Predictors of L2 word learning accuracy: A big data investigation

Elise W.M. Hopman (hopman@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Bill Thompson (biltho@mpi.nl)
Language and Cognition Department, Max Planck Institute for Psycholinguistics
Nijmegen, The Netherlands

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Gary Lupyan (lupyan@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

What makes some words harder to learn than others in a second language? Although some robust factors have been identified based on small scale experimental studies, many relevant factors are difficult to study in such experiments due to the amount of data necessary to test them. Here, we investigate what factors affect the ease of learning of a word in a second language using a large data set of users learning English as a second language through the Duolingo mobile app. In a regression analysis, we test and confirm the well-studied effect of cognate status on word learning accuracy. Furthermore, we find significant effects for both cross-linguistic semantic alignment and English semantic density, two novel predictors derived from large scale distributional models of lexical semantics. Finally, we provide data on several other psycholinguistically plausible word level predictors. We conclude with a discussion of the limits, benefits and future research potential of using big data for investigating second language learning.

Keywords: second language learning; vocabulary; big data; corpus analysis; distributional semantics;

Introduction

Spanish speakers learning English on Duolingo are more than twice as likely to err with the word 'blue' than with the word 'gray'. They are also about 1.5 times more likely to make a mistake with the word 'blue' than Italian speaking learners are.[1] What explains such differences in word learning? In this paper we investigate these questions by examining which word level factors predict accurate word learning in a large, naturalistic dataset of Spanish, Italian, and Portuguese speakers learning English.

The second language (L2) literature has identified several word level factors that predict how easy a new word in L2 will be to remember (De Groot & Keijzer, 2000). The strongest predictor is concreteness. All else equal, concrete words tend to be easier to learn. Researchers hypothesize that concrete words have richer representations in memory, and this richer representation provides more opportunities for the learner to associate the L2 word with the L1[2] word. Another predictor of learning ease is whether the L2 wordform is a cognate (largely shares form and meaning) of the L1 form (e.g., 'actor' in English and Spanish). Studies have manipulated word frequency (both L1 frequency and L2 exposure), but this predictor does not have a simple robust effect (De Groot & Keijzer, 2000). Finally, there is a body of work on translation ambiguity, showing that if a word in one language has two distinct translations into another language, it will be harder to learn (Bracken, Degani, Eddington & Tokowicz, 2017).

Other potentially interesting factors that could matter for L2 word learning have been proposed but not yet studied (De Groot & Van Hell, 2005). Because L2 word learning studies are typically conducted in small classrooms or laboratories, it has been difficult to study word level factors at scale. There have been several recent calls in the L2 literature to move beyond small classroom studies and towards more quantitatively robust analyses (e.g. Milin, Divjak, Dimitrijevic, & Baayen, 2016; Norris & Ortega, 2000). One larger scale study successfully assessed some factors that predict L2 word recognition in fluent bilinguals (Lemhöfer et al, 2008). Furthermore, there has been a broader call in the cognitive sciences to use big and natural datasets to shed light on the questions that the field has been struggling to answer with experimental studies (Paxton & Griffiths, 2017).

Here, we analyze a big and naturally occurring dataset to analyze word level factors that may affect word-learning performance of L2 English learners. The dataset we use contains a total of 36,799 people using the program Duolingo to learn English (Settles & Meeder, 2016).

[1] Example data-points based on the dataset used in this paper, described in more detail in the methods section.

[2] L1 = first language; We assume that for most users, the user interface language in Duolingo is their native language.

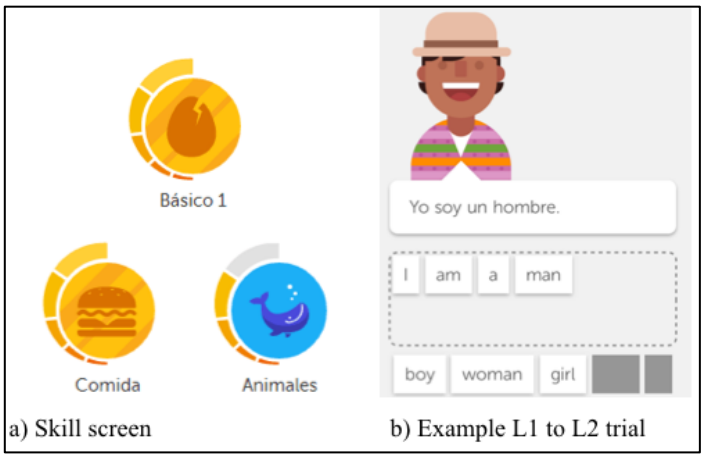

Figure 1: Screenshots from the Duolingo user interface. (a) Progress bars wrapped around each skill encourage users to do learning sessions that contain weak words. (b) Example trial where the user translates a sentence from L1 Spanish into L2 English by choosing words from a word bank. In this example, the user has already translated the sentence. Note that in this example, several words are embedded in a longer phrase, as is typical in Duolingo.

