocabulary. In particular, Groot (2000) empirically show that some of these levels can be accelerated with appropriate design of the language tasks by combining learning strategies which use both examples in context and definitions for effective learning. Our cloze-style tasks are essentially examples in context showing the word usage in a given context and the extracted rules are a proxy for human-provided definitions.

Specifically, we set up an interactive exercise where a human learner is presented with the English focus word in context, along with a set of possible L2 (Spanish or Greek) lexical choices. The learner is then required to select one of the possible lexical choices, based on which they think correctly translates the focus word in the given source context. They must also mark how confident they are in their answer ("Not at all", "Slightly", "Somewhat", "Quite" or "Very"). After they select the answer to each question, they are told the correct answer immediately. For each focus word, we ask the learner to answer up to N multiple-choice questions in sequence, which contain roughly equal number of questions for each lexical choice.

In order to evaluate how effective the extracted rules are in aiding the learning process, we perform this study in two setups, a baseline one without rules, and one using our proposed system with rules.

Baseline Setup: In this setup, the human learner does not have access to any rules and immediately starts answering questions. If the learners do not know the target language, they are likely to start out with approximately chance accuracy (e.g. 50% if there are two choices), but as they are given feedback they may be able to grasp the patterns under which one particular translation or another is used, and gradually rise above chance accuracy.

Proposed Setup: In the proposed setup, before starting the task, the learner is shown brief rules regarding when you would use each possible lexical choice $v_{y_k} \in \text{trans}(v_x, t_x)$, constructed from the rule set $R_{\langle v_x, t_x, v_{y_k} \rangle}$. They take as much time as they want to review these rules, and then move to answering questions. The interface for answering questions is the same as the baseline, but below the task screen they can review the rules of different translation choices (figures in Appendix B.2). On selecting a choice, the learner is shown the correct answer accompanied with its corresponding

[10] Individual scores per focus word listed in Appendix A.2

[11] Overall accuracy is low, with even BERT getting 70%, possibly due to lack of sufficient source-side context. OpenSubtitles comprises of movie dialogues where the sufficient context could span more than a single sentence.

human-readable rules of *only* the correct answer. Further, we highlight those individual rules that helped decide the correct answer (Figure 2) for the convenience of the learner. By highlighting it in the two bottom panes, we hope to draw the learner's attention to these hints and thus strengthen the understanding of the underlying concept.

In this setting, the annotator may achieve non-chance accuracy even at the very beginning of answering questions, as they have been given an explanation regarding the underlying rules that they can leverage in answering questions. The accuracy will likely further increase as they practice and become familiar with actual examples and how the extracted features apply to them.

6.2 Experimental Details

We select native English speakers, 7 for the Spanish study and 9 for the Greek study.[12] Each annotator is presented with the same set of English words or tasks. For each study, half of the words will be annotated using the baseline setup and remaining half with the proposed setup. To ensure an unbiased setup, we randomize whether each focus word uses rules or not, while ensuring that at least half the annotators see the proposed setup and the other half perform the same task in the baseline setup for each word. We further shuffle the order in which the words are presented. For each English word, we select up to 40 examples each for the respective lexical choices. However, as an incentive, we end a task early if the annotator correctly answers 10 questions straight in a row for each lexical choice. We explain below the selection procedure for the English words used in the experiments.

Word Selection: In an ideal situation, we would like to conduct these experiments for all identified English focus words, but this would involve annotating thousands of sentences, requiring a large time commitment from the annotators. Instead, we shortlist a handful of words using the following automated procedure: First, for a given L2 study, we sort all focus words by the number of available data points ($D_{\langle v_x, t_x \rangle}$). Next, from the trained lexical selection model $\theta_{\langle v_x, t_x \rangle}$ we compute an F1-score for each lexical choice and filter focus words where the model gets an F1 > 0.5 for each lexical choice. Finally, we select upto 10 focus words with the most data points that fit the above condition. For each word ($\langle v_x, t_x \rangle$), we then select 40 *representative* examples for each lexical choice (see paragraph below). Details on the shortlisted words can be found in Appendix B.3.

Representative Example Selection: To facilitate an effective learning process, we present examples to the learner that have *sufficient source-side context* required for correctly identifying the target-side lexical choice. This is important because there are examples in the corpus where the sufficient context requires context spanning over multiple sentences. To make our learning content both concise and effective, we focus only on context self-contained in a single sentence. Further to efficiently conduct a high-quality study, we enlist help from native speakers of the L2 language to filter the required sentences. We note, though, that the relevant sentences could also be potentially filtered automatically (left for future work).

To get such meaningful examples, we present bilingual English-Spanish and English-Greek speakers with the English sentence containing the focus word and the set of possible lexical choices in Spanish and Greek respectively. They then select the word which best suits the given context and mark their confidence in the selection. The interface for the example selection is the same as Figure 2 (but without rules). We collect these annotations from multiple native speakers and only keep those sentences on which all native speakers agree (see Appendix B.3 for details).

6.3 Results and Discussion

To confirm whether the extracted rules are effective to the learning process, we examine the following questions:

Do the extracted rules result in increased learner accuracy? We compute the learner accuracy across all learners for each L2 study. If a learner attains higher accuracy with fewer attempted examples for the experiment with rules than without, then the extracted rules could be considered effective in the learning process. However, we cannot directly use the learner accuracy as-is because of the possibility of other sources of variability such as (a) underlying learner ability, as some learners may be more proficient than others, (b) underlying task difficulty, as some words may be harder to disambiguate than others, or (c) word ordering, as learners may become proficient as they do more tasks. Therefore, we use a mixed

[12] We allow participants who know other languages but none that are familiar with the L2 or its related languages.

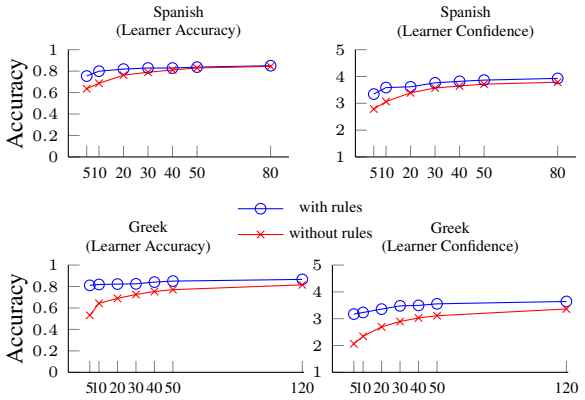


Figure 3: Learner accuracy and confidence in correct answers with and without access to rules against the number of attempted examples (x-axis). Learners achieve higher accuracy with increasing confidence with fewer examples when they have access to rules.

effects model (McLean et al., 1991), which models *random effects* and *fixed effects* to account for such random variability. Random effects are variables responsible for random variation such as task-identity, task-order and the learner, while fixed effects such as the presence of rules are the variables of interest for determining the response variable i.e. learner accuracy. A linear mixed-effect model (LME) is defined as: $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon$ where $\mathbf{y}$ is the learner accuracy, β and $\mathbf{u}$ are the fixed-effect and random-effect regression coefficients, $\mathbf{X}$ and $\mathbf{Z}$ are the respective design matrices and ϵ the noise.

We fit LME models on our data by varying the number of first n attempted examples $n = [5, 10, 20, 30, 40, 50, \text{all}]$. Each fitted LME model gives us an intercept which informs us of the learner accuracy in absence of rules, and the fixed-effect coefficient β which informs us about the gain with rules. As shown in Figure 3, it is clear that learners having access to our automatically extracted rules achieve higher accuracy with fewer examples as compared to without. As expected, with an increasing number of attempted examples the gap in accuracy between the two settings reduces. Interestingly, we find that the rules still have a significant effect on the learner's confidence even later in the learning process. This suggests that with our rules learners require fewer examples to infer the patterns governing each lexical choice and further get more confident in their understanding. This is encouraging as in true settings the learning exercise would be conducted for *every* focus word that the learner is attempting to learn, and because this process will have to be repeated many times, making it more efficient is of significant value. In

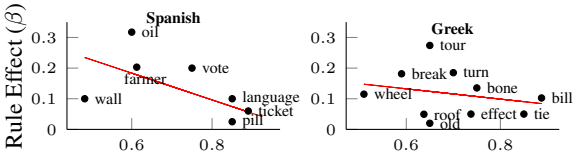


Figure 4: Rules help more for words where learners do worse. x-axis is the (avg.) learner accuracy (without rules) for first 20 examples.

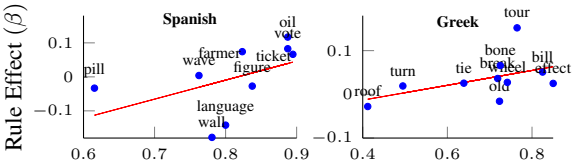


Figure 5: Rules help more for words where the model performs well. x-axis is model accuracy per word.

Appendix B.4 we report the p-value for the fitted LME models which shows that the positive gains from the presence of rules are most significant for $\leq$20 examples for Spanish and for all examples for Greek.

Overall, we find our extracted rules help both Spanish and Greek learners in their learning process. We note that the results on Greek are promising as it does not enjoy the same luxuries as Spanish in having a high-quality lemmatizer or word aligner. This is encouraging especially for researchers involved in the revival efforts of endangered languages.

Do the extracted rules result in increased learner confidence? While answering the questions we ask the learner to mark how confident they are in their answer. As before, we fit LME models for each n using annotator confidence as the response variable $\mathbf{Y}$ and presence of rules as the fixed-effect. We find that the learners' confidence in the correct answer increases more when they are provided rules (Figure 3) for both the languages.

Do the extracted rules help some words *more* over others? Since the focus words may vary in the difficulty level, we check if our extracted rules are more effective for some words over others. So, we fit a LME model on each focus word and compute the β coefficient to measure the effect of rules on learner accuracy after 20 attempted examples.[13] We plot the β coefficient with the accuracy (averaged across all learners) for each focus word when they didn't have rules in Figure 4 and find that words on which the learners performed the worst

[13]Because analysis revealed that rules are more effective earlier in the learning process.

such as *wall, oil, farmer,* and *vote* for Spanish, benefit most by our rules. Similar observations can be seen for Greek where learners are benefited more for words (*break, wheel, tour, old, roof*) on which they performed the worst. Some of these words, in fact, indeed have finer semantic subdivisions than the rest. For instance, the choices for *farmer*: *agricultor* refers to exclusively the one who works the land, harvests, sows, etc., whereas *granjero* is less formal referring to the one who manages a farm, or works or lives on it.[14] This analysis shows that, encouragingly, our rules are especially helping learners with more difficult words.

We also plot the β coefficient with the lexical model accuracy (Figure 5) and find a positive correlation, meaning that rules help more for words where the model performs well. This suggests that if we can develop more accurate models with an equal level of interpretability, the learning effect might become even stronger.

7 Related Work

Computer-assisted language learning CALL systems have been increasingly using NLP for creating learning content. Both SMILLE (Zilio et al., 2017) and WERTi (Meurers et al., 2010) aim to help the text understanding process by highlighting linguistic structures using hand-written rules and automatically acquired syntactic analysis. *Apertium* (Tyers et al., 2012), a rule-based MT system, while not aimed at language learning, does use human- and machine-readable rules, whose formalism can account for only fixed-length ordered contexts restricting their application. Further, these rules use a combination of only lemma and POS tags while our framework uses more features.

Cross-lingual word sense disambiguation CL-WSD disambiguates a word in-context by providing appropriate translation across languages. Lefever and Hoste (2010) construct a dataset (25 ambiguous English nouns across six languages) semi-automatically from parallel corpora which are then verified by expert translators. Such lexical choice tasks have been created also for evaluating MT systems (Rios Gonzales et al., 2017; Rios et al., 2018). However, these methods cover a limited set of words (mostly nouns) and require some manual intervention during the data creation process. To the best of our knowledge, our proposed pipeline is the only fully automated one that extracts several ambiguous words across multiple POS tags.

8 Future Work

While we have demonstrated the efficacy of our extracted rules in teaching new words for two languages, we plan to apply our framework on much less-resourced languages which have fewer available learning resources where learners would benefit more from an automated system. We also plan to use automated methods such as selection using model confidence to select 'representative' examples for the learning setup instead of using the native speakers. Furthermore, multimodal features have proven their utility in automatic methods for lexical acquisition (Hewitt et al., 2018), and we plan to examine their effectiveness for L2 learning.

Acknowledgements

The authors are grateful to the anonymous reviewers who took the time to provide many interesting comments that made the paper significantly better. We would also like to thank Nikolai Vogler for the original interface for data annotation, and all the learners for their participation in our study, and without whom this study would not have been possible or meaningful. This work is sponsored by the Waibel Presidential Fellowship and by the National Science Foundation under grant 1761548.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864, Online. Association for Computational Linguistics.

Edward Loper Bird, Steven and Ewan Klein. 2009. *Natural Language Processing with Python*.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. 1984. *Classification and regression trees*. CRC press.

Marine Carpuat and Dekai Wu. 2007a. How phrase sense disambiguation outperforms word sense disambiguation for statistical machine translation. *Proceedings of TMI*, pages 43–52.

[14] This is based on explanations collected from native Spanish speakers, which can also be found in Appendix B.4

Marine Carpuat and Dekai Wu. 2007b. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 61–72.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. *arXiv preprint arXiv:2101.08231*.

Nick C Ellis. 1996. Sequencing in sla: Phonological memory, chunking, and points of order. *Studies in second language acquisition*, pages 91–126.

Ismael Garcia-Varea, Franz Josef Och, Hermann Ney, and Francisco Casacuberta. 2001. Refined lexicon models for statistical machine translation using a maximum entropy approach. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 204–211.

