LingPredict

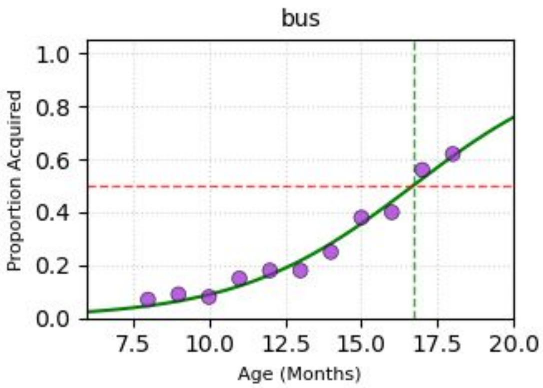
Developmental Norms and Second Language Acquisition

Q: Does L1 Age of Acquisition (AoA) influence L2 vocabulary learning?

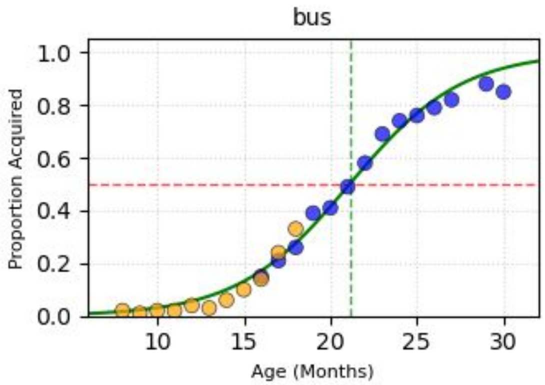
Data sources

- **Wordbank (L1)**
  - Open data repository on language acquisition in children
  - Data compiled from MacArthur-Bates Communicative Development Inventory (CDI) surveys completed by parents of young children.
  - Large-scale repository (105,290 surveys across 42 languages).
  - **CDI Forms**:
    - Words & Gestures (`WG`): comprehension/early vocabulary (8-18 months).
    - Words & Sentences (`WS`): production/grammar (16-30 months).
  - **Measures**:
    - Understands
    - Produces

Understands (WG)



Produces (WG + WS)



Data sources

- **Duolingo SLAM** (Second Language Acquisition Modeling) **2018 (L2)**
    - Data from Duolingo learners over their first 30 days of study.
    - Over 6,000 students learning English, Spanish, and French.
    - Over 2 million tokens from three exercise formats (reverse_translate, reverse_tap, listen).
    - Data features:
        - Student History**:** Anonymized **`user`** ID, **`days`** since starting
        - L2 word **`token`**, Part of Speech, and Morphological Features
        - **`time`** taken to complete the exercise
        - **`token`**-level correctness: 0 = correct; 1 = incorrect

Our contribution:
Age of acquisition (produce and understand) for native Spanish speakers

```
# prompt:Yo tengo una habitación.
# user:G86T0ut3  countries:MX  days:19.692  client:ios  session:practice  format:reverse_translate  time:13
ut0q00+s0301  I          PRON   Case=Nom|Number=Sing|Person=1|PronType=Prs|fPOS=PRON++PRP  nsubj 0 28.5  16.9
ut0q00+s0302  have       VERB   Mood=Ind|Tense=Pres|VerbForm=Fin|fPOS=VERB++VBP            ROOT  1 30.8  19.5
ut0q00+s0303  a          DET    Definite=Ind|PronType=Art|fPOS=DET++DT                     det   0 28.5  19.8
ut0q00+s0304  room       NOUN   Number=Sing|fPOS=NOUN++NN                                  dobj  1 25.8  14.7
```

Feature Engineering

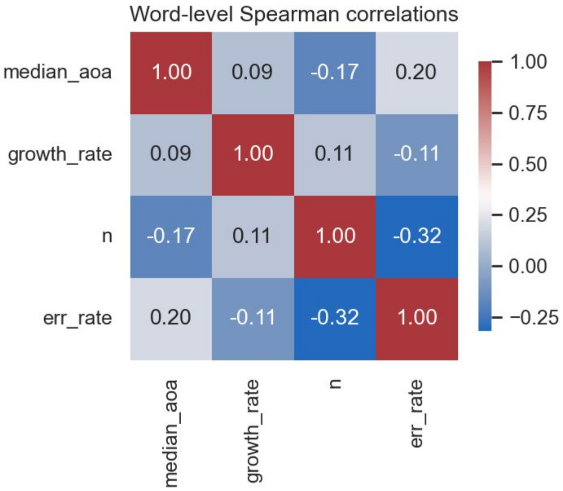
- ❑ **Age of acquisition:** fit a 2-parameter logistic curve to each word's acquisition trajectory: Age (in months) vs. Proportion of Children Knowing Word. Parameters describe **median_aoa** and **growth_rate**

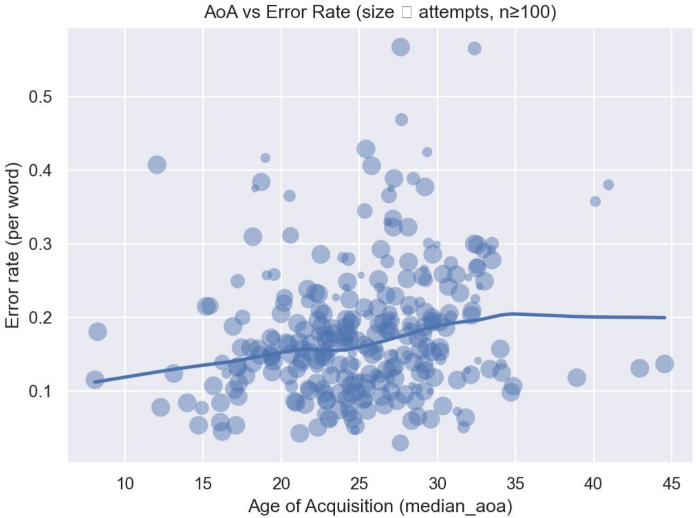
- ❑ **Category (from Wordbank):** 'sounds', 'animals', 'toys', 'food_drink',...., 'pronouns', 'question_words', 'quantifiers', 'helping_verbs'

- ❑ **Word frequency:** 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 serving as a proxy for general exposure.

