LINGPREDICT: DEVELOPMENTAL NORMS AND SECOND LANGUAGE ACQUISITION

BENARD HAUGEN, VIKRAM JAMBULAPATI, MANJEET KAUR,
SARA SANCHEZ-ALONSO, SAM SCHIAVONE

1. Introduction

1.1. **Background.** The goal of this project is to investigate how the age at which a person typically learns a word in their native language (L1) influences their ability to learn the corresponding word in a foreign language (L2). We then use this feature and others to try to predict errors of Duolingo users using the data from Duolingo's Shared Task on Second Language Acquisition Modeling (SLAM) [Set18].

The main novel feature of our project is the use of normative developmental data to model acquisition of words in L1 using the data compiled by Wordbank [FBYM17]. This data arises from MacArthur-Bates Communicative Development Inventory (CDI) surveys filled out by parents of young children. While previous groups, such as [RPH+18] also considered L1 age of acquisition as a feature, using the data from [KSGB12], this data arises from crowd-sourced adult recollections, which are not as accurate.

1.2. **Stakeholders.**

- Curriculum designers for language courses
- Duolingo developers and researchers
- Developmental linguists and psychologists

1.3. **KPIs and metrics.**

- **Correlations:** Assess relationships among linguistic and learner-level features (e.g., Age of Acquisition, exposure, error rate) to identify potential collinearity or bias.
- **Primary KPI:** Learner correctness, evaluated by Area Under the Curve (AUC), Average Precision (AP) and overall F1-score to balance precision and recall.
- **Secondary KPIs:** Precision and recall, reflecting accuracy and coverage of predicted learner errors. This assess trade-offs between accuracy and sensitivity.

2. Data sources

2.1. **Wordbank and developmental norms.** Our source for data on developmental norms and word acquisition is the online Wordbank database [FBYM21, FBYM17], available at `https://wordbank.stanford.edu/`. This site collects and aggregates data provided by parents on the language acquisition of their children. We used the data listed under English (American) and Spanish (Mexican) to model the developmental acquisition of words in English and Spanish.

- **CDIs**: The data in Wordbank comes from responses by parents to MacArthur-Bates Communicative Development Inventories (CDIs). These CDIs are surveys with lists of words; parents indicate whether or not their child knows each word.

Date: November 5, 2025.

- **Inventories**: The Wordbank data draws upon two primary CDI forms: the *Words and Gestures (WG)* form, which targets younger children (typically 8–18 months) and focuses on early word comprehension and non-verbal communication; and the *Words and Sentences (WS)* form, which targets older children (typically 16–30 months) and focuses on expressive vocabulary and emerging grammatical complexity.
- **Measures**: Acquisition of a word is measured by two linguistic measures:
  - *Understands*: child recognizes or grasps the meaning of the word.
  - *Produces*: child actively says the word (or uses the corresponding gesture/sign).

  Comprehension of a word is nearly always acquired before its production.
- **Aggregation of data**: The authors of Wordbank have tabulated the proportion of children who know each word at a given age. This by-word summary data can be downloaded from `https://wordbank.stanford.edu/data/?name=item_data`.
- **Unilemmas**: In order to study the acquisition curve for a given concept across multiple languages, the authors of Wordbank include universal lemmas, or *unilemmas*. These unilemmas provide cross-linguistic conceptual mappings and allow us to translate word tokens from L2 to L1. The unilemma data was obtained using the associated R package `wordbankr`.

2.2. **Duolingo SLAM.** In the past, language learning mainly occurred in classrooms, but with the rise of mobile learning (m-learning), platforms like Duolingo now let users learn anytime and anywhere. Using its vast learner data, Duolingo launched the SLAM challenge to study and predict how accurately users translate words while learning. We are using the data from the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM) [SBG+18], [Set18].

The Duolingo SLAM dataset consists of 2 million words that have been translated by 6,000 second language learners. Each translated word has been labeled as correctly translated or not by a specific user and its morpho-syntactic features. Each exercise contains an associated anonymous user ID, country code, days since the learner started the course, device platform, session type, exercise format and response time in seconds. This dataset is publicly available for research and teaching purposes.