Peter JM Groot. 2000. Computer assisted second language vocabulary acquisition. *Language learning & technology*, 4(1):56–76.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. 2018. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576, Melbourne, Australia. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Jan H Hulstijn, Merel Hollander, and Tine Greidanus. 1996. Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The modern language journal*, 80(3):327–339.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Els Lefever and Veronique Hoste. 2010. SemEval-2010 task 3: Cross-lingual word sense disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 15–20, Uppsala, Sweden. Association for Computational Linguistics.

Els Lefever and Véronique Hoste. 2013. SemEval-2013 task 10: Cross-lingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 158–166, Atlanta, Georgia, USA. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert A McLean, William L Sanders, and Walter W Stroup. 1991. A unified approach to mixed linear models. *The American Statistician*, 45(1):54–64.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18, Los Angeles, California. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Emily Ariel Moline. 2020. Indigenous language teaching policy in california/the us: What's left unsaid in discourse/funding. *Issues in Applied Linguistics*, 21(1).

Zena Moore. 1996. *Foreign language teacher education: Multiple perspectives*. University Press of America.

ISP Nation. 2005. Teaching and learning vocabulary. In *Handbook of research in second language teaching and learning*, pages 605–620. Routledge.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning*, 1(1):81–106.

Jack C Richards, David Singleton, and Michael H Long. 1999. *Exploring the second language mental lexicon*. Cambridge University Press.

Annette Rios, Mathias Müller, and Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 588–596, Belgium, Brussels. Association for Computational Linguistics.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19, Copenhagen, Denmark. Association for Computational Linguistics.

Frankie Robertson. 2020. Show, don't tell: Visualising Finnish word formation in a browser-based reading assistant. In *Proceedings of the 9th Workshop on NLP for Computer Assisted Language Learning*, pages 37–45, Gothenburg, Sweden. LiU Electronic Press.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Francis M. Tyers, Felipe Sánchez-Martínez, and Mikel L. Forcada. 2012. Flexible finite-state lexical selection for rule-based machine translation. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, pages 213–220, Trento, Italy. European Association for Machine Translation.

Yuichi Watanabe. 1997. Input, intake, and retention: Effects of increased processing on incidental learning of foreign language vocabulary. *Studies in second language acquisition*, pages 287–307.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2017. Using NLP for enhancing second language acquisition. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 839–846, Varna, Bulgaria. INCOMA Ltd.

A Automated Evaluation

A.1 Identifying Semantic Subdivisions

In Section 3, we describe the procedure for identifying focus words in L1. For the step of *filter on entropy* within that procedure, we use a threshold of 0.69 so focus words having an entorpy $H > 0.69$ are selected in that step. For binary lexical choices, an ambiguous word would be aligned to each choice with uniform distribution and the entropy in that case would be 0.69. Hence we are interested in words that exceed this minimum threshold. In Figure 6 we show the distribution of number of lexical choices for all extracted focus words, filtered by each POS tag for Spanish and Greek. We check the CEFR levels[15] which measure the reading proficiency in a language. We use the automated tool provided by Duolingo[16] (currently available only for Spanish and English) to get these levels and find that 60% of the extracted Spanish lexical choices belong to the B level which is the intermediate level and 20% belong to the advanced level. This suggests that the identified words are indeed more challenging.

A.2 Model Hyperparameters and Results

For the LinearSVM and DTree models, we clean the data to remove punctuation and extract features within a 3-word window of the focus word. As mentioned before, we train a lexical selection model for each focus word and in Table 5, 6, 7, 8 we report the individual accuracy for the test accuracy for LinearSVM, DTree, BERT and the baseline method across both Spanish and Greek. We also provide the train accuracy for our main model, LinearSVM.

LinearSVM We perform a grid search over the following hyperparameters: `C` = [0.001, 0.01], `class weight` =['balanced', None].

DTree We also experimented with other interpretable models such as decision trees (Quinlan, 1986) using the CART algorithm (Breiman et al., 1984), however we found them to be performing worse than the SVM model. We used the following hyperparameters: `criterion` = [gini, entropy], `max depth` = [6,15], `min impurity decrease` = $1e^{-3}$.

[15] https://en.wikipedia.org/wiki/Common_European_Framework_of_Reference_for_Languages

[16] https://cefr.duolingo.com/

BERT We compare the interpretable models LinearSVM and DTree with more complex neural model based on the popular BERT (Devlin et al., 2019). We retain the same hyperparameters as the original paper using 768 dimentions for the encoder representations. We train the model for 20 epochs, using the AdamW optimzer with a learning rate of $5e-5$.

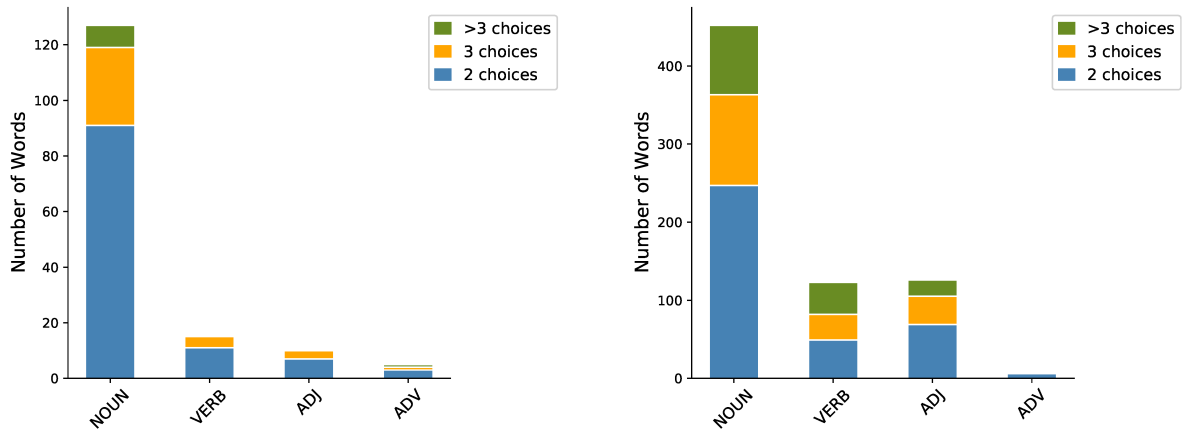
B Human Evaluation

B.1 Rule Templates

The human-understandable "rules" are essentially those features from the training set which the model thought were important for determining the correct label (i.e. features that were given higher weights for a given label). In particular, for each label (i.e. *muro* or *pared*), we choose the top-20 features. We then group these individual features together by their feature types, for instance, all the bigram features are grouped under the category called "Short Phrase", lemma features are grouped under the category "Words", and the WSD features are first expanded into their natural language form using the WordNet (Miller, 1995) knowledge base and then grouped under the category "Concepts", as shown in Table 2.

B.2 Language Learning Interface

In our proposed language learning setup, the learner is first presented with a screen showing concise explanations for each lexical choice (Figure 7(a)). They can take as much time as they require for reviewing the rules and then proceed to the tasks. Within each task, the learner is then shown an English sentence with the focus word highlighted and a set of possible lexical choices. The page also displays the concise explanations for the learner to refer if they wish to (Figure 7(b)). The learner is required to select one of the lexical choices and mark how confident they are in their answer. Once submitted, the learner is immediately shown the correct answer along with individual rules that applied to that example highlighted (Figure 2 in main text). Learners took (avg.) 3-4 hours in total to complete all tasks. Table 3 presents the tasks performed by the respective Spanish and Greek learners. Since English speakers might not be familiar with the Greek alphabet, we display the English transliteration of the respective Greek words. Some of the lexical choices (e.g. *muralla/muro/muros*) contain multiple inflec-



(a) Spanish

(b) Greek

Figure 6: Distribution of number of lexical choices for each POS tag.

Lexical Choice	**feature** $\longrightarrow$ $\langle$ **rule name** $\rangle$ $\langle$ **feature value** $\rangle$
muro	Bigram $\longrightarrow$ Short phrases: ('climb', 'wall'), ('city', 'wall'), ('brick', 'wall') Lemma $\longrightarrow$ Words: break, climb WSD $\longrightarrow$ Concepts: 'city' as in a large and densely populated urban area (city.n.01)
pared	Bigram $\longrightarrow$ Short phrases: ('face', 'wall'), ('hang', 'wall'), ('picture', 'wall') Lemma $\longrightarrow$ Words: ear, hang, room

Table 2: Human-readable rules extracted for the ambiguous word *wall* (top-6 rules per lexical choice).

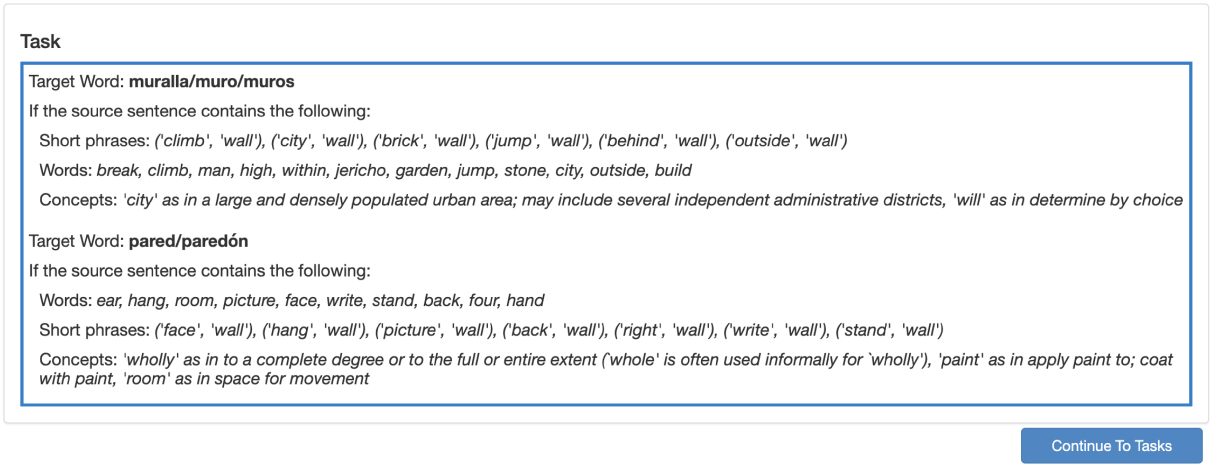
tions of the same lemma (*muro*). This is due to errors in the automatic Spanish lemmatizer which failed to correctly map the inflections to a single lemma. We therefore run an edit-distance based post-processing to combine lexical choices having the same prefix. We note that this simple heuristic might not be ideal for languages such as Indonesian that use affixes and/or reduplication with far-from-perfect lemmatizers; nevertheless such a post-processing method, when applied carefully, helps fix many of the erroneous lemmatization issues.

B.3 Representative Example Selection

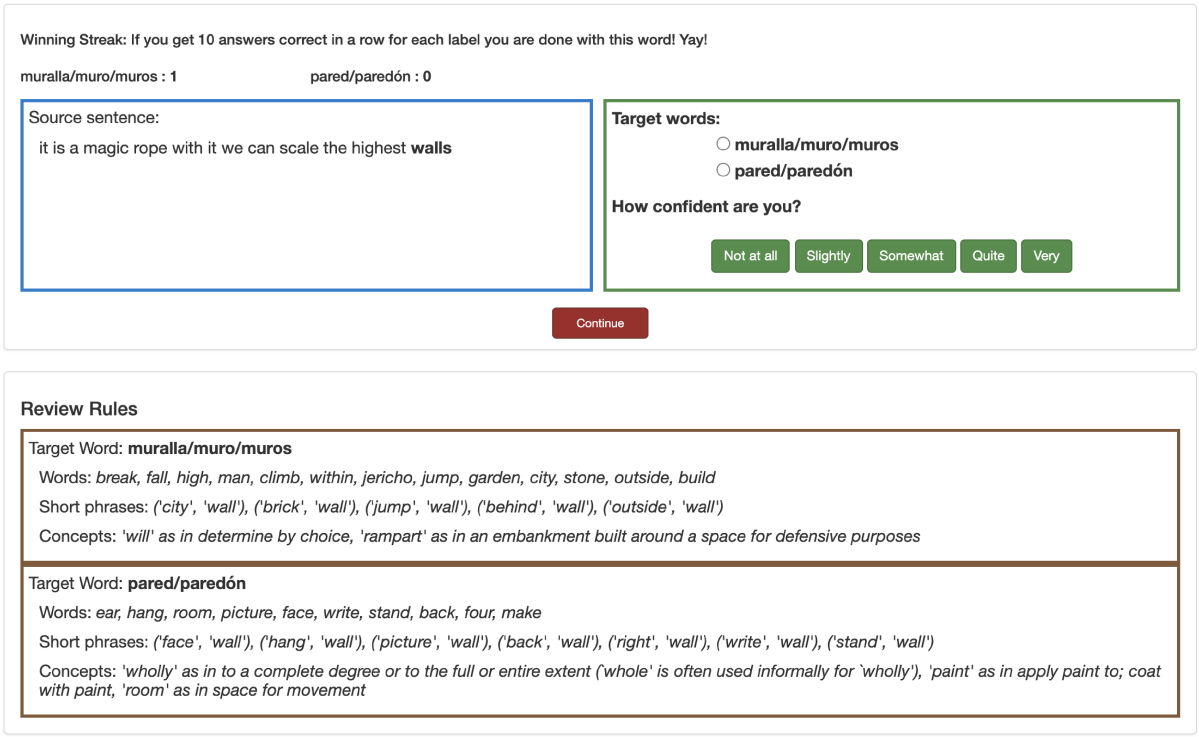
The shortlisted words for both the Spanish and Greek study can be found in Table 3. We use native speakers to filter sentences that have sufficient context for correctly identifying a lexical choice. We enlist 3 Spanish native speakers who each annotate roughly 200 examples each for 10 English focus words. The inter-annotator agreement for Spanish, computed using Fleiss' kappa is 0.77. For Greek, we use 2 native speakers to annotate 10 English words. For 7 out of 10 words we did not always have access to 2 native speakers so we relied on a single expert annotator. The (avg.) inter-annotator agreement for the remaining 3 words (*tour, tie, bill*) between the two annotators is 0.83. Of the 10 selected words, we discard words/lexical choices which have $<$ 10 examples on which all native speakers agree (Table 3) giving us 9 English words for the Spanish study and 10 English for the Greek study.

