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

Elise W.M. Hopman[*,a] (hopman@wisc.edu), Yiwen Wang[b],
Bill Thompson[c], Gary Lupyan[a] & Joseph L. Austerweil[a]

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Harvard John A. Paulson School of Engineering and Applied Sciences

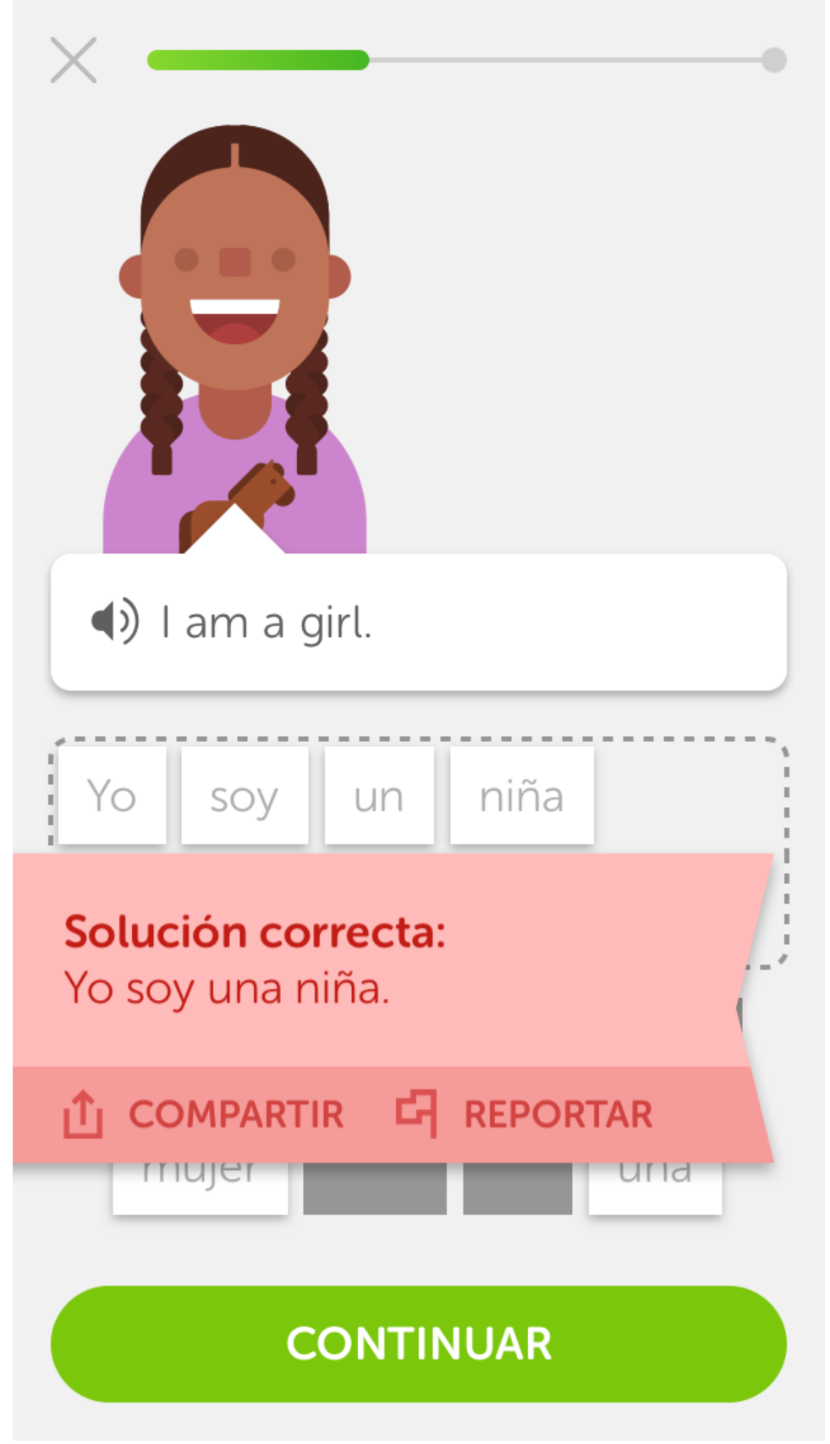
PRINCETON UNIVERSITY

Why are some L2 words harder to learn?