Duolingo is a popular free online program that gamifies second language learning. It combines several best practices from learning research in its design including explanations, implicit instruction and mastery learning (see Fig. 1 and Settles & Meeder). In the app, a user does practice sessions revolving around a skill (e.g. 'food' or 'animals', Fig. 1a). A practice session consists of multiple trials, often involving several words embedded in larger phrases (Fig. 1b). Different trials embed different language learning tasks —written translation, fill in the gap sentences, multiple choice, matching tasks — in both the L2 to L1 and the L1 to L2 direction. Duolingo applies a carefully constructed curriculum, so different skills are learned in a specified order with the goal of aiding learning. To do so, Duolingo's algorithm predicts when words are becoming weak in memory and should be studied again by the user (Fig. 1a).

We focus our analyses on two types of word-level predictors. The first is orthographic similarity between the L1 and L2 wordform — a measure that can be viewed as a continuous estimate of cognate status. We expect words with similar L1/L2 forms to be learned more easily. The second is a novel measure of semantic similarity. For each pair of words that are translation-equivalents, we compute semantic similarity based on distributional semantics of L1 and L2. This measure — described in more detail below — allows us to ask whether words whose meanings are more alignable in L1 and L2 are learned more easily.

Method[3]

Duolingo dataset manipulation

The English learning part of the Duolingo dataset (Settles & Meeder, 2016) as originally released contains 5.01 million instances of data. Each instance contains accuracy data for a single user during a single Duolingo session on a single lexeme, though that lexeme may have been seen multiple times during that session. A lexeme is a word with specified morphological form. For example, 'girl' and 'girls' are separate lexemes, as one is singular and the other plural. While the Duolingo dataset has separate entries for different lexemes that correspond to the same word (e.g. the lexemes 'cat' and 'cats' both correspond to the lemma 'cat'), most of the other corpora that we used to obtain word-level predictors did not. Thus, we aggregated the Duolingo dataset by word, collapsing from 2983 different lexemes to 1412 different words, with data on each word originating from 1-21 different lexemes.

The Duolingo dataset contained data collected over two consecutive weeks, resulting in several (sometimes thousands of) instances for a given learner being presented with a given word. In addition to accuracy data for the current session, the data instance also included a timestamp and aggregated accuracy data on all previous encounters that the learner had with that word. Since we are not interested here in the time course of learning, we aggregated accuracy data to get a single datapoint per learner per word that detailed how often the learner had seen the word in total and how often they had been correct on that word. This aggregated dataset contained 1.86 million user-word data points taken from English courses with three different user interface languages: Portuguese, Spanish and Italian.

Because the order and frequency with which words are presented to a learner are not random in Duolingo, we used two user level measures as control predictors. The first, which we call *word experience*, is the total number of times that a user saw a given word in Duolingo. The second, *user experience*, is the sum of word experience for all words a user has practiced. Both of these predictors were log-scaled since their distributions were highly skewed, with many words practiced only a few times and some words practiced many times, and many users practicing English only a little and some users practicing it a lot.

We expect that user experience is predictive of word learning accuracy, with more experienced users doing better than less experienced users. For word experience, the naive prediction would be that the more experience a user has with a certain word, the better they should do on it. However, Duolingo will present a word more often after a user has made a mistake on it, meaning that words that a user has more difficulty on will be practiced more often. Due to this biased sampling procedure, average accuracy will be lower for words that have been practiced more often.

[3] An OSF archive containing all data and analysis scripts for this paper will be made publicly available at osf.io/uwdcm

Word level predictors

We used Google Translate to obtain word translations of the English words into three different user interface languages, since the dataset released by Duolingo did not include the L1 translation equivalents used in the app. We then calculated the minimum edit[4] (Levenshtein) distance between the English word and its translation equivalent as a continuous measure of cognate status, which we refer to as *translation distance*. Earlier L2 studies have mainly used cognate status as a binary predictor, comparing perfect cognates to non-cognates. Our prediction for this measure of cognate status goes in the same direction: the smaller the translation distance between the two words in a translation pair is, the easier it should be to learn the word.