B.4 Results

In Table 4 we report the p-values for the linear mixed effect (LME) models fitted on predicting learner accuracy with rules as fixed-effect. The results show that the positive effect of rules on accuracy is statistical significant up to first 20 attempted examples for Spanish and up to all examples for Greek.



(a) Rules for "pared" vs "muro".



(b) Rules for the correct answer are displayed to the learner after each question. Individual rules which apply to the given example are highlighted for the convenience of the learner.

Figure 7: User interface for human language learning experiment.

Spanish		Greek	
(en) word	(es) lexical choices	(en) word	(el) lexical choices
wall.*N*	muralla/muro/muros: 33, pared/paredón: 60	bill.*N*	χαρτονόμισμα: 40, λογαριασμός: 40, νόμος/νομοσχέδιο: 40 (chartonómisma, logariasmós, nómos/nomoschédio)
farmer.*N*	agricultor: 29, granjero: 48	tour.*N*	ϑητεία:23, περιοδεία: 29, ξενάγηση: 33 (thitía, periodeía, xenágisi)
figure.*N*	cifra/cifras: 87, figura: 85	break.*JJ*	σπάω: 40, ράγομαι: 40, ξεσπάω: 40, διαρρηγνύω: 40 (spáo, rágomai, ksespáo, diarrignío)
vote.*N*	votemos/voto: 77, votación: 75	turn.*JJ*	στρίβω: 40, χαμηλώνω: 40, απορρίπτω: 40, καταδίδω: 40, σβήνω: 34 (strívo,chamilòno, aporrípto,katadído, svíno)
oil.*N*	aceite: 81, óleo/petróleo/petrolera/petrolero: 74	roof.*N*	ταράτσα: 40,οροφή: 40, στέγη: 39 (tarátsa, orofí, stégi)
wave.*N*	onda: 55, ola: 40, *oleado: 0*	wheel.*N*	τροχός: 40, ρόδα: 40, τιμόν: 40 (trohós, róda, timóni)
pill.*N*	pastilla: 41, somnífero: 27, *píldora: 3*	old.*JJ*	αρχαίος: 40, κλασικ: 21, έτος: 40, ηλικιωμένος: 40, παραδοσιακός: 36 (archaios, klasikos, etos, elikiomenos,paradosiakos)
language.*N*	idioma: 52, lenguaje: 68	turn.*JJ*	στρίβω: 40, χαμηλώνω: 40, απορρίπτω: 40, καταδίδω: 40, σβήνω: 34 (strívo, chamilóno, aporrípto, katadído,svíno)
ticket.*N*	multa: 24, boleto: 23, *pasaje: 0*	effect.*N*	παρενέργεια: 40, επίδραση: 40, εφέ: 40 (parenírgeia, epídrasi, efé)
servant.N	sirvienta/sirviente: 39, *servidor/servidora: 8*, *siervo/siervos: 10*	bone.*N*	μυελός: 40, οστό: 40, Μπόουν: 40 (myelós, ostó, bone)

Table 3: Example tasks with their lexical choices selected for Spanish and Greek learning setup. Words/choices marked in *red* are discarded from the language learning setup as they have $\leq$ 10 filtered examples from the represenative example selection step.

Number	Fixed-effect coefficient (β)	Spanish p-value	Greek p-value
5	0.118	0.013**	$4.50e^{-09}$***
10	0.112	0.009***	$1.64e^{-07}$***
20	0.056	0.070*	$1.32e^{-06}$***
30	0.039	0.131	$4.23e^{-05}$***
40	0.017	0.462	$7.22e^{-05}$***
50	0.007	0.718	0.00015***
All	0.006	0.739	0.00173**

Table 4: p-value tests show that the fixed-effect of presence of rules for predicting learner accuracy is statistical significant up to first 20 attempted examples for Spanish and up to all examples for Greek. Significance codes: '***': 0.01, '**': 0.05, '*': 0.1.

Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT	Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT	Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT
specifically.*RB*	0.67 - 0.7 - 0.67 / 0.67 - 0.72	block.*N*	0.58 - 0.69 - 0.92 / 0.79 - 0.86	desert.*V*	0.53 - 0.55 - 0.96 / 0.73 - 0.95
transfer.*N*	0.64 - 0.71 - 0.92 / 0.69 - 0.72	slipper.*N*	0.7 - 0.7 - 0.7 / 0.7 - 0.63	mushroom.*N*	0.42 - 0.39 - 0.91 / 0.44 - 0.49
pen.*N*	0.73 - 0.74 - 0.73 / 0.73 - 0.73	foundation.*N*	0.43 - 0.5 - 0.96 / 0.64 - 0.69	mercy.*N*	0.58 - 0.67 - 0.83 / 0.72 - 0.77
fry.*V*	0.57 - 0.71 - 0.84 / 0.69 - 0.88	bug.*N*	0.57 - 0.6 - 0.82 / 0.6 - 0.48	toast.*N*	0.64 - 0.72 - 0.96 / 0.87 - 0.94
cord.*N*	0.52 - 1.0 - 0.98 / 1.0 - 1.0	waste.*N*	0.48 - 0.53 - 0.87 / 0.74 - 0.8	opponent.*N*	0.66 - 0.66 - 0.67 / 0.66 - 0.59
figure.*N*	0.53 - 0.55 - 0.93 / 0.76 - 0.93	hood.*N*	0.55 - 0.61 - 0.94 / 0.74 - 0.92	steak.*N*	0.66 - 0.34 - 0.66 / 0.66 - 0.61
poker.*N*	0.62 - 0.62 - 0.62 / 0.62 - 0.54	heel.*N*	0.53 - 0.72 - 0.9 / 0.71 - 0.85	plot.*N*	0.6 - 0.6 - 0.99 / 0.76 - 0.88
plumber.*N*	0.57 - 0.57 - 0.57 / 0.57 - 0.65	replacement.*N*	0.6 - 0.6 - 0.61 / 0.6 - 0.67	thick.*JJ*	0.59 - 0.73 - 0.98 / 0.73 - 0.69
pee.*V*	0.61 - 0.61 - 0.62 / 0.61 - 0.58	greedy.*JJ*	0.68 - 0.68 - 0.68 / 0.68 - 0.75	barber.*N*	0.75 - 0.69 - 0.76 / 0.75 - 0.74
puppet.*N*	0.53 - 0.53 - 0.9 / 0.51 - 0.46	basket.*N*	0.66 - 0.66 - 0.69 / 0.66 - 0.66	marble.*N*	0.51 - 0.71 - 0.98 / 0.8 - 0.93
bowl.*N*	0.6 - 0.6 - 0.96 / 0.72 - 0.72	dump.*N*	0.61 - 0.64 - 0.9 / 0.64 - 0.67	lipstick.*N*	0.51 - 0.49 - 0.93 / 0.4 - 0.53
appeal.*N*	0.75 - 0.8 - 0.95 / 0.82 - 0.93	promote.*V*	0.74 - 0.73 - 0.74 / 0.74 - 0.64	brush.*N*	0.55 - 0.58 - 0.96 / 0.66 - 0.87
fan.*N*	0.31 - 0.3 - 0.89 / 0.43 - 0.59	disappoint.*V*	0.68 - 0.68 - 0.68 / 0.68 - 0.74	persuade.*V*	0.7 - 0.7 - 0.74 / 0.7 - 0.58
mob.*N*	0.7 - 0.74 - 0.98 / 0.74 - 0.93	romance.*N*	0.58 - 0.64 - 0.98 / 0.73 - 0.79	raincoat.*N*	0.65 - 0.65 - 0.64 / 0.65 - 0.56
properly.*RB*	0.42 - 0.42 - 0.42 / 0.42 - 0.41	jungle.*N*	0.55 - 0.55 - 0.77 / 0.56 - 0.53	nail.*N*	0.6 - 0.67 - 0.93 / 0.81 - 0.9
hobby.*N*	0.58 - 0.58 - 0.68 / 0.58 - 0.63	pupil.*N*	0.53 - 0.56 - 0.96 / 0.67 - 0.79	regularly.*RB*	0.64 - 0.66 - 0.96 / 0.7 - 0.5
stew.*N*	0.59 - 0.41 - 0.6 / 0.59 - 0.62	farmer.*N*	0.66 - 0.7 - 0.93 / 0.77 - 0.84	pipe.*N*	0.5 - 0.52 - 0.93 / 0.62 - 0.83
bait.*N*	0.54 - 0.62 - 0.58 / 0.59 - 0.5	eve.*N*	0.96 - 0.96 - 0.99 / 0.96 - 0.96	transitional.*JJ*	0.56 - 0.56 - 0.9 / 0.56 - 0.52
trunk.*N*	0.74 - 0.76 - 0.77 / 0.74 - 0.81	oil.*N*	0.55 - 0.63 - 0.95 / 0.85 - 0.89	ancestor.*N*	0.56 - 0.56 - 0.95 / 0.55 - 0.58
port.*N*	0.41 - 0.61 - 0.98 / 0.79 - 0.96	lock.*N*	0.8 - 0.85 - 0.91 / 0.85 - 0.89	cripple.*N*	0.63 - 0.63 - 0.63 / 0.63 - 0.63
shell.*N*	0.41 - 0.41 - 0.95 / 0.59 - 0.55	servant.*N*	0.64 - 0.7 - 0.78 / 0.71 - 0.73	bean.*N*	0.59 - 0.7 - 0.94 / 0.79 - 0.68
cap.*N*	0.89 - 0.91 - 1.0 / 0.91 - 0.99	teddy.*N*	0.52 - 0.45 - 0.91 / 0.48 - 0.64	scale.*N*	0.53 - 0.59 - 0.95 / 0.59 - 0.82
bra.*N*	0.52 - 0.5 - 0.9 / 0.48 - 0.45	lump.*N*	0.52 - 0.87 - 0.97 / 0.91 - 0.87	shuttle.*N*	0.85 - 0.85 - 0.84 / 0.85 - 0.7
park.*V*	0.53 - 0.54 - 0.86 / 0.54 - 0.58	comfort.*N*	0.78 - 0.78 - 0.78 / 0.78 - 0.77	fuse.*N*	0.58 - 0.72 - 0.96 / 0.83 - 0.78
drunk.*JJ*	0.56 - 0.69 - 0.77 / 0.68 - 0.76	barn.*N*	0.75 - 0.75 - 0.78 / 0.78 - 0.79	radioactive.*JJ*	0.54 - 0.54 - 0.98 / 0.69 - 0.63
vegetable.*N*	0.59 - 0.61 - 0.74 / 0.61 - 0.71	peanut.*N*	0.51 - 0.57 - 0.8 / 0.6 - 0.59	pretend.*V*	0.56 - 0.54 - 0.81 / 0.57 - 0.69
opening.*N*	0.55 - 0.6 - 0.85 / 0.59 - 0.7	cabbage.*N*	0.63 - 0.63 - 0.7 / 0.65 - 0.74	custom.*N*	0.52 - 0.65 - 0.99 / 0.81 - 0.98
rule.*V*	0.68 - 0.7 - 0.87 / 0.72 - 0.88	sandwich.*N*	0.62 - 0.62 - 0.61 / 0.62 - 0.6	log.*V*	0.81 - 0.81 - 0.82 / 0.81 - 0.76
rifle.*N*	0.7 - 0.71 - 0.96 / 0.78 - 0.85	link.*N*	0.5 - 0.78 - 0.96 / 0.83 - 0.94	riot.*N*	0.57 - 0.57 - 0.85 / 0.57 - 0.56
herd.*N*	0.51 - 0.51 - 0.93 / 0.67 - 0.62	supply.*N*	0.59 - 0.59 - 0.85 / 0.61 - 0.65	sweater.*N*	0.56 - 0.54 - 0.56 / 0.56 - 0.47
language.*N*	0.59 - 0.65 - 0.89 / 0.75 - 0.81	privacy.*N*	0.56 - 0.58 - 0.9 / 0.6 - 0.66	intrusion.*N*	0.65 - 0.61 - 0.64 / 0.65 - 0.65
parking.*N*	0.68 - 0.65 - 0.67 / 0.68 - 0.65	alien.*JJ*	0.64 - 0.64 - 0.96 / 0.64 - 0.6	fighter.*N*	0.56 - 0.57 - 0.85 / 0.62 - 0.68
approach.*N*	0.63 - 0.64 - 0.9 / 0.62 - 0.62	pit.*N*	0.63 - 0.63 - 0.64 / 0.63 - 0.63	cover.*N*	0.5 - 0.63 - 0.92 / 0.68 - 0.87
bracelet.*N*	0.54 - 0.53 - 0.88 / 0.55 - 0.65	gossip.*N*	0.55 - 0.55 - 0.81 / 0.57 - 0.65	transfer.*V*	0.63 - 0.64 - 0.63 / 0.63 - 0.64
horn.*N*	0.57 - 0.49 - 0.89 / 0.64 - 0.74	dick.*N*	0.56 - 0.56 - 0.54 / 0.56 - 0.56	reflection.*N*	0.6 - 0.67 - 0.98 / 0.79 - 0.86
razor.*N*	0.69 - 0.71 - 0.89 / 0.76 - 0.68	cleaner.*N*	0.56 - 0.98 - 1.0 / 0.98 - 0.98	speaker.*N*	0.72 - 0.83 - 0.93 / 0.84 - 0.95
computer.*N*	0.65 - 0.65 - 0.65 / 0.65 - 0.61	survivor.*N*	0.55 - 0.56 - 0.81 / 0.59 - 0.47	condolence.*N*	0.57 - 0.59 - 0.56 / 0.54 - 0.54
flock.*N*	0.61 - 0.7 - 0.95 / 0.8 - 0.91	dutch.*JJ*	0.69 - 0.81 - 0.98 / 0.83 - 0.86	ounce.*N*	0.57 - 0.57 - 0.79 / 0.62 - 0.52
cliff.*N*	0.69 - 0.31 - 0.7 / 0.69 - 0.74	bite.*N*	0.57 - 0.66 - 0.85 / 0.64 - 0.77	spread.*V*	0.37 - 0.44 - 0.95 / 0.52 - 0.51
prayer.*N*	0.62 - 0.7 - 0.75 / 0.69 - 0.63	match.*N*	0.52 - 0.52 - 0.52 / 0.52 - 0.56	twenty.*JJ*	0.4 - 0.46 - 0.84 / 0.46 - 0.79
promotion.*N*	0.56 - 0.57 - 0.89 / 0.6 - 0.68	retire.*V*	0.52 - 0.52 - 0.88 / 0.64 - 0.69	tribute.*N*	0.62 - 0.71 - 0.79 / 0.71 - 0.68
vote.*N*	0.53 - 0.76 - 0.87 / 0.83 - 0.92	honesty.*N*	0.66 - 0.66 - 0.66 / 0.66 - 0.59	jar.*N*	0.59 - 0.59 - 0.6 / 0.59 - 0.62
record.*N*	0.34 - 0.41 - 0.92 / 0.59 - 0.79	twenty.*N*	0.74 - 0.76 - 0.77 / 0.74 - 0.93	advance.*N*	0.36 - 0.41 - 0.87 / 0.5 - 0.7
hunch.*N*	0.65 - 0.65 - 0.67 / 0.65 - 0.6	praise.*N*	0.71 - 0.68 - 0.97 / 0.76 - 0.83	jean.*N*	0.58 - 0.58 - 0.8 / 0.62 - 0.54
skull.*N*	0.81 - 0.83 - 0.98 / 0.82 - 0.88	wall.*N*	0.66 - 0.69 - 0.86 / 0.69 - 0.75	restore.*V*	0.67 - 0.63 - 0.93 / 0.7 - 0.73
essentially.*RB*	0.74 - 0.74 - 0.75 / 0.74 - 0.67	mud.*N*	0.61 - 0.61 - 0.6 / 0.61 - 0.5	alien.*N*	0.55 - 0.56 - 0.87 / 0.52 - 0.59
requirement.*N*	0.7 - 0.7 - 0.72 / 0.7 - 0.62	driver.*N*	0.63 - 0.68 - 0.68 / 0.68 - 0.69	chin.*N*	0.64 - 0.64 - 0.93 / 0.56 - 0.59
pneumonia.*N*	0.64 - 0.36 - 0.65 / 0.64 - 0.61	relevant.*JJ*	0.66 - 0.66 - 0.98 / 0.68 - 0.61	wave.*N*	0.66 - 0.73 - 0.89 / 0.78 - 0.89
greed.*N*	0.57 - 0.57 - 0.98 / 0.51 - 0.49	pill.*N*	0.46 - 0.49 - 0.72 / 0.52 - 0.59	unfortunately.*RB*	0.36 - 0.36 - 0.59 / 0.34 - 0.42
dagger.*N*	0.67 - 0.67 - 0.95 / 0.7 - 0.68	riddle.*N*	0.71 - 0.29 - 0.71 / 0.71 - 0.7	encourage.*V*	0.42 - 0.43 - 0.98 / 0.51 - 0.51
editor.*N*	0.54 - 0.63 - 0.9 / 0.69 - 0.86	rude.*JJ*	0.76 - 0.76 - 0.75 / 0.76 - 0.7	belly.*N*	0.41 - 0.42 - 0.89 / 0.52 - 0.5
maid.*N*	0.36 - 0.54 - 0.67 / 0.56 - 0.56	calf.*N*	0.64 - 0.66 - 0.96 / 0.67 - 0.6		
temper.*N*	0.29 - 0.5 - 0.55 / 0.48 - 0.49	ticket.*N*	0.58 - 0.64 - 0.77 / 0.67 - 0.66		
		Overall Average:	59.43 - 62.40 - 66.87 - 70.72		