Classroom/experimental studies identified several predictors[1,2,3], but…

- predictors often dichotomized
- predictors usually studies in isolation
- few predictors studied

What is Duolingo?



I am a girl.

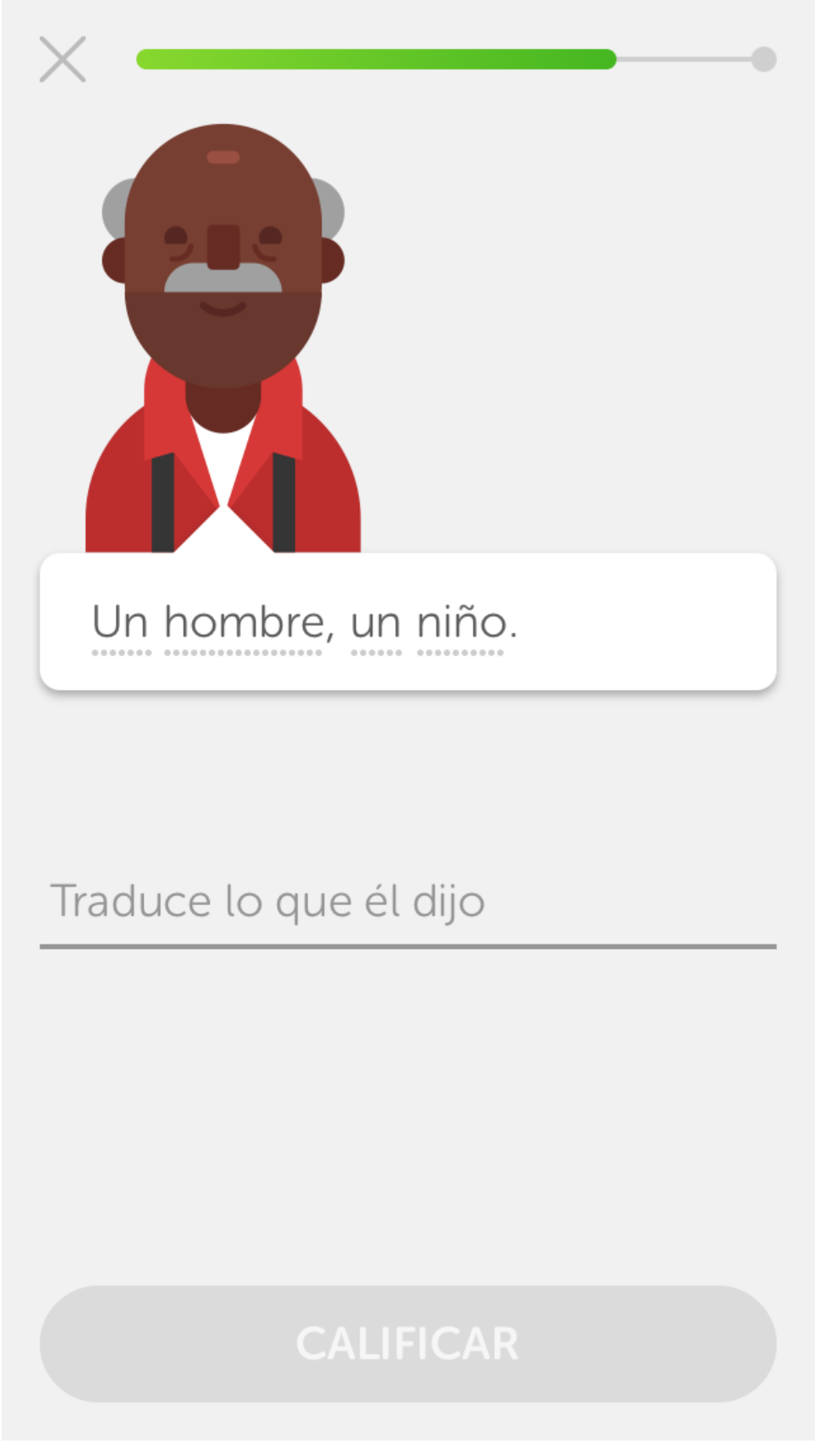
Yo | soy | un | niña

Solución correcta:
Yo soy una niña.

COMPARTIR REPORTAR

mujer | una

CONTINUAR



Un hombre, un niño.

Traduce lo que él dijo

CALIFICAR

Translation "Algorithm"

Google Translate

1. machine translations

2. bilinguals hand-check

What is in the Duolingo dataset?