3. Feature engineering

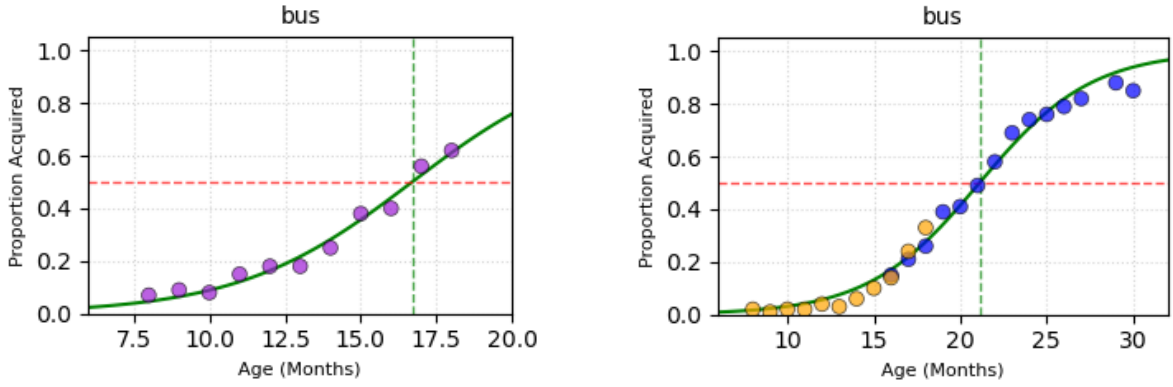
In our inferential analysis and predictive models we considered the following data features.

- **Age of acquisition**: The age of acquisition (AoA) data is modeled using a 2-parameter logistic function applied to the observed proportions of children who have acquired a given word at specific ages:

$$y = \frac{1}{1 + e^{-k(x - x_0)}} \, .$$

  Here y is the proportion of children who have acquired the given word, x is their age in months, and the fitted parameters are the growth rate k and the median AoA x_0. The curve fitting is performed using the `curve_fit` function from the SciPy library. See Figure 1 for an example of the logistic models for these acquisition curves.

  We separate and model the Understands and Produces data independently to capture the expected time lag between the two measures. This results in two distinct pairs of parameters (k_p and $x_{0,p}$ for Produces and k_u, and $x_{0,u}$ for Understands) for each word, which we then incorporate into our design matrix.
- **Category**: The Wordbank data includes classification of words into broad conceptual categories, such as `animals`, `clothing`, `connecting_words`, and `quantifiers`.
- **Levenshtein edit distance fraction**: One expects that a word in L2 is easier to learn if it is similar to the corresponding word in L1. To measure this effect, we compute the



(A) Understands (WG)

(B) Produces (WG and WS)

FIGURE 1. Logistic models of the acquisition curves for the word "bus" for the Understands and Produces measures.

edit distance (also known as the *Levenshtein distance*) between the L1 an L2 words. This counts how many letter-by-letter changes are required to transform the L2 word into the L1 word. We compute edit distances using the Python package editdistance. To normalize for the length of the word, we divide by the maximum of the lengths of the L1 and L2 words.

- **Word frequencies**: This feature quantifies how frequently a given word occurs in everyday language use. It is obtained from large-scale language corpora (e.g., Common Crawl, Wikipedia, news sources) using the wordfreq library. Word frequencies were computed for both source and target languages.
- **Token frequency**: This feature represents raw count of how often the exact word token appeared within the specific Duolingo SLAM dataset.
- **Token morphology**: This feature describes morphological features of the word token, such as gender and number for nouns and adjectives, and mood, tense, number, and person for verbs.
- **Token part of speech**: This feature records the part of speech of the word token, including determiners (e.g., articles), nouns, verbs, adjectives, and prepositions.
- **Token dependency edges**: A labeled, directed link between two words in a dependency parse-connecting a head (governing token) to a dependent with a relation label (e.g., nsubj, obj, amod).
- **Response time**: The time (in seconds) it took the learner to submit a response.
- **Days in course**: Number of days since the learner started learning the language on Duolingo.
- **Omitted features**: The features client, countries and session were excluded because we were interested in modeling token-specific properties and the learning context, so these features were not relevant. User ID was not included because we wanted the model to generalize across users. Finally, token was not included because it did not account for much variance once token-level properties were added to the model.

4. RESULTS

4.1. **Inferential results.** The SLAM dataset comes pre-split in TRAIN/DEV/TEST portions. We kept these splits while investigating token-level error (token wrong) as a function of token-level properties (e.g., part of speech, morphology, dependency edges, conceptual category, age of

acquisition), and learner context (e.g., days in course, response latency). We conducted exploratory data analysis on TRAIN to characterize coverage, missingness, and the relationship between our feature set and token-level errors. Figure 2 shows a positive association between AoA and error rate ($\rho \approx 0.20$), indicating that later-acquired items tend to elicit more errors. Error rate is lower for high-exposure words ($\rho(n, \text{error rate}) \approx -0.32$), consistent with practice and frequency effects. Furthermore, AoA is modestly anticorrelated with exposure ($\rho \approx -0.17$). Correlations with growth rate are weak ($|\rho| \leq 0.11$), suggesting limited collinearity among predictors.