Table 5: Lexical selection model test accuracies for all 157 English-Spanish words.

Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT	Focus word	FreqBaseline - DTree - LinearSVM Train/Test- BERT	Focus word	FreqBaseline - DTree - LinearSVM - BERT
free.*JJ*	0.7 - 0.73 - 0.73 / 0.73 - 0.68	peaceful.*JJ*	0.68 - 0.7 - 0.82 / 0.72 - 0.76	fighter.*N*	0.29 - 0.31 - 0.78 / 0.31 - 0.47
roof.*N*	0.37 - 0.39 - 0.63 / 0.39 - 0.46	sharp.*JJ*	0.53 - 0.62 - 0.71 / 0.63 - 0.65	crew.*N*	0.83 - 0.88 - 0.91 / 0.85 - 0.94
sword.*N*	0.76 - 0.77 - 0.76 / 0.76 - 0.7	tie.*V*	0.34 - 0.5 - 0.82 / 0.63 - 0.89	convinced.*JJ*	0.66 - 0.68 - 0.73 / 0.66 - 0.65
storm.*N*	0.85 - 0.84 - 0.87 / 0.84 - 0.83	puzzle.*N*	0.61 - 0.69 - 0.81 / 0.71 - 0.7	point.*V*	0.38 - 0.42 - 0.55 / 0.43 - 0.61
break.*V*	0.16 - 0.42 - 0.73 / 0.55 - 0.75	fan.*N*	0.61 - 0.65 - 0.87 / 0.69 - 0.8	broke.*JJ*	0.37 - 0.4 - 0.72 / 0.35 - 0.4
bitch.*N*	0.32 - 0.45 - 0.46 / 0.45 - 0.42	shake.*V*	0.32 - 0.66 - 0.83 / 0.68 - 0.85	sail.*V*	0.55 - 0.55 - 0.8 / 0.52 - 0.62
set.*V*	0.3 - 0.48 - 0.59 / 0.51 - 0.71	capital.*N*	0.71 - 0.78 - 0.9 / 0.77 - 0.98	civil.*JJ*	0.53 - 0.84 - 0.96 / 0.86 - 0.91
wheel.*N*	0.38 - 0.59 - 0.84 / 0.68 - 0.74	sixth.*JJ*	0.78 - 0.78 - 0.78 / 0.78 - 0.71	beef.*N*	0.37 - 0.37 - 0.77 / 0.39 - 0.43
bone.*N*	0.33 - 0.46 - 0.85 / 0.62 - 0.82	illusion.*N*	0.67 - 0.68 - 0.78 / 0.68 - 0.68	deadline.*N*	0.8 - 0.8 - 0.8 / 0.8 - 0.76
tunnel.*N*	0.67 - 0.67 - 0.68 / 0.67 - 0.61	wet.*JJ*	0.41 - 0.68 - 0.77 / 0.68 - 0.71	makeup.*N*	0.61 - 0.6 - 0.7 / 0.62 - 0.49
cool.*JJ*	0.48 - 0.49 - 0.51 / 0.48 - 0.5	costume.*N*	0.65 - 0.65 - 0.65 / 0.65 - 0.57	text.*N*	0.84 - 0.84 - 0.92 / 0.84 - 0.86
bedroom.*N*	0.63 - 0.63 - 0.67 / 0.64 - 0.6	plague.*N*	0.69 - 0.68 - 0.84 / 0.69 - 0.67	feed.*V*	0.32 - 0.46 - 0.88 / 0.48 - 0.76
mountain.*N*	0.94 - 0.95 - 0.94 / 0.94 - 0.95	vault.*N*	0.48 - 0.48 - 0.81 / 0.48 - 0.53	bubble.*N*	0.54 - 0.57 - 0.86 / 0.55 - 0.56
cake.*N*	0.93 - 0.99 - 0.98 / 0.99 - 0.99	deadly.*JJ*	0.72 - 0.78 - 0.75 / 0.73 - 0.73	drum.*N*	0.52 - 0.68 - 0.83 / 0.75 - 0.76
coat.*N*	0.87 - 0.88 - 0.88 / 0.88 - 0.87	collect.*V*	0.53 - 0.61 - 0.86 / 0.65 - 0.83	drill.*N*	0.52 - 0.59 - 0.91 / 0.74 - 0.83
bunch.*N*	0.78 - 0.79 - 0.84 / 0.79 - 0.73	cliff.*N*	0.58 - 0.59 - 0.85 / 0.59 - 0.63	musical.*JJ*	0.55 - 0.7 - 0.93 / 0.76 - 0.91
turn.*V*	0.12 - 0.19 - 0.66 / 0.34 - 0.83	scale.*N*	0.73 - 0.75 - 0.9 / 0.77 - 0.82	burn.*N*	0.68 - 0.73 - 0.88 / 0.74 - 0.83
effect.*N*	0.38 - 0.8 - 0.77 / 0.8 - 0.82	horn.*N*	0.62 - 0.71 - 0.82 / 0.73 - 0.79	lamb.*N*	0.77 - 0.76 - 0.86 / 0.77 - 0.77
farm.*N*	0.95 - 0.95 - 0.95 / 0.95 - 0.94	stick.*N*	0.46 - 0.52 - 0.52 / 0.51 - 0.47	frame.*N*	0.37 - 0.35 - 0.97 / 0.57 - 0.57
tie.*N*	0.67 - 0.76 - 0.89 / 0.81 - 0.93	porn.*N*	0.79 - 0.79 - 0.79 / 0.79 - 0.77	column.*N*	0.77 - 0.78 - 0.9 / 0.8 - 0.88
tour.*N*	0.48 - 0.57 - 0.81 / 0.65 - 0.66	range.*N*	0.58 - 0.66 - 0.86 / 0.67 - 0.69	brilliant.*JJ*	0.43 - 0.43 - 0.56 / 0.44 - 0.51
band.*N*	0.79 - 0.81 - 0.85 / 0.81 - 0.79	host.*N*	0.56 - 0.64 - 0.88 / 0.74 - 0.87	explode.*V*	0.51 - 0.59 - 0.85 / 0.58 - 0.9
self.*N*	0.33 - 0.38 - 0.87 / 0.46 - 0.85	grow.*V*	0.46 - 0.62 - 0.92 / 0.69 - 0.92	fort.*N*	0.61 - 0.62 - 0.7 / 0.62 - 0.55
bill.*N*	0.57 - 0.64 - 0.91 / 0.74 - 0.88	expose.*V*	0.62 - 0.62 - 0.62 / 0.62 - 0.68	impact.*N*	0.47 - 0.56 - 0.92 / 0.63 - 0.8
cookie.*N*	0.79 - 0.77 - 0.8 / 0.79 - 0.73	upset.*JJ*	0.47 - 0.53 - 0.54 / 0.53 - 0.46	scarf.*N*	0.45 - 0.45 - 0.63 / 0.45 - 0.45
dozen.*N*	0.58 - 0.66 - 0.83 / 0.71 - 0.81	lightning.*N*	0.51 - 0.66 - 0.8 / 0.68 - 0.65	fail.*V*	0.71 - 0.71 - 0.74 / 0.7 - 0.99
fruit.*N*	0.79 - 0.86 - 0.95 / 0.88 - 0.93	laptop.*N*	0.79 - 0.79 - 0.82 / 0.79 - 0.77	skill.*N*	0.54 - 0.57 - 0.82 / 0.6 - 0.71
pen.*N*	0.78 - 0.8 - 0.79 / 0.79 - 0.79	response.*N*	0.57 - 0.59 - 0.61 / 0.58 - 0.64	mail.*N*	0.9 - 0.95 - 0.97 / 0.96 - 0.95
trigger.*N*	0.87 - 0.88 - 0.97 / 0.88 - 0.93	obvious.*JJ*	0.52 - 0.54 - 0.6 / 0.55 - 0.58	issue.*V*	0.5 - 0.54 - 0.86 / 0.58 - 0.79
ring.*N*	0.42 - 0.63 - 0.79 / 0.72 - 0.81	distant.*JJ*	0.67 - 0.77 - 0.96 / 0.89 - 0.92	involve.*V*	0.28 - 0.31 - 0.69 / 0.32 - 0.38
drag.*V*	0.27 - 0.33 - 0.65 / 0.43 - 0.64	niece.*N*	0.76 - 0.76 - 0.76 / 0.76 - 0.72	label.*N*	0.55 - 0.66 - 0.8 / 0.68 - 0.79
old.*JJ*	0.4 - 0.66 - 0.86 / 0.73 - 0.91	string.*N*	0.38 - 0.66 - 0.85 / 0.7 - 0.77	psychic.*JJ*	0.61 - 0.74 - 0.93 / 0.79 - 0.84
bug.*N*	0.35 - 0.42 - 0.54 / 0.46 - 0.52	promote.*V*	0.59 - 0.62 - 0.92 / 0.75 - 0.83	stamp.*N*	0.61 - 0.7 - 0.91 / 0.72 - 0.83
campaign.*N*	0.53 - 0.53 - 0.83 / 0.53 - 0.58	straight.*JJ*	0.43 - 0.53 - 0.86 / 0.61 - 0.76	dump.*N*	0.42 - 0.62 - 0.73 / 0.64 - 0.7
match.*N*	0.47 - 0.54 - 0.79 / 0.58 - 0.79	worthy.*JJ*	0.75 - 0.75 - 0.75 / 0.75 - 0.81	shell.*N*	0.36 - 0.53 - 0.9 / 0.56 - 0.7
beat.*V*	0.22 - 0.29 - 0.53 / 0.34 - 0.54	rocket.*N*	0.74 - 0.74 - 0.76 / 0.74 - 0.74	disease.*N*	0.61 - 0.63 - 0.89 / 0.65 - 0.75
bottom.*N*	0.69 - 0.78 - 0.85 / 0.78 - 0.8	tub.*N*	0.7 - 0.89 - 0.88 / 0.89 - 0.87	rule.*V*	0.33 - 0.4 - 0.9 / 0.59 - 0.76
solve.*V*	0.41 - 0.51 - 0.67 / 0.53 - 0.66	heir.*N*	0.66 - 0.68 - 0.87 / 0.68 - 0.67	appeal.*N*	0.76 - 0.78 - 0.85 / 0.8 - 0.84
sell.*V*	0.44 - 0.49 - 0.75 / 0.57 - 0.75	circumstance.*N*	0.94 - 0.94 - 0.94 / 0.94 - 0.94	can.*N*	0.55 - 0.68 - 0.87 / 0.7 - 0.71
butter.*N*	0.59 - 0.82 - 0.86 / 0.82 - 0.93	trunk.*N*	0.4 - 0.52 - 0.7 / 0.57 - 0.66	glorious.*JJ*	0.74 - 0.74 - 0.75 / 0.74 - 0.7
culture.*N*	0.48 - 0.53 - 0.89 / 0.58 - 0.64	engagement.*N*	0.64 - 0.84 - 0.91 / 0.84 - 0.85	focused.*JJ*	0.75 - 0.78 - 0.75 / 0.75 - 0.73
season.*N*	0.62 - 0.67 - 0.81 / 0.68 - 0.66	remain.*N*	0.37 - 0.38 - 0.91 / 0.35 - 0.38	delicious.*JJ*	0.54 - 0.54 - 0.63 / 0.54 - 0.58
tank.*N*	0.5 - 0.66 - 0.8 / 0.68 - 0.63	crash.*N*	0.48 - 0.61 - 0.72 / 0.6 - 0.64	addict.*N*	0.52 - 0.52 - 0.52 / 0.52 - 0.54
wash.*V*	0.5 - 0.59 - 0.77 / 0.62 - 0.85	cellar.*N*	0.84 - 0.84 - 0.84 / 0.84 - 0.84	player.*N*	0.39 - 0.97 - 0.99 / 0.98 - 0.98
ball.*N*	0.51 - 0.54 - 0.58 / 0.54 - 0.63	opening.*N*	0.37 - 0.62 - 0.82 / 0.64 - 0.76	lethal.*JJ*	0.75 - 0.79 - 0.85 / 0.79 - 0.78
painful.*JJ*	0.48 - 0.49 - 0.5 / 0.49 - 0.46	seal.*N*	0.6 - 0.66 - 0.95 / 0.77 - 0.9	server.*N*	0.67 - 0.67 - 0.83 / 0.68 - 0.62
general.*JJ*	0.68 - 0.69 - 0.95 / 0.78 - 0.99	leak.*V*	0.56 - 0.58 - 0.85 / 0.6 - 0.76	welcome.*JJ*	0.46 - 0.48 - 0.81 / 0.6 - 0.77
evil.*JJ*	0.5 - 0.5 - 0.66 / 0.52 - 0.48	harmless.*JJ*	0.51 - 0.56 - 0.56 / 0.55 - 0.49	penny.*N*	0.93 - 0.92 - 0.93 / 0.92 - 0.98
degree.*N*	0.5 - 0.62 - 0.93 / 0.86 - 0.93	demand.*N*	0.66 - 0.76 - 0.95 / 0.8 - 0.94	immune.*JJ*	0.57 - 0.93 - 0.94 / 0.93 - 0.97