Example raw user-word observations[4]

Interface Language	UserID	Word	Word Experience	*User Experience*	Times correct	*Proportion correct*
Spanish	u:0Fa	Blue	24	*81*	16	*0.66*
Spanish	u:0Fa	Spider	7	*81*	7	*1*
Spanish	u:0Fa	Eat	20	*81*	18	*0.9*
Spanish	u:0Fa	Until	30	*81*	24	*0.8*

Overall dataset characteristics

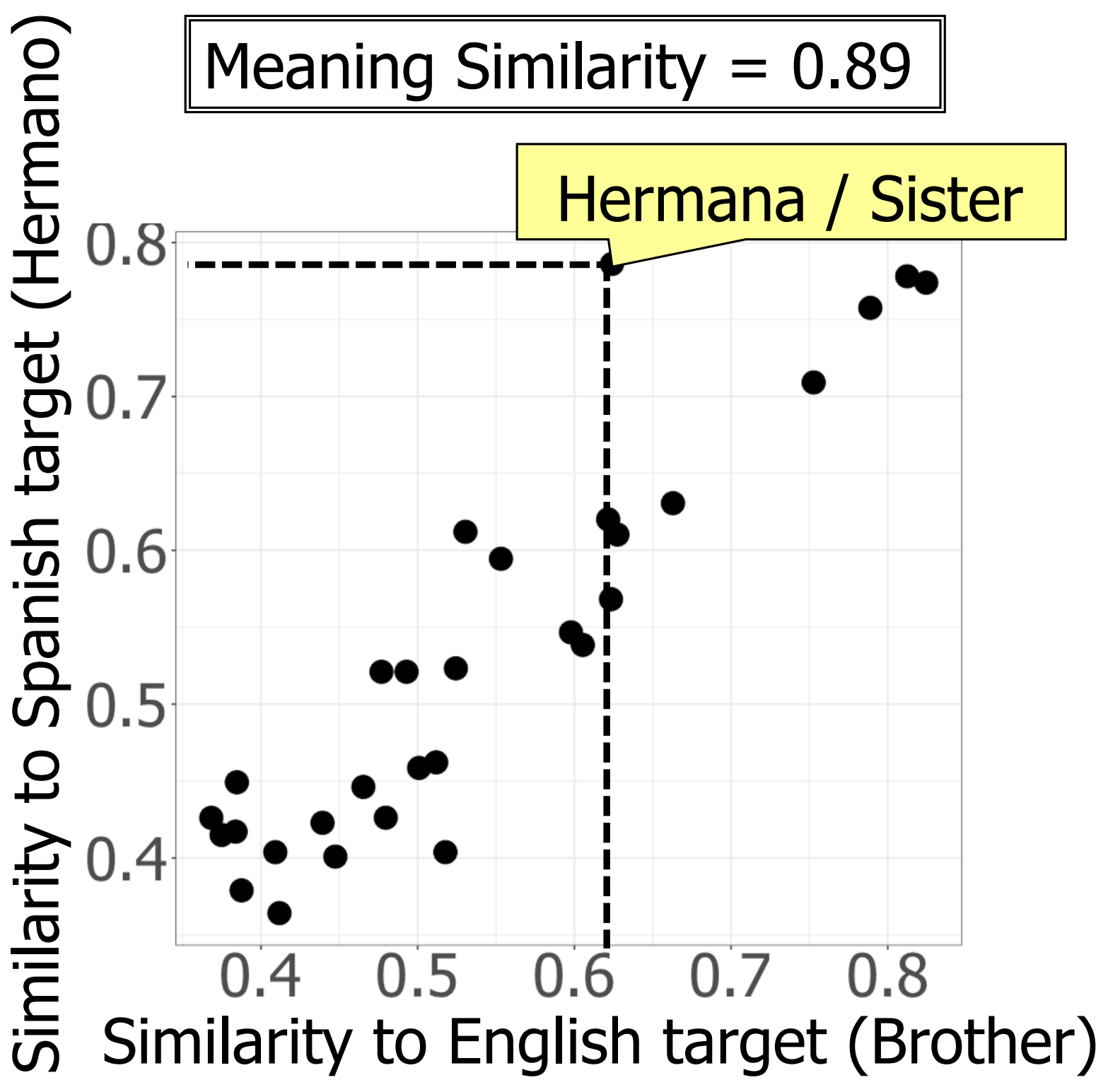
Interface Language	Language Learned	Number of Users	Number of Words	Range of Word Experience	Range of User Experience	Number of observations
Spanish	English	28,107	1,411	3 – 22,336	41 – 392,683	1,197,890
English	Spanish	27,248	1,737	3 – 4,737	41 – 75,664	1,182,191
Portuguese	English	7,713	1,398	3 – 7,991	41 – 40,052	312,088
English	Portuguese	2,395	1,517	3 – 1,540	41 – 13,971	99,633
Italian	English	2,959	1,411	3 – 1,1577	41 – 32,304	152,523
English	Italian	5,522	1,330	3 – 1,104	41 – 17,802	222,925