We focus on two main semantic predictors[5], both derived from large scale distributional models of lexical semantics. The first, which we call *semantic alignment,* measures one aspect of how similarly a word is situated in the semantic representations of two languages. In intuitive terms, this captures the quality of a translation pair from a distributional perspective: If a translation pair keep similar semantic associations in two languages, then their meanings can be understood to be more aligned, and perhaps easier to learn. To compute semantic alignment (denoted rho), we obtained distributional models of lexical semantics for English, Spanish, Portuguese, and Italian. These networks were recently released by Facebook Artificial Intelligence Research (Bojanowski et al., 2017). Each model encodes the vocabulary of its language as points (or vectors) in an abstract semantic space. The configuration of these points is estimated by training a neural network to predict the words that appear often with a word in a large dataset of text derived from Wikipedia, and using this neural network to assign vectors to words. Close proximity in this vector space, as measured by cosine distance, implies a close semantic relationship between words.

To compute the semantic alignment of a translation pair, we found the semantic neighborhood of each word - the *N* closest words within their respective semantic spaces (here *N* = 40). We only included in each space those words from the Duolingo dataset for which we were able to translate between the two languages using Google Translate. We then compared the cosine similarity of the words in the semantic neighborhoods of the translation pair. To do so, we first identified common neighbors, i.e, words which are in the semantic neighborhood of the target in English, and whose translation equivalent appears in the semantic neighborhood of the target in the other language. For each of these common neighbors, we calculated the cosine similarity between the target and the neighbor, in both semantic spaces. We calculated total structural alignment by computing the rank correlation statistic (Spearman's rho) between these aligned target-neighbor cosine similarities[6]. If the similarity structure is closely aligned between the semantic spaces of the two languages near the translation pair, rho will approach 1, and words with a large rho should be easier to learn. If the similarity structure were inverted between languages (i.e. the closest common neighbor in English is the most distant common neighbor Spanish), rho would approach -1; such words should be harder to learn.

Thus, while our new measure semantic alignment is conceptually related to translation ambiguity, it is calculated in an automated way that does not rely on human similarity judgments, making it easier to calculate for a wide variety of words between different language pairs. Our prediction that words that are less semantically aligned should be harder to learn corresponds to experimental findings that words that are more translation ambiguous are harder to learn (Bracken et al., 2017).

In addition to semantic alignment, we also recorded the mean edit (Levenshtein) distance between the target English word and its N = 40 closest neighbors in English semantic space. This measure identifies, for example, whether a word has many morphological variants, which are close in orthographic and semantic space. We call this variable *English morphological density*.

The third semantic predictor is English *semantic density* (a measure distinct from phonological or orthographic density used in some past studies). By identifying the *N* nearest-neighbors of a target English word, we can obtain a measure of how concentrated the region of semantic space the word occupies. Some words are surrounded by many other words with similar or related meanings, while others occupy isolated territory with few close associations. This aspect of a word can be quantified by the mean cosine distance to its closest 40 neighbors (in this dataset), and can be understood as a close analog of local clustering coefficients used in network analysis (how many words connected to a target word are connected to each other). An English word from a dense semantic neighborhood may be more confusable with its neighbors and thus harder to learn.

Besides our three measures of core interest (translation distance, semantic alignment and semantic density), we added several other exploratory word level predictors[7]. We added *English word recognition*, a z-scored Reaction Time measure for native English speakers in a lexical decision task as a predictor (Balota et al, 2007). We added this measure from the English Lexicon Project to see if and how native speaker ease of processing of a word might be related to early L2 English word learning accuracy.

[4] Minimum number of edits needed to change one word into another. For example, the Levenshtein distance between 'cat' and 'gato' is 2: the 'c' needs to become a 'g' and an 'o' needs to be added.

[5] The method for calculating the new semantic predictors mentioned in this section is described in more detail in Thompson, Roberts & Lupyan (2018).

[6] Neighbor overlap and semantic alignment are correlated statistics; in ongoing work, we are exploring this relationship further.

[7] Our dataset includes additional predictors, e.g. measures of phonological neighborhood density; predictors that didn't significantly contribute to model fit in initial analyses were left out of later analyses.