captain.*N*	0.52 - 0.54 - 0.52 / 0.52 - 0.67	hip.*N*	0.65 - 0.66 - 0.86 / 0.68 - 0.85	vet.*N*	0.73 - 0.8 - 0.94 / 0.82 - 0.91
serial.*JJ*	0.63 - 0.84 - 0.85 / 0.85 - 0.82	pride.*N*	0.61 - 0.64 - 0.8 / 0.65 - 0.85	define.*V*	0.46 - 0.49 - 0.87 / 0.51 - 0.59
infection.*N*	0.7 - 0.7 - 0.7 / 0.7 - 0.66	file.*V*	0.4 - 0.43 - 0.57 / 0.46 - 0.58	desperate.*JJ*	0.73 - 0.72 - 0.74 / 0.73 - 0.78
fight.*N*	0.48 - 0.52 - 0.76 / 0.51 - 0.62	burn.*V*	0.5 - 0.58 - 0.88 / 0.6 - 0.78	move.*V*	0.31 - 0.47 - 0.9 / 0.62 - 0.89
gut.*N*	0.56 - 0.74 - 0.87 / 0.79 - 0.84	charm.*N*	0.75 - 0.85 - 0.88 / 0.85 - 0.88	acre.*N*	0.72 - 0.72 - 0.72 / 0.72 - 0.72
spring.*N*	0.92 - 0.92 - 0.92 / 0.92 - 0.97	partner.*N*	0.53 - 0.63 - 0.93 / 0.72 - 0.88	star.*N*	0.28 - 0.67 - 0.91 / 0.77 - 0.87
spread.*V*	0.38 - 0.41 - 0.66 / 0.44 - 0.65	youth.*N*	0.39 - 0.45 - 0.85 / 0.47 - 0.6	claim.*N*	0.47 - 0.49 - 0.89 / 0.58 - 0.69
hot.*JJ*	0.5 - 0.71 - 0.88 / 0.75 - 0.83	raise.*V*	0.25 - 0.41 - 0.87 / 0.55 - 0.77	leadership.*N*	0.77 - 0.81 - 0.87 / 0.8 - 0.81
cup.*N*	0.7 - 0.68 - 0.72 / 0.7 - 0.67	depressed.*JJ*	0.83 - 0.83 - 0.82 / 0.83 - 0.81	collar.*N*	0.69 - 0.69 - 0.85 / 0.7 - 0.68
clown.*N*	0.95 - 0.95 - 0.95 / 0.95 - 0.95	toast.*N*	0.73 - 0.73 - 0.74 / 0.73 - 0.6	donate.*V*	0.55 - 0.55 - 0.87 / 0.55 - 0.68
lieutenant.*N*	0.41 - 0.48 - 0.71 / 0.52 - 0.63	burger.*N*	0.52 - 0.5 - 0.52 / 0.52 - 0.44	suspend.*V*	0.53 - 0.52 - 0.78 / 0.55 - 0.7
original.*JJ*	0.46 - 0.59 - 0.82 / 0.65 - 0.77	bury.*V*	0.74 - 0.76 - 0.76 / 0.75 - 0.88	shade.*N*	0.58 - 0.76 - 0.94 / 0.88 - 0.92
grass.*N*	0.51 - 0.52 - 0.83 / 0.58 - 0.64	fatal.*JJ*	0.52 - 0.57 - 0.89 / 0.55 - 0.52	sketch.*N*	0.79 - 0.81 - 0.8 / 0.79 - 0.93
radiation.*N*	0.63 - 0.66 - 0.88 / 0.66 - 0.64	abuse.*N*	0.61 - 0.72 - 0.79 / 0.72 - 0.77	hood.*N*	0.55 - 0.7 - 0.94 / 0.82 - 0.88
sad.*JJ*	0.5 - 0.64 - 0.85 / 0.73 - 0.81	soft.*JJ*	0.75 - 0.82 - 0.88 / 0.82 - 0.87	build.*V*	0.36 - 0.4 - 0.94 / 0.45 - 0.76
necklace.*N*	0.79 - 0.79 - 0.79 / 0.79 - 0.76	invade.*V*	0.74 - 0.74 - 0.74 / 0.74 - 0.86	tear.*V*	0.39 - 0.49 - 0.86 / 0.51 - 0.75
grade.*N*	0.42 - 0.68 - 0.88 / 0.89 - 0.9	crack.*N*	0.46 - 0.59 - 0.91 / 0.64 - 0.83	determine.*V*	0.62 - 0.63 - 0.93 / 0.7 - 0.94
beast.*N*	0.66 - 0.67 - 0.8 / 0.67 - 0.66	maid.*N*	0.59 - 0.77 - 0.8 / 0.78 - 0.74	mourn.*V*	0.56 - 0.55 - 0.84 / 0.6 - 0.62
blade.*N*	0.77 - 0.84 - 0.83 / 0.84 - 0.83	spare.*V*	0.32 - 0.46 - 0.81 / 0.52 - 0.63	abandon.*V*	0.48 - 0.55 - 0.86 / 0.59 - 0.75
rise.*V*	0.23 - 0.43 - 0.88 / 0.6 - 0.82	inappropriate.*JJ*	0.65 - 0.65 - 0.81 / 0.66 - 0.6	lottery.*N*	0.51 - 0.63 - 0.78 / 0.61 - 0.59
autopsy.*JJ*	0.54 - 0.53 - 0.79 / 0.51 - 0.48	trouble.*N*	0.32 - 0.49 - 0.81 / 0.59 - 0.86	pole.*N*	0.44 - 0.58 - 0.91 / 0.67 - 0.65
jacket.*N*	0.65 - 0.65 - 0.66 / 0.66 - 0.55	daily.*JJ*	0.7 - 0.83 - 0.96 / 0.83 - 0.83	exhausted.*JJ*	0.62 - 0.6 - 0.61 / 0.62 - 0.49
oil.*N*	0.83 - 0.9 - 0.95 / 0.89 - 0.88	recording.*N*	0.6 - 0.56 - 0.82 / 0.59 - 0.59	rubber.*N*	0.42 - 0.42 - 0.88 / 0.44 - 0.51
compliment.*N*	0.57 - 0.26 - 0.58 / 0.57 - 0.51	sink.*N*	0.7 - 0.7 - 0.7 / 0.7 - 0.68	nipple.*N*	0.76 - 0.76 - 0.76 / 0.76 - 0.73
pulse.*N*	0.7 - 0.76 - 0.78 / 0.77 - 0.82	custom.*N*	0.58 - 0.66 - 0.89 / 0.71 - 0.97	sew.*V*	0.48 - 0.52 - 0.87 / 0.56 - 0.72
nephew.*N*	0.76 - 0.76 - 0.77 / 0.76 - 0.72	brandy.*N*	0.76 - 0.76 - 0.76 / 0.76 - 0.74	mental.*JJ*	0.44 - 0.38 - 0.84 / 0.45 - 0.47
step.*N*	0.35 - 0.61 - 0.74 / 0.68 - 0.73	souvenir.*N*	0.48 - 0.49 - 0.57 / 0.5 - 0.46	janitor.*N*	0.8 - 0.8 - 0.8 / 0.8 - 0.77
suffer.*V*	0.72 - 0.81 - 0.89 / 0.8 - 0.92	pepper.*N*	0.67 - 0.84 - 0.9 / 0.84 - 0.97	efficient.*JJ*	0.7 - 0.7 - 0.7 / 0.7 - 0.66
run.*V*	0.19 - 0.32 - 0.84 / 0.59 - 0.84	distraction.*N*	0.53 - 0.64 - 0.75 / 0.62 - 0.73	speaker.*N*	0.45 - 0.51 - 0.84 / 0.58 - 0.8
fight.*V*	0.21 - 0.29 - 0.61 / 0.34 - 0.65	remarkable.*JJ*	0.44 - 0.5 - 0.78 / 0.52 - 0.52	lawn.*N*	0.57 - 0.57 - 0.82 / 0.53 - 0.54
store.*N*	0.24 - 0.66 - 0.81 / 0.69 - 0.7	mob.*N*	0.58 - 0.62 - 0.91 / 0.69 - 0.86	therapist.*N*	0.69 - 0.69 - 0.69 / 0.69 - 0.65
fire.*V*	0.76 - 0.81 - 0.91 / 0.83 - 0.95	casualty.*N*	0.78 - 0.78 - 0.78 / 0.78 - 0.7	administration.*N*	0.52 - 0.52 - 0.9 / 0.56 - 0.67
bright.*JJ*	0.71 - 0.86 - 0.89 / 0.86 - 0.87	increase.*V*	0.63 - 0.66 - 0.96 / 0.68 - 0.9	reckless.*JJ*	0.63 - 0.6 - 0.77 / 0.62 - 0.56
inch.*N*	0.5 - 0.56 - 0.69 / 0.54 - 0.5	melt.*V*	0.63 - 0.64 - 0.85 / 0.65 - 0.83	ham.*N*	0.73 - 0.12 - 0.73 / 0.73 - 0.69
barn.*N*	0.79 - 0.79 - 0.79 / 0.79 - 0.75	thick.*JJ*	0.47 - 0.55 - 0.88 / 0.57 - 0.63	settle.*V*	0.39 - 0.51 - 0.83 / 0.64 - 0.91
gas.*N*	0.45 - 0.81 - 0.89 / 0.85 - 0.9	mole.*N*	0.34 - 0.36 - 0.84 / 0.42 - 0.54	headline.*N*	0.52 - 0.52 - 0.86 / 0.55 - 0.49
cover.*N*	0.65 - 0.66 - 0.89 / 0.71 - 0.91	football.*N*	0.39 - 0.41 - 0.75 / 0.49 - 0.52	estate.*N*	0.43 - 0.71 - 0.82 / 0.71 - 0.76
pot.*N*	0.38 - 0.49 - 0.68 / 0.6 - 0.56	cattle.*N*	0.27 - 0.27 - 0.35 / 0.28 - 0.34	smooth.*JJ*	0.38 - 0.45 - 0.88 / 0.45 - 0.6
alley.*N*	0.44 - 0.44 - 0.81 / 0.49 - 0.51	special.*JJ*	0.85 - 0.89 - 0.93 / 0.89 - 0.88	worker.*N*	0.75 - 0.99 - 0.97 / 0.99 - 0.99
liver.*N*	0.65 - 0.73 - 0.88 / 0.75 - 0.74	high.*JJ*	0.6 - 0.66 - 0.69 / 0.66 - 0.63	gallon.*N*	0.69 - 0.68 - 0.72 / 0.69 - 0.71
escape.*V*	0.43 - 0.46 - 0.86 / 0.5 - 0.72	clear.*JJ*	0.31 - 0.38 - 0.86 / 0.49 - 0.81	scan.*V*	0.48 - 0.52 - 0.59 / 0.52 - 0.58
beard.*N*	0.6 - 0.6 - 0.6 / 0.6 - 0.6	paperwork.*N*	0.55 - 0.55 - 0.64 / 0.55 - 0.56	lure.*V*	0.43 - 0.42 - 0.82 / 0.42 - 0.69
moon.*N*	0.6 - 0.95 - 0.97 / 0.96 - 0.96	dry.*V*	0.74 - 0.8 - 0.86 / 0.8 - 0.89	sophisticated.*JJ*	0.35 - 0.35 - 0.66 / 0.37 - 0.45
crown.*N*	0.81 - 0.85 - 0.96 / 0.86 - 0.89	benefit.*N*	0.39 - 0.45 - 0.87 / 0.53 - 0.62	offensive.*JJ*	0.72 - 0.82 - 0.92 / 0.81 - 0.89
arrest.*V*	0.49 - 0.54 - 0.59 / 0.54 - 0.82	scratch.*N*	0.56 - 0.59 - 0.57 / 0.56 - 0.56	contribute.*V*	0.59 - 0.63 - 0.84 / 0.63 - 0.72
male.*JJ*	0.6 - 0.61 - 0.86 / 0.63 - 0.69	peanut.*N*	0.45 - 0.82 - 0.82 / 0.82 - 0.78	management.*N*	0.55 - 0.6 - 0.89 / 0.61 - 0.7
gum.*N*	0.5 - 0.5 - 0.77 / 0.52 - 0.69	immunity.*N*	0.8 - 0.85 - 0.91 / 0.85 - 0.88	straw.*N*	0.51 - 0.62 - 0.91 / 0.66 - 0.86
approve.*V*	0.43 - 0.48 - 0.85 / 0.54 - 0.83	cancel.*V*	0.83 - 0.83 - 0.83 / 0.83 - 0.85	donkey.*N*	0.61 - 0.6 - 0.7 / 0.62 - 0.63
candy.*N*	0.51 - 0.61 - 0.82 / 0.62 - 0.65	wing.*N*	0.85 - 0.96 - 0.99 / 0.96 - 0.99	delicate.*JJ*	0.7 - 0.7 - 0.7 / 0.7 - 0.68
fifth.*JJ*	0.61 - 0.66 - 0.73 / 0.67 - 0.72	camp.*N*	0.81 - 0.85 - 0.87 / 0.85 - 0.89	brutal.*JJ*	0.39 - 0.42 - 0.43 / 0.4 - 0.28
egg.*N*	0.63 - 0.78 - 0.86 / 0.81 - 0.89	chip.*N*	0.87 - 1.0 - 0.99 / 1.0 - 1.0	sunny.*JJ*	0.6 - 0.6 - 0.94 / 0.71 - 0.81