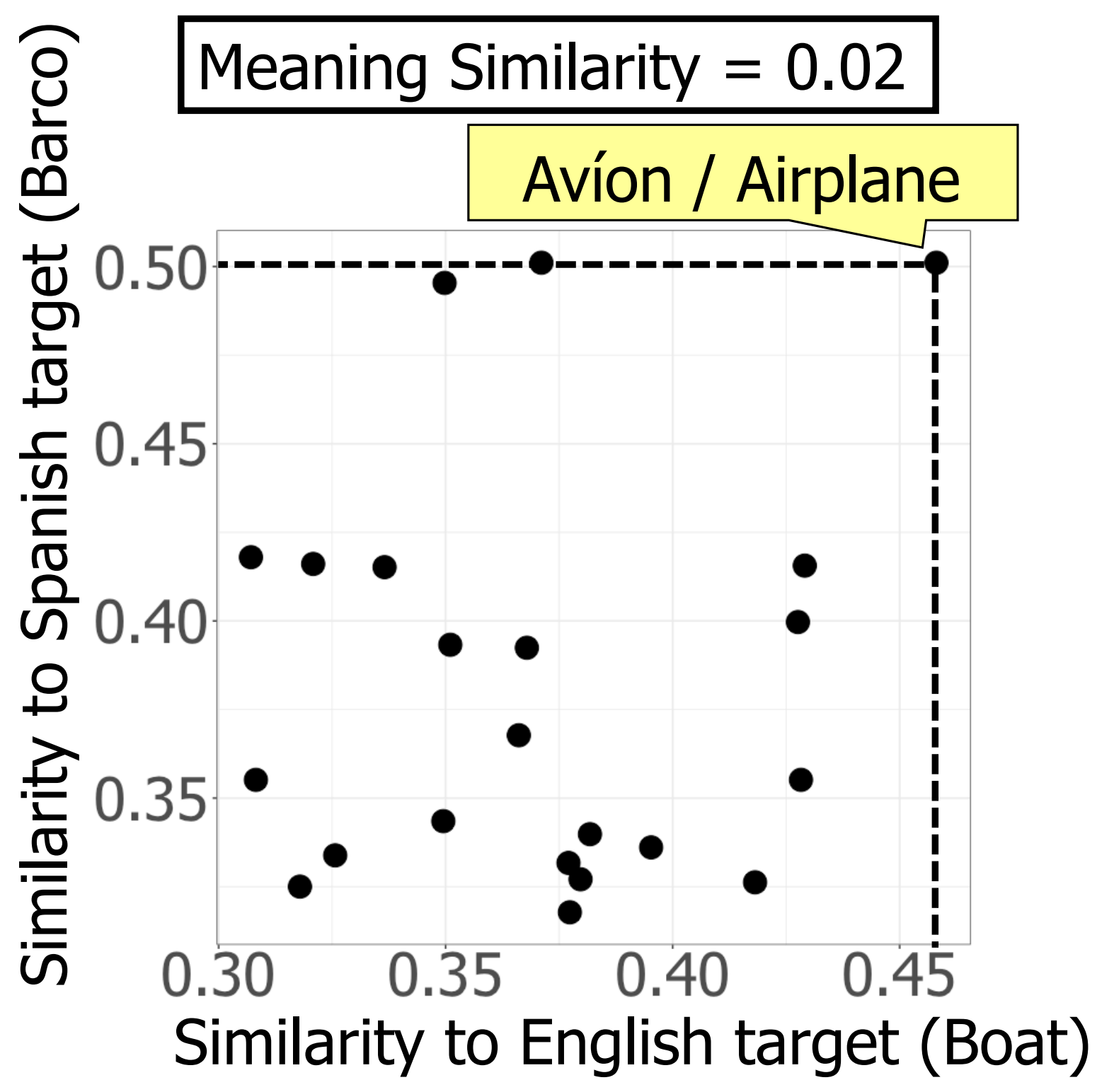
Do translations mean the same?