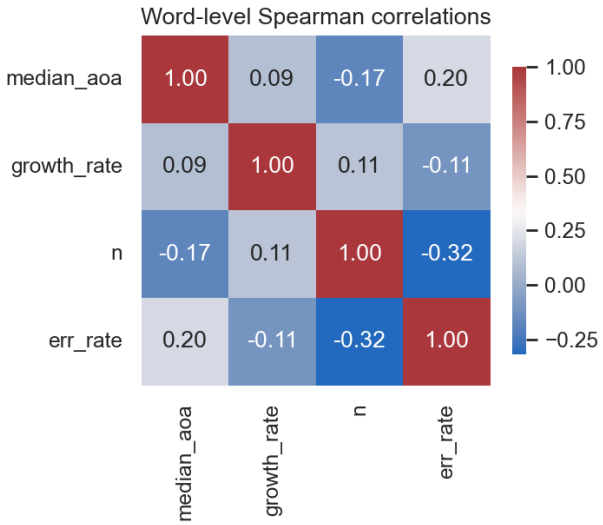


FIGURE 2. Word-level Spearman correlations among median age-of-acquisition (median AoA), growth rate, attempt count (n), and error rate. Values are Spearman ρ (monotonic associations). AoA correlates positively with error rate ($\rho = 0.20$), while error rate decreases with exposure ($|\rho| = -0.32$). AoA and exposure are modestly anticorrelated ($|\rho| = -0.17$), and growth rate shows only weak links to the other variables. These patterns motivate weighting or controlling for n and including AoA as a primary predictor, with minimal concern about multicollinearity.

Similarly, Figure 3 shows that later-acquired words tend to have higher error rates. The LOWESS curve rises steadily through the mid-20s to low-30s in AoA and then flattens, suggesting diminishing increases at the upper end. Variation is larger in the mid-AoA range, and high-exposure words (large bubbles) cluster at lower error rates, consistent with practice/frequency effects.

We aggregated the data to the word level ($\geq$ 100 attempts) to analyze error rate vs median aoa with part of speech (POS controls). As shown in Figure 4, we observed a modest positive association between AoA and error rate, a weak positive association between AoA and growth rate, and a strongly right-skewed, heavy-tailed distribution of attempt counts (n), with most words low-frequency and a few very high-frequency outliers. Next, we fit a frequency-weighted binomial GLM with logit link at the word level to model error rate as a function of median AoA, part-of-speech (POS), and their interactions. The overall AoA slope was positive and statistically reliable ($\beta = 0.0158$, $z = 7.35$, $p < 0.001$), which implies that for the reference POS each one-unit increase in AoA is associated with about a 1.6% increase in the odds of error (OR ≈ 1.016). Critically, POS moderates the AoA$\times$POS interactions indicate steeper increases for several categories—for example, DET (β total=0.1396 $\rightarrow$ OR ≈ 1.15, +15% odds per AoA unit, $p < 0.001$), PRON (.0813 $\rightarrow$ OR ≈ 1.085, +8.5%, $p < 0.001$), AUX (0.0717 $\rightarrow$ OR ≈ 1.074, +7.4%, $p < 0.001$),