Table 6: Part-1: Lexical selection model test accuracies for a English-Greek words.

Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT	Focus word	FreqBaseline - DTree - LinearSVM Train/Test- BERT	Focus word	FreqBaseline - DTree - LinearSVM - BERT
light.*JJ*	0.84 - 0.89 - 0.99 / 0.92 - 0.99	camping.*N*	0.69 - 0.69 - 0.68 / 0.69 - 0.63	suspension.*N*	0.42 - 0.48 - 0.96 / 0.58 - 0.73
current.*JJ*	0.9 - 0.92 - 0.9 / 0.9 - 0.91	cheat.*V*	0.34 - 0.33 - 0.64 / 0.3 - 0.67	notorious.*JJ*	0.7 - 0.72 - 0.7 / 0.7 - 0.67
humiliating.*JJ*	0.51 - 0.47 - 0.78 / 0.51 - 0.47	act.*V*	0.51 - 0.53 - 0.87 / 0.51 - 0.65	relative.*JJ*	0.59 - 0.64 - 0.98 / 0.76 - 0.95
drone.*N*	0.38 - 0.39 - 0.89 / 0.41 - 0.52	rose.*N*	0.47 - 0.51 - 0.9 / 0.57 - 0.76	honor.*V*	0.82 - 0.88 - 0.99 / 0.93 - 0.93
bald.*JJ*	0.69 - 0.69 - 0.69 / 0.69 - 0.64	preacher.*N*	0.64 - 0.64 - 0.65 / 0.64 - 0.67	duct.*N*	0.46 - 0.65 - 0.88 / 0.68 - 0.7
wipe.*V*	0.38 - 0.4 - 0.85 / 0.46 - 0.55	doorman.*N*	0.72 - 0.69 - 0.72 / 0.72 - 0.67	scooter.*N*	0.78 - 0.78 - 0.79 / 0.78 - 0.72
institution.*N*	0.54 - 0.67 - 0.89 / 0.68 - 0.76	unite.*V*	0.53 - 0.53 - 0.91 / 0.52 - 0.7	temporal.*JJ*	0.51 - 0.93 - 0.94 / 0.89 - 0.89
retirement.*N*	0.54 - 0.58 - 0.87 / 0.63 - 0.57	flock.*N*	0.51 - 0.54 - 0.88 / 0.56 - 0.59	sis.*N*	0.55 - 0.55 - 0.85 / 0.58 - 0.42
bunny.*N*	0.5 - 0.47 - 0.83 / 0.5 - 0.53	remark.*N*	0.59 - 0.58 - 0.6 / 0.59 - 0.51	terrace.*N*	0.63 - 0.63 - 0.63 / 0.63 - 0.7
despair.*N*	0.53 - 0.53 - 0.95 / 0.55 - 0.56	expel.*V*	0.5 - 0.62 - 0.68 / 0.62 - 0.74	tenth.*JJ*	0.77 - 0.77 - 0.76 / 0.77 - 0.73
cult.*N*	0.75 - 0.75 - 0.75 / 0.75 - 0.71	dizzy.*JJ*	0.45 - 0.51 - 0.66 / 0.53 - 0.56	banner.*N*	0.5 - 0.52 - 0.91 / 0.55 - 0.67
trainer.*N*	0.44 - 0.54 - 0.83 / 0.51 - 0.57	vulture.*N*	0.53 - 0.19 - 0.83 / 0.62 - 0.62	mixture.*N*	0.53 - 0.54 - 0.95 / 0.5 - 0.41
genius.*N*	0.66 - 0.66 - 0.81 / 0.66 - 0.67	simulation.*N*	0.78 - 0.78 - 0.79 / 0.78 - 0.79	swelling.*N*	0.73 - 0.71 - 0.79 / 0.7 - 0.73
competition.*N*	0.87 - 0.86 - 0.87 / 0.87 - 0.87	contempt.*N*	0.74 - 0.85 - 0.88 / 0.86 - 0.81	clip.*N*	0.37 - 0.68 - 0.93 / 0.83 - 0.83
thread.*N*	0.63 - 0.65 - 0.81 / 0.64 - 0.67	dean.*N*	0.8 - 0.81 - 0.8 / 0.8 - 0.76	mix.*V*	0.57 - 0.6 - 0.86 / 0.64 - 0.84
coach.*N*	0.83 - 0.82 - 0.84 / 0.83 - 0.94	beg.*V*	0.79 - 0.82 - 0.94 / 0.8 - 0.91	eliminate.*V*	0.46 - 0.51 - 0.52 / 0.46 - 0.63
trance.*N*	0.82 - 0.84 - 0.82 / 0.82 - 0.99	grieve.*V*	0.67 - 0.67 - 0.69 / 0.67 - 0.63	bouquet.*N*	0.63 - 0.63 - 0.73 / 0.64 - 0.57
respectable.*JJ*	0.52 - 0.51 - 0.53 / 0.52 - 0.34	pea.*N*	0.59 - 0.56 - 0.78 / 0.59 - 0.62	dig.*V*	0.53 - 0.63 - 0.65 / 0.63 - 0.85
notebook.*N*	0.78 - 0.78 - 0.78 / 0.78 - 0.78	extension.*N*	0.38 - 0.51 - 0.94 / 0.56 - 0.68	abs.*N*	0.55 - 0.58 - 0.93 / 0.64 - 0.97
recommendation.*N*	0.54 - 0.73 - 0.83 / 0.74 - 0.7	intercept.*V*	0.55 - 0.65 - 0.87 / 0.72 - 0.87	desperation.*N*	0.59 - 0.59 - 0.9 / 0.52 - 0.48
wire.*N*	0.61 - 0.71 - 0.93 / 0.73 - 0.77	homemade.*JJ*	0.77 - 0.79 - 0.77 / 0.77 - 0.85	slogan.*N*	0.77 - 0.77 - 0.77 / 0.77 - 0.79
humiliation.*N*	0.6 - 0.62 - 0.69 / 0.63 - 0.56	domestic.*JJ*	0.48 - 0.7 - 0.93 / 0.62 - 0.75	raven.*N*	0.65 - 0.68 - 0.88 / 0.65 - 0.89
dock.*N*	0.48 - 0.61 - 0.8 / 0.6 - 0.74	spear.*N*	0.43 - 0.42 - 0.8 / 0.38 - 0.58	waste.*V*	0.58 - 0.86 - 0.96 / 0.83 - 0.94
recover.*V*	0.38 - 0.48 - 0.87 / 0.52 - 0.79	printer.*N*	0.78 - 0.77 - 0.78 / 0.78 - 0.86	honorable.*JJ*	0.75 - 0.79 - 0.97 / 0.77 - 0.75
fortress.*N*	0.71 - 0.69 - 0.71 / 0.71 - 0.61	carnival.*N*	0.72 - 0.75 - 0.72 / 0.72 - 0.64	clarity.*N*	0.69 - 0.69 - 0.7 / 0.69 - 0.69
furious.*JJ*	0.67 - 0.67 - 0.68 / 0.67 - 0.62	intervene.*V*	0.73 - 0.7 - 0.73 / 0.73 - 0.68	minority.*N*	0.66 - 0.66 - 0.91 / 0.67 - 0.72
light.*V*	0.62 - 0.73 - 0.94 / 0.74 - 0.83	concrete.*N*	0.63 - 0.65 - 0.82 / 0.67 - 0.7	frustrating.*JJ*	0.73 - 0.73 - 0.76 / 0.73 - 0.61
sign.*V*	0.82 - 0.82 - 0.83 / 0.82 - 0.98	argument.*N*	0.58 - 0.58 - 0.57 / 0.58 - 0.55	resident.*N*	0.7 - 0.73 - 0.69 / 0.7 - 0.85
crop.*N*	0.63 - 0.69 - 0.86 / 0.69 - 0.63	extensive.*JJ*	0.72 - 0.72 - 0.72 / 0.72 - 0.71	spaghetti.*N*	0.56 - 0.56 - 0.55 / 0.56 - 0.48
spill.*V*	0.56 - 0.65 - 0.85 / 0.7 - 0.81	kiss.*V*	0.36 - 0.68 - 0.92 / 0.77 - 0.89	relic.*N*	0.6 - 0.6 - 0.78 / 0.58 - 0.75
congressman.*N*	0.59 - 0.62 - 0.74 / 0.61 - 0.52	harvest.*N*	0.61 - 0.64 - 0.74 / 0.64 - 0.71	shaft.*N*	0.5 - 0.65 - 0.83 / 0.62 - 0.75
sale.*N*	0.53 - 0.71 - 0.92 / 0.76 - 0.99	foreman.*N*	0.43 - 0.51 - 0.81 / 0.58 - 0.77	breathe.*RB*	0.67 - 0.67 - 0.6 / 0.67 - 0.61
advanced.*JJ*	0.67 - 0.73 - 0.84 / 0.73 - 0.78	pier.*N*	0.82 - 0.82 - 0.82 / 0.82 - 0.78	contraction.*N*	0.69 - 0.69 - 0.69 / 0.69 - 0.67
publish.*V*	0.52 - 0.53 - 0.62 / 0.55 - 0.74	ignorant.*JJ*	0.6 - 0.56 - 0.6 / 0.6 - 0.5	outbreak.*N*	0.6 - 0.6 - 0.6 / 0.6 - 0.53
popcorn.*N*	0.77 - 0.23 - 0.77 / 0.77 - 0.71	fart.*V*	0.68 - 0.73 - 0.68 / 0.68 - 0.9	record.*V*	0.83 - 0.85 - 0.85 / 0.83 - 0.97
thunder.*N*	0.62 - 0.61 - 0.68 / 0.62 - 0.68	sabotage.*N*	0.78 - 0.78 - 0.78 / 0.78 - 0.68	ranger.*N*	0.55 - 0.7 - 0.86 / 0.73 - 0.78
wreck.*N*	0.35 - 0.39 - 0.51 / 0.38 - 0.58	stripe.*N*	0.54 - 0.63 - 0.89 / 0.65 - 0.59	cathedral.*N*	0.55 - 0.52 - 0.86 / 0.57 - 0.57
shine.*V*	0.33 - 0.65 - 0.88 / 0.71 - 0.84	rooftop.*N*	0.66 - 0.66 - 0.66 / 0.66 - 0.66	boredom.*N*	0.48 - 0.48 - 0.52 / 0.48 - 0.5
bite.*N*	0.76 - 0.87 - 0.97 / 0.83 - 0.92	destroyer.*N*	0.59 - 0.64 - 0.95 / 0.79 - 0.86	manual.*JJ*	0.73 - 1.0 - 1.0 / 1.0 - 1.0
contaminate.*V*	0.44 - 0.43 - 0.57 / 0.43 - 0.8	whine.*V*	0.41 - 0.41 - 0.53 / 0.39 - 0.32	interfere.*V*	0.76 - 0.78 - 0.75 / 0.76 - 0.85
intern.*N*	0.56 - 0.56 - 0.57 / 0.56 - 0.55	muffin.*N*	0.59 - 0.59 - 0.76 / 0.6 - 0.55	quality.*N*	0.57 - 0.55 - 0.84 / 0.62 - 0.58
willing.*JJ*	0.69 - 0.83 - 0.82 / 0.83 - 0.77	delusion.*N*	0.59 - 0.59 - 0.59 / 0.59 - 0.55	obstruction.*N*	0.65 - 0.56 - 0.66 / 0.65 - 0.56
fountain.*N*	0.63 - 0.63 - 0.69 / 0.63 - 0.59	tornado.*N*	0.76 - 0.76 - 0.76 / 0.76 - 0.77	modification.*N*	0.65 - 0.67 - 0.69 / 0.65 - 0.72
compete.*V*	0.56 - 0.58 - 0.86 / 0.62 - 0.71	rock.*N*	0.81 - 0.96 - 1.0 / 0.96 - 1.0	marrow.*N*	0.71 - 0.72 - 0.89 / 0.72 - 0.8
mentally.*RB*	0.59 - 0.71 - 0.82 / 0.73 - 0.73	courier.*N*	0.52 - 0.52 - 0.89 / 0.58 - 0.48	loser.*N*	0.66 - 0.66 - 0.68 / 0.66 - 0.75
swim.*N*	0.5 - 0.71 - 0.84 / 0.68 - 0.76	rat.*N*	0.41 - 0.67 - 0.86 / 0.68 - 0.86	branch.*N*	0.53 - 0.54 - 0.9 / 0.59 - 0.9
birth.*N*	0.37 - 0.82 - 0.84 / 0.83 - 0.89	gather.*V*	0.62 - 0.83 - 0.95 / 0.9 - 0.9	bankruptcy.*N*	0.5 - 0.59 - 0.89 / 0.61 - 0.56
sequence.*N*	0.72 - 0.77 - 0.85 / 0.78 - 0.74	countryside.*N*	0.65 - 0.71 - 0.65 / 0.65 - 0.79	profitable.*JJ*	0.54 - 0.48 - 0.91 / 0.58 - 0.5
dirty.*JJ*	0.4 - 0.84 - 0.97 / 0.87 - 0.93	train.*N*	0.49 - 0.74 - 0.94 / 0.86 - 0.94	broker.*N*	0.62 - 0.67 - 0.74 / 0.67 - 0.65