We used a multilingual WordNet (Bond & Paik, 2012) to obtain a measure for the number of *distinct meanings* a word has. We added *concreteness* (Brysbaert et al., 2014) as a predictor because the L2 word learning literature identifies this as the most robust predictor for early L2 word learning accuracy, with more concrete words easier to learn.

We also add user interface language *word frequency* (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2012). Some, but not all previous studies investigating early L2 word learning accuracy find that more frequent words are easier to learn. Since users' frequency of exposure to a word in English is already captured by our *word experience* variable, we opted to use frequency estimates based on the user's L1.

Finally, to control for part of speech, we included the dominant part of speech based on a parse of SUBTLEX-US in our dataset (Brysbaert et al, 2012). Since most closed part of speech categories (e.g. articles) only consist of a handful of words in our dataset, we collapsed this into 4 big open categories (noun, verb, adjective, adverb) and a 5th miscellaneous category containing all other word types.

Final corpus descriptives

Combining all word level predictors with the Duolingo dataset, we were able to collect a measure of each predictor on 1064 different English words. We removed data from users who completed less than 41 data instances (15% of total users). We also excluded words for individual users that had been practiced fewer than three times (2.6 % of user-word data points). Table 1 shows, Mean, SD, and range for all predictors in our final corpus.

Table 1. Descriptive statistics (Mean, Standard Deviation and Range) for the predictors in data sets for the three different user interface languages.

Predictor	Language of dataset	Mean (SD)	Range	n[a]
user experience	Spanish	5.59 (1.15)	3.71 to 12.69	26628
	Portuguese	5.6 (1.13)	3.71 to 10.3	7329
	Italian	5.72 (1.17)	3.71 to 10.17	2843
word experience	Spanish	2.3 (0.82)	1.1 to 9.51	1007092
	Portuguese	2.3 (0.79)	1.1 to 8.51	258609
	Italian	2.23 (0.79)	1.1 to 7.14	124471
translation distance	Spanish	3.99 (2.23)	0 to 12	979
	Portuguese	4.1 (2.11)	0 to 12	952
	Italian	4.3 (2.37)	0 to 13	956
semantic alignment	Spanish	0.37 (0.3)	-0.86 to 0.94	979
	Portuguese	0.36 (0.29)	-0.75 to 1	952
	Italian	0.35 (0.3)	-0.77 to 0.96	956
English morphological density	Spanish	5.91 (1.28)	3.4 to 11.32	979
	Portuguese	5.94 (1.29)	3.4 to 11.32	952
	Italian	5.91 (1.27)	3.4 to 11	956
English semantic density	Spanish	0.42 (0.07)	0.25 to 0.68	979
	Portuguese	0.42 (0.07)	0.25 to 0.71	952
	Italian	0.42 (0.08)	0.25 to 0.71	956
English word recognition	Spanish	-0.59 (0.19)	-1 to 0.17	979
	Portuguese	-0.59 (0.19)	-1 to 0.17	952
	Italian	-0.59 (0.18)	-1 to 0.17	956
distinct meanings	Spanish	1.48 (0.82)	0 to 3.56	911
	Portuguese	1.46 (0.69)	0 to 3.5	896
	Italian	1.41 (0.65)	0 to 3.66	892
concreteness	Spanish	3.32 (1.11)	1.12 to 5	979
	Portuguese	3.32 (1.11)	1.12 to 5	952
	Italian	3.31 (1.12)	1.12 to 5	956
word frequency	Spanish	7.86 (1.75)	0.69 to 14.16	911
	Portuguese	8.25 (1.75)	0.69 to 14.38	896
	Italian	8.54 (1.8)	0 to 14.67	892

[a]The number of data points n used to estimate the reported statistics of each predictor are different due to properties of the data set. For example, user experience is determined based on the number of users in a dataset, concreteness is determined based on the number of English words in a dataset, and word experience is based on unique user-word data-points. Finally, since there were many more Spanish users, some words had enough users to make it into the Spanish but not the Portuguese or Italian datasets.