1. find the 40 closest neighbors[5] to a translation pair
2. find overlapping neighbors between the two languages

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

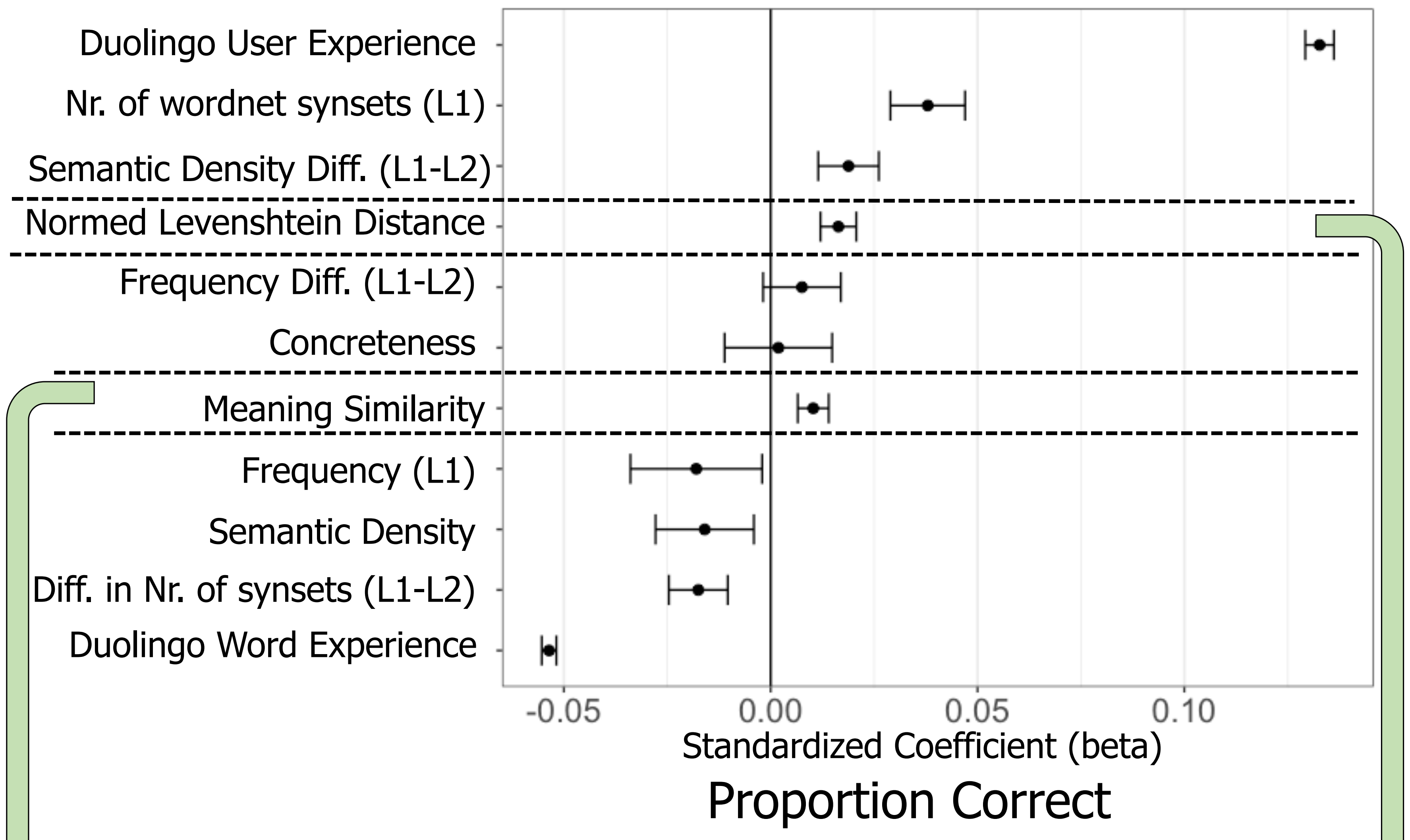
3. Meaning Similarity[6]: overlapping neighbors' rank correlation