attempt.*V*	0.8 - 0.8 - 0.8 / 0.8 - 0.85	guest.*N*	0.52 - 0.74 - 0.89 / 0.76 - 0.87	official.*N*	0.69 - 0.7 - 0.96 / 0.69 - 0.8
link.*N*	0.57 - 0.74 - 0.96 / 0.78 - 0.94	tonic.*N*	0.78 - 0.92 - 0.83 / 0.9 - 0.99	distract.*V*	0.58 - 0.58 - 0.84 / 0.65 - 0.87
request.*N*	0.55 - 0.6 - 0.8 / 0.61 - 0.68	yen.*N*	0.67 - 0.65 - 0.65 / 0.67 - 0.67	remote.*N*	0.53 - 0.49 - 0.86 / 0.53 - 0.53
cautious.*JJ*	0.79 - 0.79 - 0.78 / 0.79 - 0.73	equal.*V*	0.56 - 0.61 - 0.6 / 0.57 - 0.62	ignition.*N*	0.77 - 0.82 - 0.95 / 0.79 - 0.89
accessory.*N*	0.61 - 0.71 - 0.96 / 0.7 - 0.88	isolated.*JJ*	0.7 - 0.94 - 0.91 / 0.94 - 0.97	weed.*N*	0.47 - 0.57 - 0.92 / 0.59 - 0.71
ounce.*N*	0.58 - 0.53 - 0.8 / 0.53 - 0.58	arrogance.*N*	0.78 - 0.78 - 0.78 / 0.78 - 0.73	vigilante.*N*	0.7 - 0.7 - 0.7 / 0.7 - 0.68
spinal.*JJ*	0.52 - 0.68 - 0.81 / 0.71 - 0.62	native.*N*	0.51 - 0.39 - 0.79 / 0.39 - 0.54	proportion.*N*	0.62 - 0.62 - 0.98 / 0.7 - 0.81
editor.*N*	0.59 - 0.64 - 0.92 / 0.59 - 0.65	redemption.*N*	0.81 - 0.19 - 0.8 / 0.81 - 0.74	pedophile.*N*	0.63 - 0.63 - 0.72 / 0.63 - 0.58
shocking.*JJ*	0.75 - 0.75 - 0.74 / 0.75 - 0.69	rookie.*N*	0.36 - 0.36 - 0.65 / 0.35 - 0.27	scenery.*N*	0.68 - 0.88 - 0.91 / 0.8 - 0.88
miserable.*JJ*	0.74 - 0.75 - 0.75 / 0.74 - 0.73	robber.*N*	0.38 - 0.64 - 0.76 / 0.67 - 0.67	ballot.*N*	0.55 - 0.68 - 0.98 / 0.7 - 0.64
scan.*N*	0.66 - 0.77 - 0.86 / 0.78 - 0.73	decency.*N*	0.59 - 0.56 - 0.86 / 0.54 - 0.54	nightfall.*N*	0.52 - 0.5 - 0.9 / 0.55 - 0.78
rider.*N*	0.6 - 0.62 - 0.85 / 0.64 - 0.7	mold.*N*	0.63 - 0.76 - 0.93 / 0.77 - 0.97	lookout.*N*	0.53 - 0.53 - 0.83 / 0.56 - 0.78
choose.*V*	0.59 - 0.68 - 0.84 / 0.7 - 0.87	brothel.*N*	0.52 - 0.52 - 0.84 / 0.48 - 0.52	violet.*N*	0.62 - 0.66 - 0.98 / 0.75 - 0.94
push.*V*	0.54 - 0.58 - 0.61 / 0.58 - 0.7	breathe.*V*	0.64 - 0.72 - 0.9 / 0.7 - 0.86	caleb.*JJ*	0.74 - 0.74 - 0.74 / 0.74 - 0.65
pajama.*N*	0.6 - 0.6 - 0.6 / 0.6 - 0.58	deliberately.*RB*	0.69 - 0.68 - 0.7 / 0.69 - 0.64	static.*JJ*	0.65 - 0.72 - 0.9 / 0.69 - 0.89
thorough.*JJ*	0.39 - 0.43 - 0.6 / 0.4 - 0.34	adjustment.*N*	0.63 - 0.65 - 0.81 / 0.73 - 0.72	baptize.*V*	0.55 - 0.59 - 0.86 / 0.55 - 0.62
stubborn.*JJ*	0.86 - 0.86 - 0.86 / 0.86 - 0.85	component.*N*	0.51 - 0.61 - 0.82 / 0.54 - 0.87	mark.*V*	0.52 - 0.65 - 0.92 / 0.72 - 0.83
penetrate.*V*	0.57 - 0.57 - 0.93 / 0.65 - 0.69	lust.*N*	0.58 - 0.58 - 0.81 / 0.6 - 0.57	modest.*JJ*	0.58 - 0.51 - 0.83 / 0.55 - 0.58
guard.*N*	0.23 - 0.69 - 0.95 / 0.77 - 0.93	classy.*JJ*	0.38 - 0.38 - 0.47 / 0.38 - 0.27	insight.*N*	0.42 - 0.42 - 0.81 / 0.46 - 0.5
decorate.*V*	0.59 - 0.77 - 0.89 / 0.78 - 0.83	unsolved.*JJ*	0.63 - 0.63 - 0.63 / 0.63 - 0.67	chart.*N*	0.86 - 0.86 - 1.0 / 0.88 - 0.94
cord.*N*	0.44 - 0.77 - 0.9 / 0.8 - 0.8	bravery.*N*	0.66 - 0.7 - 0.95 / 0.68 - 0.59	shrink.*N*	0.6 - 0.6 - 0.58 / 0.6 - 0.53
eighth.*JJ*	0.76 - 0.76 - 0.76 / 0.76 - 0.7	radius.*N*	0.29 - 0.46 - 0.88 / 0.66 - 0.8	whorehouse.*N*	0.51 - 0.46 - 0.94 / 0.41 - 0.56
poll.*N*	0.6 - 0.79 - 0.89 / 0.77 - 0.78	register.*V*	0.29 - 0.38 - 0.93 / 0.47 - 0.73	memorial.*N*	0.53 - 0.58 - 0.98 / 0.77 - 0.72
pathetic.*JJ*	0.49 - 0.27 - 0.53 / 0.49 - 0.44	paint.*V*	0.61 - 0.65 - 0.93 / 0.61 - 0.89	mineral.*N*	0.52 - 0.85 - 0.9 / 0.85 - 0.89
prey.*N*	0.55 - 0.61 - 0.91 / 0.64 - 0.6	powerless.*JJ*	0.53 - 0.47 - 0.81 / 0.51 - 0.65	tracker.*N*	0.45 - 0.5 - 0.67 / 0.52 - 0.69
settlement.*N*	0.63 - 0.63 - 0.89 / 0.64 - 0.79	exhaust.*V*	0.78 - 0.78 - 0.78 / 0.78 - 0.65	rebuild.*V*	0.6 - 0.6 - 0.97 / 0.75 - 0.64
bow.*N*	0.33 - 0.64 - 0.9 / 0.72 - 0.88	bend.*V*	0.3 - 0.53 - 0.88 / 0.65 - 0.76	consumption.*N*	0.81 - 0.83 - 0.97 / 0.83 - 0.87
dorm.*N*	0.69 - 0.69 - 0.68 / 0.69 - 0.64	porter.*N*	0.41 - 0.41 - 0.79 / 0.43 - 0.65	viper.*N*	0.57 - 0.57 - 0.97 / 0.61 - 0.78
serve.*V*	0.66 - 0.66 - 0.9 / 0.74 - 0.89	guinea.*N*	0.52 - 1.0 - 0.98 / 1.0 - 1.0	milligram.*N*	0.73 - 0.76 - 0.76 / 0.73 - 0.64
contribution.*N*	0.7 - 0.7 - 0.7 / 0.7 - 0.67	serpent.*N*	0.58 - 0.42 - 0.56 / 0.58 - 0.52	hanging.*N*	0.72 - 0.76 - 0.96 / 0.85 - 0.65
grocery.*N*	0.36 - 0.4 - 0.73 / 0.41 - 0.38	purity.*N*	0.77 - 0.77 - 0.78 / 0.77 - 0.79	expansion.*N*	0.83 - 0.88 - 0.95 / 0.94 - 0.88
sunshine.*N*	0.48 - 0.57 - 0.8 / 0.63 - 0.65	college.*N*	0.44 - 0.56 - 0.91 / 0.71 - 0.78	reluctant.*JJ*	0.64 - 0.62 - 0.68 / 0.64 - 0.6
sponsor.*N*	0.56 - 0.56 - 0.78 / 0.59 - 0.65	guarantee.*V*	0.68 - 0.68 - 0.67 / 0.68 - 0.86	declaration.*N*	0.56 - 0.82 - 0.94 / 0.82 - 0.71
broken.*JJ*	0.43 - 0.86 - 0.94 / 0.86 - 0.9	camp.*V*	0.68 - 0.71 - 0.86 / 0.72 - 0.84	regional.*JJ*	0.6 - 0.64 - 0.89 / 0.6 - 0.55
convoy.*N*	0.28 - 0.28 - 0.79 / 0.33 - 0.39	promising.*JJ*	0.77 - 0.77 - 0.77 / 0.77 - 0.61	sterile.*JJ*	0.6 - 0.6 - 0.92 / 0.69 - 0.9
unbearable.*JJ*	0.44 - 0.46 - 0.66 / 0.49 - 0.39	detector.*N*	0.54 - 0.8 - 0.87 / 0.83 - 0.79	skinny.*JJ*	0.65 - 0.62 - 0.66 / 0.65 - 0.56
vague.*JJ*	0.6 - 0.6 - 0.6 / 0.6 - 0.56	elect.*V*	0.6 - 0.61 - 0.82 / 0.63 - 0.73	amendment.*N*	0.57 - 0.57 - 0.8 / 0.57 - 0.61
torch.*N*	0.44 - 0.44 - 0.44 / 0.44 - 0.44	paramedic.*N*	0.41 - 0.42 - 0.75 / 0.44 - 0.58	binocular.*N*	0.65 - 0.65 - 0.64 / 0.65 - 0.53
boxer.*N*	0.76 - 0.76 - 0.77 / 0.76 - 0.74	shiny.*JJ*	0.79 - 0.79 - 0.79 / 0.79 - 0.76	crutch.*N*	0.61 - 0.59 - 0.7 / 0.61 - 0.64
chin.*N*	0.61 - 0.23 - 0.61 / 0.61 - 0.59	racial.*JJ*	0.56 - 0.69 - 0.85 / 0.69 - 0.56	grill.*N*	0.47 - 0.47 - 0.64 / 0.47 - 0.49
cube.*N*	0.61 - 0.95 - 0.97 / 0.95 - 0.95	vagina.*N*	0.55 - 0.55 - 0.55 / 0.55 - 0.48	cone.*N*	0.53 - 0.65 - 0.89 / 0.69 - 0.78
lean.*V*	0.3 - 0.34 - 0.85 / 0.39 - 0.65	poor.*JJ*	0.74 - 0.74 - 0.73 / 0.74 - 0.71	militia.*N*	0.64 - 0.64 - 0.65 / 0.64 - 0.68
carrier.*N*	0.59 - 0.65 - 0.94 / 0.73 - 0.95	neural.*JJ*	0.54 - 0.58 - 0.93 / 0.68 - 0.65	count.*V*	0.42 - 0.42 - 0.86 / 0.49 - 0.91
release.*V*	0.56 - 0.68 - 0.91 / 0.74 - 0.93	substitute.*N*	0.57 - 0.64 - 0.79 / 0.65 - 0.74	voluntarily.*RB*	0.63 - 0.37 - 0.63 / 0.63 - 0.54
dot.*N*	0.58 - 0.65 - 0.89 / 0.78 - 0.81	bolt.*N*	0.29 - 0.56 - 0.69 / 0.57 - 0.54	gallow.*N*	0.62 - 0.66 - 0.73 / 0.66 - 0.5
pantie.*N*	0.46 - 0.46 - 0.66 / 0.46 - 0.44	oral.*JJ*	0.74 - 0.88 - 0.91 / 0.75 - 0.96	hangover.*N*	0.6 - 0.6 - 0.77 / 0.6 - 0.53
roadblock.*N*	0.66 - 0.66 - 0.66 / 0.66 - 0.53	forensic.*JJ*	0.51 - 0.58 - 0.94 / 0.65 - 0.5	operating.*N*	0.74 - 1.0 - 1.0 / 1.0 - 1.0
blackout.*N*	0.62 - 0.66 - 0.82 / 0.66 - 0.53	stench.*N*	0.55 - 0.55 - 0.81 / 0.55 - 0.51	pasta.*N*	0.56 - 0.56 - 0.9 / 0.6 - 0.55
superstition.*N*	0.37 - 0.37 - 0.5 / 0.37 - 0.33	carpenter.*N*	0.72 - 0.72 - 0.73 / 0.72 - 0.63	overrate.*V*	0.53 - 0.53 - 0.82 / 0.53 - 0.61
silver.*N*	0.85 - 0.91 - 0.98 / 0.92 - 0.97	mri.*N*	0.78 - 0.78 - 0.77 / 0.78 - 0.78	charge.*V*	0.6 - 0.88 - 0.97 / 0.9 - 0.93
complex.*JJ*	0.53 - 0.54 - 0.96 / 0.54 - 0.66	deport.*V*	0.63 - 0.63 - 0.68 / 0.65 - 0.66	rift.*N*	0.68 - 0.78 - 0.78 / 0.73 - 0.84