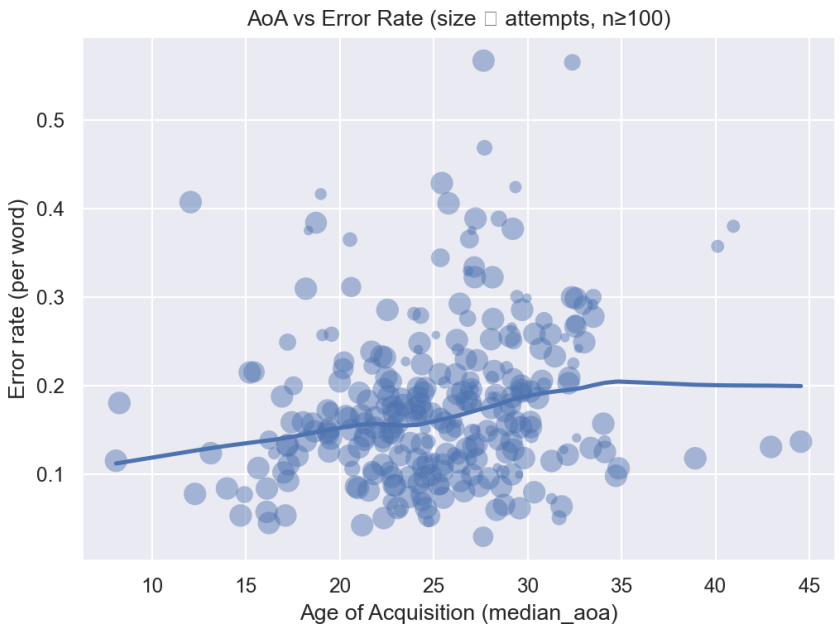


FIGURE 3. Age of Acquisition (AoA) vs. word-level error rate (point size $\propto$ number of attempts; $n \geq 100$). Each point represents a word and bubble size reflects exposure. The LOWESS smoother indicates a positive, mostly monotonic association between AoA and error rate, with a plateau beyond $\sim$ 33–35 AoA. High-exposure items generally exhibit lower error rates, and dispersion is greatest for mid-AoA words. This pattern aligns with the word-level Spearman correlation (AoA–error $\rho \approx 0.20$) and motivates including AoA in subsequent models.

ADV ($0.0629 \rightarrow$ OR ≈ 1.065, $+6.5\%$, $p < 0.001$), and CONJ ($0.2143 \rightarrow$ OR ≈ 1.239, $+23.9\%$, $p < 0.001$).

The inference analyses indicated a positive association between AoA and error rate, which is largely consistent across major POS categories. Guided by these patterns, we fit two model families: a regularized logistic-regression and a non-linear HistGradientBoosting classifier, which naturally captures non-linear effects and POS/dep interactions without manual feature engineering.

4.2. **Modeling.**

4.2.1. *Summary of past models.* The Duolingo baseline consist of a regularized logistic regression estimated separately for each language track achieving a F_1-score of 0.190, 0.175 and 0.281 in English, Spanish and French respectively. Teams in the original competition employed a diverse range of models. Past SLAM approaches explored how learners acquire a new language using different algorithms and data features. Results from the SLAM challenge showed that advanced models like RNNs and GBDTs were more effective at tracking learner progress over time. In contrast, simpler models such as logistic regression focused on psychological features but were limited by the dataset's short duration and uniform learner background.

4.2.2. *Baseline SLAM model.* The model provided by Duolingo was a simple, L^2-regularized logistic regression trained with Stochastic Gradient Descent (SGD). It served as a benchmark for all participating teams to measure their own, more complex models [SBG+18].

FIGURE 4. Word-level pairplot of median age-of-acquisition (AoA), growth rate, attempt count (n), and error rate, colored by part-of-speech (POS). Histograms on the diagonal summarize marginal distributions; off-diagonal panels show bivariate relationships. AoA exhibits a positive relationship with error rate and a mild positive trend with growth rate. Attempt counts are heavy-tailed, with most words low-frequency and a small number of high-frequency outliers. Overlap across POS indicates that these patterns are largely consistent across categories, though some POS (e.g., DET/PRON) show clusters with higher error rates.

4.2.3. *Logistic Regression.* We built a token-level error classifier using regularized logistic regression. The goal was to predict whether a learner would mark a given token wrong. We split the data into TRAIN, DEV, and TEST: the model is fit on TRAIN, the decision threshold is chosen on DEV, and TEST is held out and evaluated once at the end to avoid leakage. Baseline and Enhanced models were used for comparison. The **Baseline model** used as categorical features part-of-speech, dependency and morphology tags, dependency edges, and a set of numeric features (response time, days in course). The **Enhanced model** added age of acquisition, semantic category, word frequency and edit distance. Edit or Levenhstein distance is calculated as the minimum number of edits (operations) to convert the target word into the semantically closest word in the native language. Comparing these two models allowed us to assess whether the age at which a word is acquired in the native language, along with word frequency, L1-L2 relatedness and conceptual information improved ranking beyond the baseline. Our implementation is based on the code associated to [RPH+18], adapted to include our new features.

All preprocessing was done using a single scikit-learn pipeline that kept TRAIN/DEV/TEST datasets separate. Numeric features were median-imputed, time and days were log-transformed (log1p) and all numeric features were standardized. Categorical features were imputed (most-frequent) and one-hot encoded with unknowns ignored, so the model could handle categories not seen during training. Because the dataset is imbalanced, the classifier used class-balanced weights.