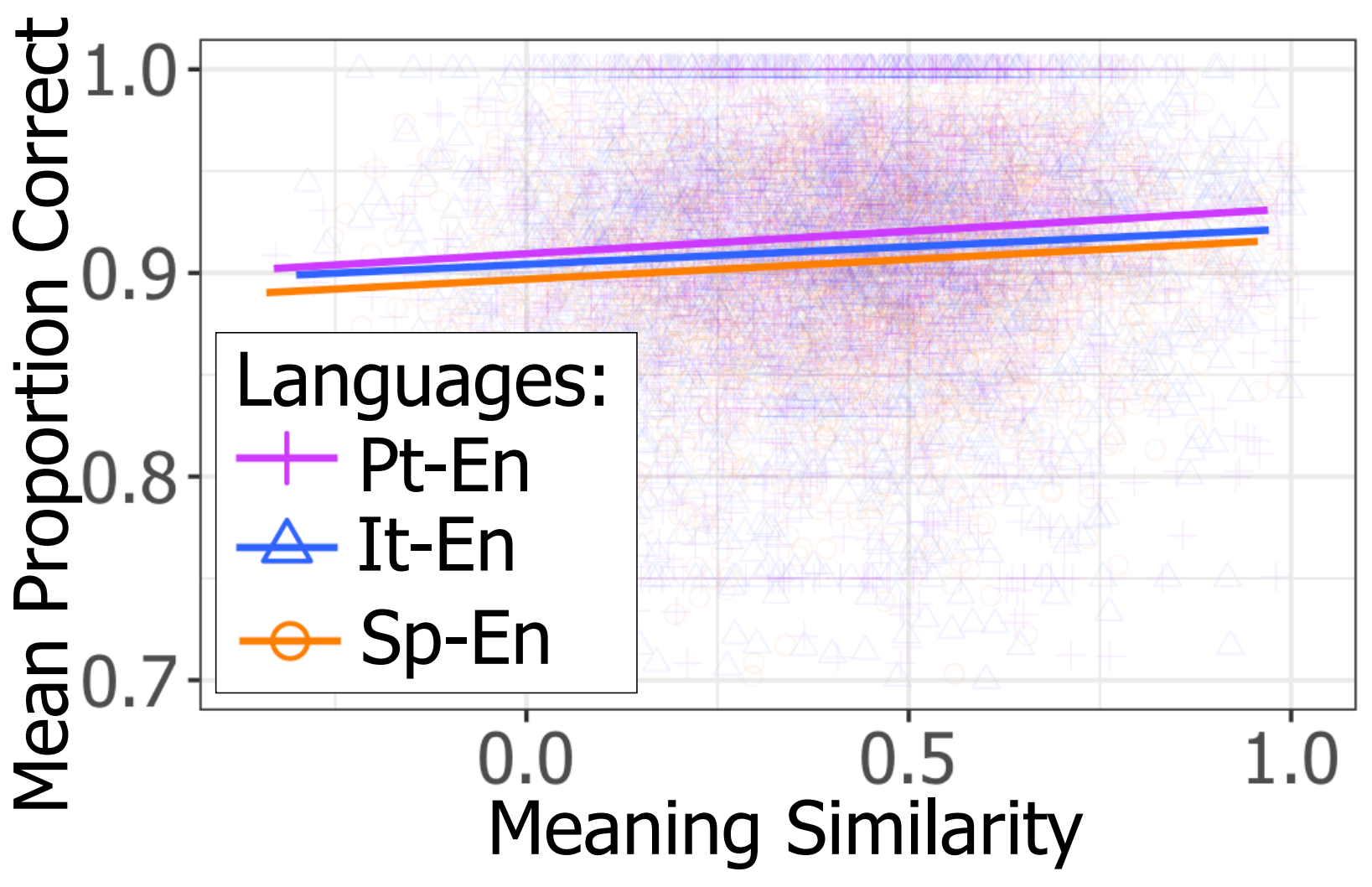
Predicting word learning accuracy

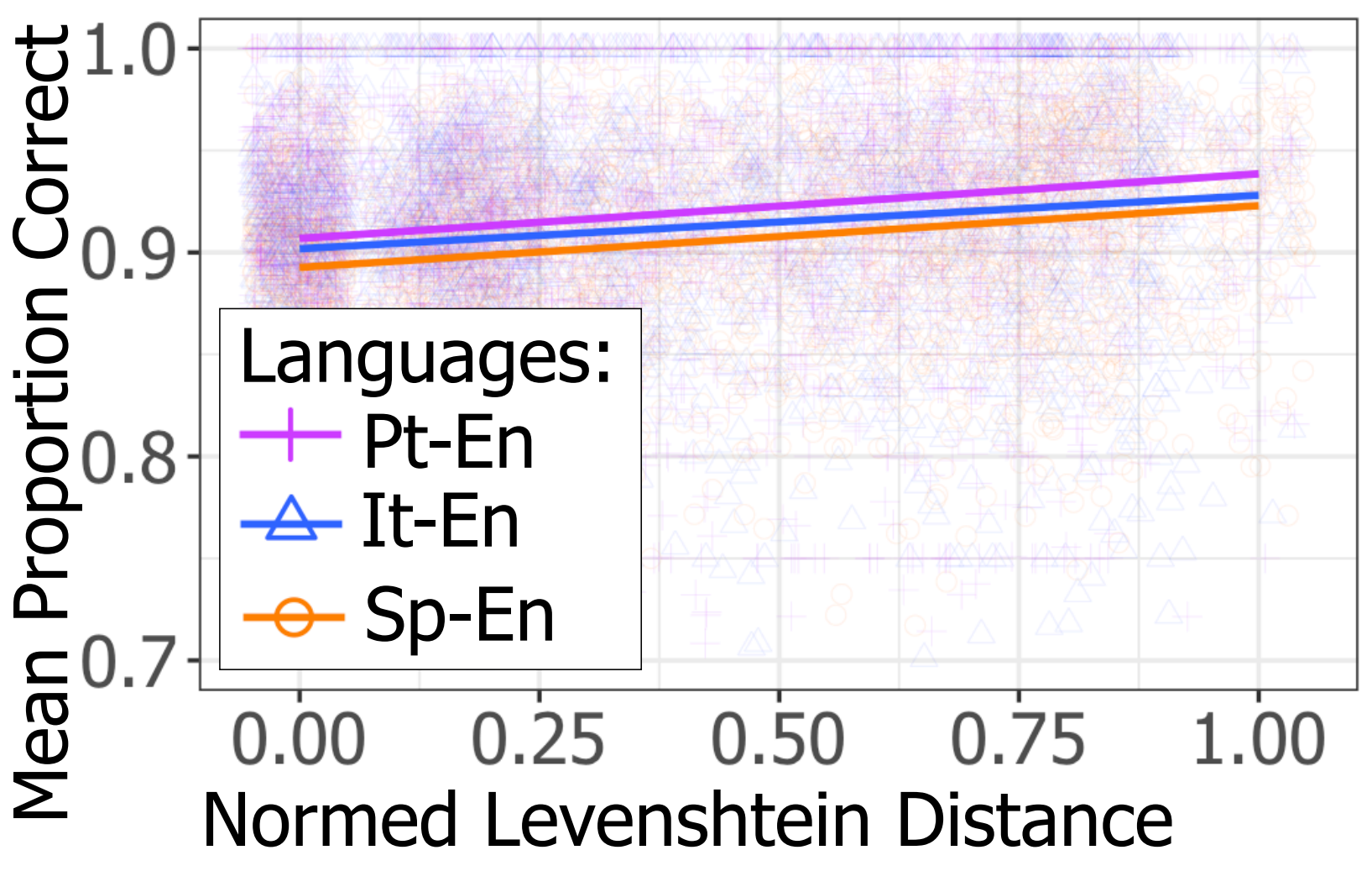
Linear mixed effects regression model results:



higher meaning similarity ⇨ easier to learn

more cognateness ⇨ easier to learn





Our new variable **Meaning Similarity** is predictive of word learning accuracy!

A continuous measure of **Cognateness** is predictive of word learning accuracy!

References

[1] De Groot, A.M.B., & 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).

[2] 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.

[3] 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)*.

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

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

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

[a] Department of Psychology, University of Wisconsin-Madison

[b] Institute for Applied Computational Science, Harvard University

[c] Department of Psychology, Princeton University

*Research was supported by a Menzies and Royalty research award from the Psychology Department at the University of Wisconsin - Madison

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?