Investigating early second language word learning accuracy in a large-scale dataset

Elise W.M. Hopman*,a (hopman@wisc.edu), Yiwen Wangb, Bill Thompsonc, Gary Lupyanb & Joseph L. Austerweilb

Why are some L2 words harder to learn?
Classroom/experimental studies identified several predictors1,2,3, but...
- predictors often dichotomized
- predictors usually studied in isolation
- few predictors studied

What is Duolingo?
Translation “Algorithm”
1. machine translations
2. bilinguals hand-check

Do translations mean the same?
1. find the 40 closest neighbors to a translation pair
2. find overlapping neighbors between the two languages

<table>
<thead>
<tr>
<th>Rank</th>
<th>Brother</th>
<th>Similarity</th>
<th>Hermana</th>
<th>Similarity</th>
<th>Boat</th>
<th>Similarity</th>
<th>Barco</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>uncle</td>
<td>0.82</td>
<td>hermana</td>
<td>0.79</td>
<td>ship</td>
<td>0.56</td>
<td>avión</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>father</td>
<td>0.81</td>
<td>padre</td>
<td>0.78</td>
<td>airplane</td>
<td>0.46</td>
<td>viaje</td>
<td>0.49</td>
</tr>
<tr>
<td>3</td>
<td>son</td>
<td>0.79</td>
<td>tío</td>
<td>0.78</td>
<td>navy</td>
<td>0.43</td>
<td>puerto</td>
<td>0.42</td>
</tr>
<tr>
<td>4</td>
<td>grandfather</td>
<td>0.75</td>
<td>hijo</td>
<td>0.77</td>
<td>swim</td>
<td>0.43</td>
<td>tren</td>
<td>0.42</td>
</tr>
<tr>
<td>5</td>
<td>daughter</td>
<td>0.66</td>
<td>abuelo</td>
<td>0.76</td>
<td>car</td>
<td>0.43</td>
<td>armada</td>
<td>0.42</td>
</tr>
<tr>
<td>6</td>
<td>friend</td>
<td>0.63</td>
<td>hija</td>
<td>0.71</td>
<td>fish</td>
<td>0.42</td>
<td>capitán</td>
<td>0.42</td>
</tr>
<tr>
<td>7</td>
<td>sister</td>
<td>0.62</td>
<td>esposa</td>
<td>0.63</td>
<td>beach</td>
<td>0.42</td>
<td>cochec</td>
<td>0.40</td>
</tr>
<tr>
<td>8</td>
<td>mother</td>
<td>0.62</td>
<td>quien</td>
<td>0.62</td>
<td>river</td>
<td>0.40</td>
<td>isla</td>
<td>0.39</td>
</tr>
<tr>
<td>9</td>
<td>wife</td>
<td>0.62</td>
<td>amigo</td>
<td>0.61</td>
<td>ton</td>
<td>0.38</td>
<td>vehículo</td>
<td>0.39</td>
</tr>
<tr>
<td>10</td>
<td>aunt</td>
<td>0.61</td>
<td>marido</td>
<td>0.61</td>
<td>wooden</td>
<td>0.38</td>
<td>hotel</td>
<td>0.38</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>40</td>
<td>family</td>
<td>0.36</td>
<td>abogado</td>
<td>0.36</td>
<td>port</td>
<td>0.36</td>
<td>cuando</td>
<td>0.32</td>
</tr>
</tbody>
</table>

3. Meaning Similiarity6: overlapping neighbors’ rank correlation

References
De Groot, A.M.B., & Kayne, R. (2003). What is hard to learn is easy to forget: The role of word concreteness, cognitive status, and word frequency in foreign language vocabulary learning and forgetting. Language Learning, 53(1).
Department of Psychology, University of Wisconsin-Madison
Institute for Applied Computational Science, Harvard University
Department of Psychology, Princeton University
Research was supported by a Merit and Equity research award from the Psychology Department at the University of Wisconsin - Madison

What is in the Duolingo dataset?
Example raw user-word observations4

<table>
<thead>
<tr>
<th>Interface Language</th>
<th>User ID</th>
<th>Word</th>
<th>Experience</th>
<th>User Experience</th>
<th>Times correct</th>
<th>Proportion correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish u:0Fa</td>
<td>Blue</td>
<td>24</td>
<td>81</td>
<td>16</td>
<td>0.66</td>
<td></td>
</tr>
<tr>
<td>Spanish u:0Fa</td>
<td>Spider</td>
<td>7</td>
<td>81</td>
<td>7</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Spanish u:0Fa</td>
<td>Eat</td>
<td>20</td>
<td>81</td>
<td>18</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>Spanish u:0Fa</td>
<td>Until</td>
<td>30</td>
<td>81</td>
<td>24</td>
<td>0.8</td>
<td></td>
</tr>
</tbody>
</table>

Overall dataset characteristics

Predicting word learning accuracy
Linear mixed effects regression model results:

Duolingo User Experience
Nr. of wordnet synsets (L1)
Semantic Density Diff. (L1-L2)
Normed Levenshtein Distance
Frequency Diff. (L1-L2)
Concreteness
Meaning Similarity
Frequency (L1)
Semantic Density
Diff. in Nr. of synsets (L1-L2)
Duolingo Word Experience

Challenges and open questions
- Duolingo’s algorithm oversamples words people find hard
- Given biased statistics in this dataset, what are appropriate statistical models?
- Combinatorial explosion of potential interactions
- How do these analyses inform cognitive theories of word learning?
- What can asymmetries in learnability tell us?