Modeling used logistic regression with the SAGA solver and L^2 regularization. We conducted cross-validation on TRAIN only (StratifiedKFold, scored by Average Precision) to pick among a small grid of regularization strengths, then identified the best estimator on full TRAIN. After training, we selected a threshold on DEV that maximizes recall subject to a minimum precision target (default 0.30). Hyper-parameters (penalty, C) were tuned by CV on TRAIN only (best C=0.2, L2 for both), and the operating threshold on DEV was chosen to maximize recall subject to precision ≥ 0.30.

On DEV, the Enhanced model outperformed the Baseline on ranking metrics (AUC 0.618 vs 0.595; AP 0.226 vs 0.212) and at the selected operating point increased recall (6.4% vs 4.6%) at similar precision (0.30). Held-out TEST confirmed these gains: AUC 0.616 vs 0.594 and AP 0.234 vs 0.221. At the DEV-selected thresholds applied to TEST, recall improved from 5.6% $\rightarrow$ 7.1% ($\approx$+1.5 pp, +27% relative) while precision stayed near the target (0.320 $\rightarrow$ 0.313). Accuracy was similar ($\approx$ 0.83 Baseline vs 0.83 Enhanced). Confusion matrices reflect the intended trade-off: on TEST, true positives rose (571 $\rightarrow$ 721) with a controlled increase in false positives (1215 $\rightarrow$ 1579) while meeting the precision floor. Overall, the Enhanced feature set produces modest but consistent improvements in both ranking quality and retrieval at the target precision.

TABLE 1. Logistic regression results. Hyperparameters tuned on TRAIN (best: C=0.2, L2). Threshold selected on DEV to meet precision ≥ 0.30 and then fixed for TEST (Baseline τ=0.7336, Enhanced τ=0.7281).

Metric	**LG-Baseline**	**LG-Enhanced**
AUC (%)	0.59	0.62
AP (%)	0.22	0.23
Accuracy (%)	83.1	82.8
Precision (%)	32.0	31.3
Recall (%)	5.6	7.1
F1-Score (%)	9.5	11.5
Overall Rank	2	1

4.2.4. *Histogram-Based Gradient Boosting.* We decided to use a Histogram-Based Gradient Boosting (HBG) model for comparison with the results of the logistic regression (LG) model for three main reasons. First, the novel engineered features included in the enhanced model (age of acquisition and token frequency) are non-linear, that is, the error probability does not change linearly. Boosted trees fit non-linearity automatically and are therefore better fitted for this type of data. Second, HBGs learn interaction terms implicitly (e.g., POS $\times$ dependency label $\times$ lexical frequency), whereas these interaction terms would need to be explicitly added to the LG model. Finally, HBG models work well with mixed data types and high cardinality. Specifically, using OrdinalEncoder for categorical variables avoids large one-hot matrices, which is particularly useful for token-specific features, such as part of speech (POS) or token morphology.

We trained a non-linear token-error classifier using scikit-learn's HistGradientBoostingClassifier. The objective was to predict whether a learner would mark a token as wrong. Similar to the LR model, we split the dataset into nto TRAIN, DEV, and TEST parquet files. The model was fit on

TRAIN, the decision threshold was chosen on DEV to satisfy a minimum precision, and TEST was reserved for a final, single evaluation.

The Baseline model and Enhanced models were as described in the LG model. Categorical columns were imputed (most frequent) and Ordinal-encoded with a safe unknown code to ensure the feature space stays compact and to avoid huge, sparse matrices. Numeric columns were median-imputed, with optional log1p on timing variables (time, days), then passed to the model. All preprocessing steps were conducted inside a single pipeline, which ensured no leakage across splits.

We used HGB with class imbalance handled via class weight="balanced". The default configuration ensured fast, well-regularized trees: learning rate ≈ 0.08, up to 31 leaf nodes per tree, min samples leaf=50, and about 300 boosting iterations (no internal early stopping because DEV serves as the external validation set). As noted, since trees naturally capture non-linearities and interactions, this setup models non-linear effects and cross-feature interactions that a linear model would miss.