Results

Overall accuracy was 90% and was remarkably similar for the three L1 languages in the dataset: Italian (90.9%), Spanish (90.0%), Portuguese (89.9%). Not surprisingly, performance was better for users who used Duolingo more, b=.10, 95% CI = [0.094, 0.104], t=41.2. Despite the robustness of user experience as a predictor of performance, absolute differences in performance were quite small. Users at the lowest quartile of usage had 90.5% accuracy, while users at the highest quartile of usage had 91.3% accuracy. This highly restricted range of accuracies speaks to the adaptive nature of Duolingo's platform. When users make mistakes, they are more likely to practice the words later, keeping overall accuracy high and relatively constant. Accordingly, controlling for overall user experience, greater experience with a given word is associated with *lower* accuracy b=-.04, 95% CI=[-0.052, -0.047], t=-36.8, most likely because users get increased exposure to a word *because* they made mistakes with it. This adaptive-sampling property of Duolingo makes it difficult to predict accuracy from word-level properties, but as we describe below, we can nevertheless account for what makes some words more difficult than others.

We modeled accuracy for each user-word datapoint with mixed-effects regression, running separate models for each of three user interface languages (Italian, Spanish, and Portuguese). This model included a random intercept for user (since each user had seen multiple words), English word (since each word was seen by multiple users) and major part of speech (to ensure that some predictors like concreteness are not confounded by differences between parts of speech). We show standardized coefficients and 95% CIs for each L1 language model in Fig. 2. For example, a 1SD increase in user experience for Portuguese users leads to .12 SD increase in overall accuracy. Corresponding p-values can be inferred from the displayed 95% CIs.

Controlling for both user and word experience, we find that translation distance between L1 and English is negatively associated with accuracy. This relationship is significant for Spanish-English, Italian-English and is marginal for Portuguese-English. Accuracy in all three languages is associated positively with semantic alignment. The larger the semantic alignment between L1 and English for a given translation-pair, the more likely people are to be accurate (controlling for all other factors in the model). Accuracy in all three L1s models is associated negatively with the density of the English word's *semantic* neighborhood. Words having high density neighborhoods such as "something", "anything" and "anybody" pose greater learning challenges (controlling for other factors) compared to words such as "register", "profile", and "special", which reside in neighborhoods with lower semantic density.

Aside from these three predictors, we found some other effects which we did not explicitly predict and which should be interpreted with caution. Words being learned with larger English morphological density were associated with larger accuracy for Portuguese-English users, but this was not a significant predictor in the Italian and Spanish datasets. Concreteness is not reliably associated with accuracy when we take part of speech into account. Puzzlingly, more concrete words were associated with marginally lower accuracy for Portuguese-English users. Finally, longer lexical decision times from native-English speakers (Balota et al., 2007) were associated with numerically lower accuracy. However, this predictor was only significant for the Portuguese-English users.

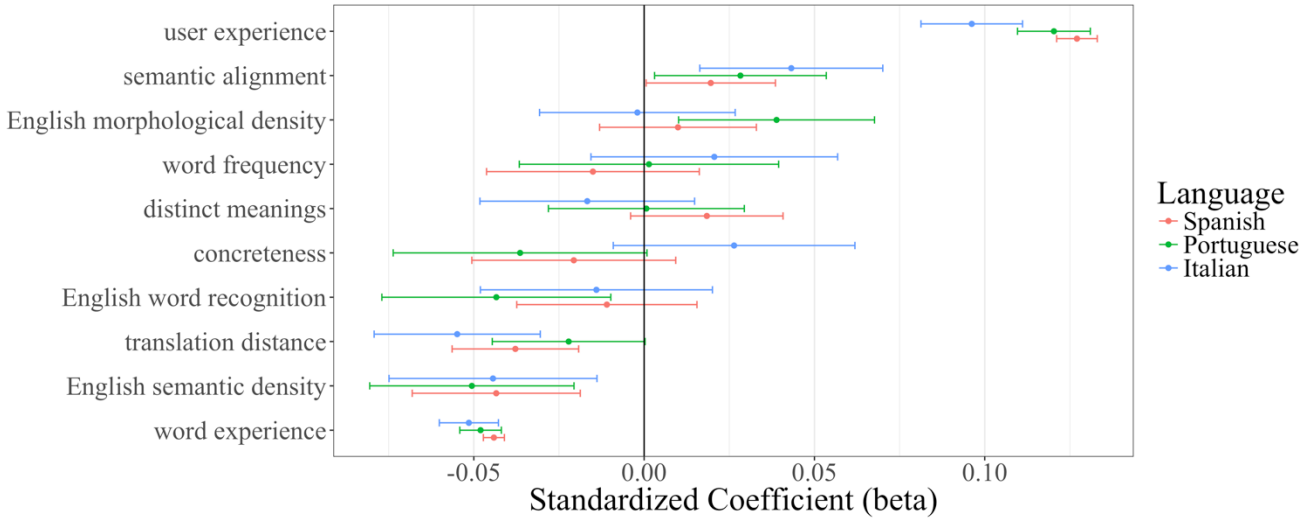

Figure 2: Regression results for accuracy. Standardized coefficients and 95% CIs are plotted for the predictors in each L1 language model. Standardized coefficients are interpreted as in the following example: a 1SD increase in user experience for Portuguese users leads to .12 SD increase in overall accuracy.