Table 7: Part-2: Lexical selection model test accuracies for a English-Greek words.

Focus word	FreqBaseline - DTree - LinearSVM Train/Test - BERT	Focus word	FreqBaseline - DTree - LinearSVM Train/Test- BERT	Focus word	FreqBaseline - DTree - LinearSVM - BERT
descent.*N*	0.64 - 0.64 - 0.94 / 0.67 - 0.86	suit.*N*	0.63 - 0.63 - 0.95 / 0.7 - 1.0	shrine.*N*	0.51 - 0.51 - 0.85 / 0.56 - 0.61
coaster.*N*	0.68 - 0.85 - 0.93 / 0.85 - 0.94	vamp.*N*	0.53 - 0.53 - 0.96 / 0.57 - 0.57	wisely.*RB*	0.56 - 0.58 - 0.84 / 0.63 - 0.58
teacher.*N*	0.4 - 0.91 - 0.96 / 0.91 - 0.96	tangible.*JJ*	0.52 - 0.57 - 0.74 / 0.57 - 0.57	boxing.*N*	0.59 - 0.6 - 0.71 / 0.6 - 0.53
blow.*V*	0.84 - 0.81 - 0.91 / 0.89 - 0.96	jap.*N*	0.61 - 0.61 - 0.63 / 0.61 - 0.59	podium.*N*	0.62 - 0.62 - 0.94 / 0.75 - 0.75
fairy.*N*	0.69 - 1.0 - 1.0 / 1.0 - 1.0	jelly.*N*	0.52 - 0.59 - 0.72 / 0.59 - 0.59	ruler.*N*	0.48 - 0.48 - 0.91 / 0.42 - 0.7
diversity.*N*	0.5 - 0.41 - 0.88 / 0.45 - 0.41	proud.*JJ*	0.64 - 0.64 - 0.65 / 0.64 - 0.7	goat.*N*	0.62 - 0.62 - 0.7 / 0.63 - 0.77
postmortem.*N*	0.48 - 0.48 - 0.96 / 0.57 - 0.64	extend.*V*	0.62 - 0.76 - 0.9 / 0.76 - 0.98	bakery.*N*	0.45 - 0.45 - 0.73 / 0.38 - 0.5
clamp.*N*	0.62 - 0.62 - 0.92 / 0.56 - 0.53	remote.*JJ*	0.49 - 0.49 - 0.88 / 0.59 - 0.65	dusty.*JJ*	0.79 - 0.85 - 0.97 / 0.82 - 0.88
reconsider.*V*	0.61 - 0.61 - 0.6 / 0.61 - 0.58	rethink.*V*	0.79 - 0.79 - 0.79 / 0.79 - 0.79	agne.*N*	0.62 - 0.62 - 0.96 / 0.5 - 0.69
knit.*V*	0.63 - 0.65 - 0.78 / 0.63 - 0.74	conductor.*N*	0.47 - 0.5 - 0.91 / 0.61 - 0.82	memorable.*JJ*	0.62 - 0.62 - 0.72 / 0.65 - 0.6
medium.*JJ*	0.52 - 0.6 - 0.94 / 0.78 - 0.68	elite.*JJ*	0.65 - 0.65 - 0.95 / 0.67 - 0.69	holodeck.*N*	0.52 - 0.52 - 0.97 / 0.7 - 0.67
bastard.*N*	0.38 - 0.68 - 0.95 / 0.76 - 0.89	helm.*N*	0.81 - 0.83 - 0.79 / 0.81 - 0.78	cradle.*N*	0.66 - 0.76 - 0.97 / 0.84 - 0.95
vacant.*JJ*	0.69 - 0.77 - 0.99 / 0.82 - 0.95	piss.*V*	0.69 - 0.69 - 0.69 / 0.69 - 0.7	motto.*N*	0.48 - 0.46 - 0.46 / 0.48 - 0.45
parliament.*N*	0.59 - 0.59 - 0.61 / 0.59 - 0.57	hideout.*N*	0.56 - 0.56 - 0.56 / 0.56 - 0.46	bond.*N*	0.63 - 0.66 - 0.9 / 0.61 - 0.98
pretzel.*N*	0.7 - 0.7 - 0.69 / 0.7 - 0.62	deek.*N*	0.61 - 0.61 - 0.62 / 0.61 - 0.48	abnormal.*JJ*	0.66 - 0.61 - 0.67 / 0.66 - 0.59
abrasion.*N*	0.61 - 0.39 - 0.6 / 0.61 - 0.71	countess.*N*	0.58 - 0.58 - 0.58 / 0.58 - 0.48	dioxide.*N*	0.66 - 0.66 - 0.68 / 0.66 - 0.72
walkie.*N*	0.68 - 0.68 - 0.68 / 0.68 - 0.71	tight.*N*	0.62 - 0.62 - 0.73 / 0.62 - 0.64	invaluable.*JJ*	0.65 - 0.65 - 0.64 / 0.65 - 0.65
tango.*N*	0.52 - 0.52 - 0.94 / 0.6 - 0.55	pedal.*N*	0.53 - 0.49 - 0.92 / 0.53 - 0.63	petal.*N*	0.56 - 0.83 - 0.92 / 0.78 - 0.81
expand.*V*	0.54 - 0.76 - 0.97 / 0.84 - 0.95	envy.*N*	0.56 - 0.56 - 0.92 / 0.64 - 0.71	partially.*RB*	0.57 - 0.66 - 0.87 / 0.55 - 0.66
resourceful.*JJ*	0.6 - 0.6 - 0.6 / 0.6 - 0.51	heinous.*JJ*	0.72 - 0.28 - 0.84 / 0.88 - 0.84	contusion.*N*	0.55 - 0.53 - 0.9 / 0.61 - 0.58
seeker.*N*	0.55 - 0.55 - 0.58 / 0.55 - 0.59 weep.*V*	0.62 - 0.67 - 0.61 / 0.62 - 0.95	jewel.*N*	0.7 - 0.3 - 0.7 / 0.7 - 0.7	
algae.*N*	0.55 - 0.67 - 0.85 / 0.58 - 0.58	newlywed.*N*	0.62 - 0.62 - 0.6 / 0.62 - 0.79	barge.*N*	0.56 - 0.56 - 0.91 / 0.59 - 0.75
fatigue.*N*	0.52 - 0.45 - 0.88 / 0.5 - 0.59	farmhouse.*N*	0.54 - 0.54 - 0.64 / 0.54 - 0.46	imply.*V*	0.72 - 0.72 - 0.68 / 0.72 - 0.9
gator.*N*	0.73 - 0.76 - 0.73 / 0.73 - 0.84	outskirt.*N*	0.52 - 0.52 - 0.82 / 0.52 - 0.76	gel.*N*	0.5 - 0.73 - 0.81 / 0.75 - 0.68
riddance.*N*	0.52 - 0.52 - 0.7 / 0.38 - 0.76	infectious.*JJ*	0.54 - 0.57 - 0.79 / 0.57 - 0.54	chord.*N*	0.62 - 0.72 - 0.8 / 0.88 - 0.66
hide.*V*	0.75 - 0.75 - 0.98 / 0.86 - 0.82	conquer.*V*	0.56 - 1.0 - 1.0 / 1.0 - 1.0	compression.*N*	0.76 - 0.91 - 0.91 / 0.88 - 0.85
cunning.*JJ*	0.56 - 0.56 - 0.87 / 0.62 - 0.66	plague.*V*	0.62 - 0.62 - 0.98 / 0.94 - 1.0	morbid.*N*	0.6 - 0.52 - 0.63 / 0.6 - 0.56
particle.*N*	0.55 - 0.92 - 0.97 / 0.92 - 1.0	theo.*N*	0.64 - 0.64 - 0.96 / 0.62 - 0.5	femur.*N*	0.61 - 0.61 - 0.61 / 0.61 - 0.61
commit.*V*	0.92 - 0.93 - 0.93 / 0.92 - 0.99	rivalry.*N*	0.58 - 0.5 - 0.84 / 0.54 - 0.54	pyjama.*N*	0.62 - 0.62 - 0.62 / 0.62 - 0.66
scatter.*V*	0.59 - 0.63 - 0.61 / 0.59 - 0.67	donor.*N*	0.53 - 0.88 - 0.94 / 0.91 - 0.94	smither.*N*	0.6 - 0.6 - 0.97 / 0.47 - 0.43
yuan.*N*	0.59 - 0.59 - 0.88 / 0.73 - 0.73	autopsy.*N*	0.56 - 0.56 - 0.92 / 0.6 - 0.52	gullible.*JJ*	0.52 - 0.41 - 0.89 / 0.37 - 0.63
absolution.*N*	0.59 - 0.59 - 0.9 / 0.48 - 0.69	modesty.*N*	0.58 - 0.75 - 0.83 / 0.71 - 0.71	handgun.*N*	0.58 - 0.58 - 0.59 / 0.58 - 0.42
firecracker.*N*	0.59 - 0.64 - 0.86 / 0.5 - 0.73	honorary.*JJ*	0.58 - 0.67 - 0.91 / 0.67 - 0.92	rod.*N*	0.66 - 0.66 - 0.98 / 0.74 - 0.83
kneecap.*N*	0.54 - 0.54 - 0.95 / 0.62 - 0.69	insubordination.*N*	0.56 - 0.56 - 0.89 / 0.32 - 0.64	poacher.*N*	0.52 - 0.52 - 0.98 / 0.43 - 0.43
railing.*N*	0.56 - 0.56 - 0.61 / 0.56 - 0.72	tremor.*N*	0.59 - 0.55 - 0.82 / 0.64 - 0.86	coalition.*N*	0.52 - 0.43 - 0.83 / 0.61 - 0.52
carnage.*N*	0.55 - 0.52 - 0.56 / 0.55 - 0.59	memento.*N*	0.59 - 0.38 - 0.84 / 0.38 - 0.59		
		Overall Average:	58.56 - 63.79 - 66.46 - 71.74		

Table 8: Part-3: Lexical selection model test accuracies for a English-Greek words.