We compared the *Baseline* and an *Enhanced* feature sets with fixed hyperparameters. Thresholds were chosen on DEV to satisfy a precision floor of ≈ 0.30 (Baseline τ=0.6377, Enhanced τ=0.6014) and then held fixed for TEST. On DEV, the Enhanced model improved ranking quality (AUC = 0.6783; AP = 0.3165) and, at the operating point, increased recall while maintaining precision ≈ 0.30. Held-out TEST confirmed these gains (Table 2): AUC rose from 0.6345 (Base) to 0.6791 (Enhanced), and AP from 0.2733 to 0.3234. At the fixed thresholds, recall nearly doubled ($0.194 \rightarrow 0.396$) with similar precision ($0.311 \rightarrow 0.308$), increasing true positives from 1,982 to 4,042 (with false positives rising from 4,384 to 9,089). As expected under class imbalance, accuracy decreased ($0.804 \rightarrow 0.763$) when retrieving more positives, but the higher AUC and AP indicate a better overall ranking and retrieval of difficult items at the target precision.

TABLE 2. HistGradientBoosting results. Threshold selected on DEV to meet precision ≥ 0.30 and then fixed for TEST (Base τ=0.6377, Enhanced τ=0.6014).

Metric	**HGB-Base**	**HGB-Enhanced**
AUC (%)	0.63	0.68
AP (%)	0.27	0.32
Accuracy (%)	80.4	76.3
Precision (%)	31.1	30.8
Recall (%)	19.4	39.6
F1-Score (%)	23.9	34.6
Overall Rank	2	1

Model comparison (LR vs. HGB). Across both model families, adding AoA, conceptual category and frequency features (*Enhanced*) improves ranking and retrieval, but the effect size differs markedly. With *logistic regression* (LR), TEST AUC/AP rise modestly (AUC $0.594 \rightarrow 0.616$, AP $0.221 \rightarrow 0.234$) and recall at the DEV–selected precision floor (≈ 0.30) increases from $5.6\% \rightarrow 7.1\%$, while precision remains $\sim 31\%$ and accuracy stays near 83% (see Table 1). In contrast, *HistGradientBoosting* (HGB) yields larger gains: TEST AUC $0.635 \rightarrow 0.679$, AP $0.273 \rightarrow 0.323$, and recall $19.4\% \rightarrow 39.6\%$ (roughly $\times 2$) at similar precision ($\sim 31\%$), with the expected drop in accuracy ($80.4\% \rightarrow 76.3\%$) due to retrieving more positives under class imbalance (Table 2). Overall, **HGB–Enhanced** delivers the strongest retrieval under the precision constraint, whereas **LR–Enhanced** provides smaller but consistent improvements and remains valuable for interpretability (coefficients, effect signs) and as a transparent baseline.

5. Conclusion

This project investigated the impact of native language (L1) age of acquisition (AoA) on the ability to learn corresponding words in a second language (L2). Our primary contribution was the integration of normative developmental data from the Wordbank database, which offers a more accurate measure of AoA than the adult-recollection data used in previous studies.

Future Directions and Further Improvements. Building upon the current findings, several avenues for future work could extend this research.

- **Expanding Cross-Linguistic Scope**. The current analysis focused on L1 data from the United States and Mexico. Future work could incorporate Wordbank data from a broader range of countries and language families to test the generalizability of the L1 AoA effect across different regions.
- **Enhancing Lexical Matching Precision**. To improve the validity of the L1-L2 transfer analysis, Natural Language Processing (NLP) techniques, such as lemmatization, could be employed for more robust and nuanced matching between word tokens in the Wordbank and the SLAM L2 learning datasets.
- **Integrating Larger Corpora**. The study could be extended by incorporating larger, independently collected corpora of vocabulary items that include reliable age of acquisition data, which would increase the size of the lexical set analyzed for L1 transfer effects.

References

[FBYM17] Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694, 2017. doi:10.1017/S0305000916000209. 1

[FBYM21] Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. *Variability and consistency in early language learning: The Wordbank project*. MIT Press, 2021. 1

[KSGB12] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990, 2012. 1

[RPH+18] Alexander Rich, Pamela Osborn Popp, David Halpern, Anselm Rothe, and Todd Gureckis. Modeling second-language learning from a psychological perspective. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 223–230, 2018. 1, 6

[SBG+18] Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 56–65, 2018. 2, 5

[Set18] Burr Settles. Data for the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM), 2018. doi:10.7910/DVN/8SWHNO. 1, 2

Benard Haugen,
Email address: benardhaugen@gmail.com

Vikram Jambulapati, UC San Diego
Email address: vikjam@ucsd.edu

Manjeet Kaur,
Email address: kaurranamanjeet@gmail.com

Sara Sanchez-Alonso, Yale University
Email address: sara.sanchez.alonso@yale.edu

Sam Schiavone,
Email address: sam.schiavone@gmail.com