Discussion

We predicted accuracy of learning English words by Italian, Spanish, and Portuguese users of Duolingo as measured in a large naturally collected dataset. We found evidence for several novel factors for L2 word learning. As predicted, English words having smaller orthographic distances to their translation-equivalents were easier to learn. In addition to orthographic similarity, *semantic* similarity (obtained by cross-linguistic alignment of word embeddings derived from distributional semantics) was also associated with higher accuracy. Finally, words residing in dense English semantic neighborhoods were harder to learn than words residing in less dense semantic neighborhoods, when controlling for all other predictors. We also examined several other predictors that might be of interest to researchers investigating L2 word learning (Fig. 2).

These results can form the basis for future experimental studies. Since classroom and experimental studies are often necessarily limited in the number of words they test, predictors could first be investigated in this or similar big data, so that items for controlled studies can be strategically chosen.

Limitations

There are several aspects of this naturally occurring dataset that, despite its size, limit its usefulness for answering theoretically interesting questions about L2 learning. The Duolingo curriculum is constructed to maximize accuracy. This leads to a much smaller range of accuracy scores than is usually seen in experimental studies of word-learning. Furthermore, highly biased sampling of words produces a non-random ordering that affects several of our word level predictors. This may be why we do not find effects for certain predictors that are typically strong in experimental studies. Relatedly, Duolingo's algorithm presents a word sooner for repeated study after the user gets it wrong on a trial. Finally, we have very little information about the users. In experimental studies, a participant's language background and other demographics that might influence learning abilities can be measured in questionnaires, whereas for this dataset even a user's native language is only inferred. Such limitations in using big and naturally occurring datasets should not, however, preclude their use in cognitive science (Paxton & Griffiths, 2017).

On the positive side, these data provide certain ecological validity absent from lab studies, and allow us to look at a longer slice of learning time compared to typical lab studies. Duolingo users are self-motivated to learn a second language, which is not necessarily true for learners in school classrooms who might just be meeting a curriculum requirement and participants in experimental studies. Furthermore, the size of this dataset allowed us to investigate many more word level predictors than can easily be manipulated in any one classroom study.

References

Balota, D. A., Yap, M. J., Hutchinson, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3).

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5*.

Bond, F. & Paik, K. (2012). A survey of wordnets and their licenses. *Proceedings of the 6th Global WordNet Conference (GWC 2012)*. Matsue.

Bracken, J., Degani, T., Eddington, C. & Tokowicz, N. (2017). Translation semantic variability: How semantic relatedness affects learning of translation-ambiguous words. *Bilingualism: Language and Cognition 20 (4)*.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods, 44*, 991-997.

Brysbaert, M., Warriner, A.B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *33*(2).

De Groot, A., & Keijzer, R. (2000). What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting. *Language Learning, 50*(1).

De Groot, A.M.B., & Van Hell, J.G. (2005). The learning of foreign language vocabulary. In J.F. Kroll & A.M.B. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches*. Oxford University Press.

Lemhöfer, K., Dijkstra, T., Schriefers, H., Baayen, R.H., Grainger, J. & Zwitserlood, P. (2008). Native language influences on word recognition in second language: a megastudy. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*.

Milin, P., Divjak, D., Dimitrijevic, S., & Baayen, R. H. (2016). Towards cognitive plausible data science in language research. *Cognitive Linguistics, 27*(4).

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning, 50*(3).

Paxton, A. & Griffiths, T. L. (2017). Finding the traces of behavioral and cognitive processes in big data and naturally occurring datasets. *Behavior Research Methods, 49*.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. In *Proceedings of the 54th Annual Meeting of the ACL*.

Thompson, B., Roberts, S., & Lupyan, G. (2018). Quantifying Semantic Similarity Across Languages. In *Proceedings of the 40th annual conference of the cognitive science society*.