- ❑ **Token frequency:** raw count of how often the exact word token appeared within the specific Duolingo SLAM dataset

Inferential Results



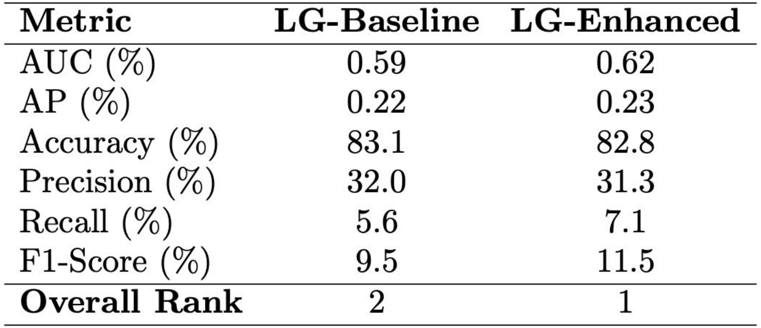


- Later-acquired words tend to elicit more errors (r =0.20).
- Error rate decreases with exposure, which is consistent with practice/frequency effects (n = 0.32).
- AoA and growth rate show limited collinearity (p =0.11).

- Positive association between AoA and error rate, with a plateau beyond ~33–35 months.
- These findings motivate including AoA in subsequent models.

Logistic Regression Results

- Logistic regression with SAGA solver and L2 regularization.
- Cross-validation conducted on TRAIN only to select regularization strengths.
- After training, we selected a threshold on DEV that maximizes recall subject to a minimum precision target (0.30).
- Hyper-parameters (penalty, C) were tuned by CV on TRAIN only (best C=0.2, L2 for both).



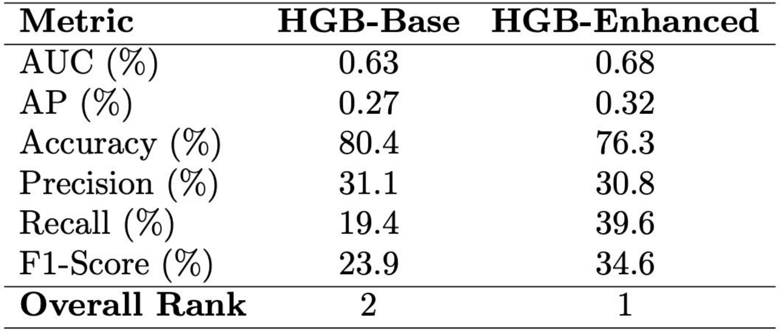
Metric	**LG-Baseline**	**LG-Enhanced**
AUC (%)	0.59	0.62
AP (%)	0.22	0.23
Accuracy (%)	83.1	82.8
Precision (%)	32.0	31.3
Recall (%)	5.6	7.1
F1-Score (%)	9.5	11.5
Overall Rank	2	1

Enhanced feature set produced modest but consistent improvements:

On TEST, Enhanced model outperformed the Baseline on ranking metrics (AUC, AP) and increased recall at the selected precision target.

Histogram-Based Gradient Boosting

- Class imbalance in token error was handled via class weight = “balanced”.
- Parameters: Learning rate 0.08, up to 31 leaf nodes per tree, min samples per leaf=50, 300 boosting iterations.
- In contrast to LR model, this setup models non-linear effects and cross-feature interactions implicitly.
- As in the LR model, thresholds were chosen on DEV to satisfy a precision floor of ~0.30.



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

Enhanced model improves raking quality (AUC, AP) and increases recall:

- On TEST: AUC increased from 0.63 to 0.68 and AP from 0.27 to 0.32. Recall nearly doubled from 0.19 to 0.40.
- Accuracy decreased as expected under class imbalance when retrieving more positives.

Conclusion/Next Steps

- Model comparison (LR vs. HGB):
  - Adding AoA, conceptual category and frequency features **improved ranking and retrieval**.
  - LR: AUC/AP rose modestly (AUC 0.59 to 0.62, AP 0.22 to 0.23) and recall increased from 5.6% to 7.1%
  - **Larger gains in HGB**: AUC/AP increased (0.64 to 0.68, AP: 0.27 to 0.32) and recalled doubled from 19.4% to 39.6%.
  - HGB provides **strongest retrieval**, whereas LR provides smaller but consistent improvements. LR model is still valuable for **interpretability** and as a transparent baseline.
- Integrating Wordbank developmental norms with Duolingo SLAM data shows that early-acquired L1 words are generally easier to learn in L2.
- Our gradient-boosting model, using features like AoA, word frequency, and Levenshtein distance, modestly improved predictive accuracy over the SLAM baseline.
- Future work should refine word matching, add longitudinal learner data, and expand developmental datasets to strengthen cross-linguistic insights.