Language models show cross-language similarities rather than differences between L1 and L2 speakers

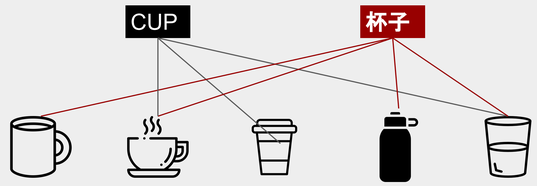
Jingyuan Selena She[1,2], Benjamin Zinszer[1]

1. Swarthmore College, 2. Haverford College

Introduction

Bilingualism & lexical semantics

- Second language learners gradually adapt L2 word meanings to be more similar to native-speakers' judgments of **word-word similarity** [2]
- L2 word meanings follow a **years-long trajectory** & are shaped by **both L1 and L2** knowledge [6]
- Bilingual lexical semantics show **cross-language convergence** [1]



Translation equivalence & inequivalence

Brain response patterns show **cross-language consistency** between word meanings in English and Chinese [3][4]

Speakers of different languages derive **unique semantic information** from translation-equivalent words [6]

Do deep learning language models reflect these nuances in word meaning?

➔ **Word embedding model:** context-independent model that vectorizes individual words into static vectors

➔ **Transformers model:** context-dependent model that dynamically computes word vectors in different sentences

Suppose language models are effective models of word meanings. We predict that they will correlate with different groups of speakers in systematic ways:

❖ L1 English speakers ~ English models
>
L2 English speakers ~ English models

❖ L2 English speakers ~ Mandarin models
>
L1 English speakers ~ Mandarin models

❖ L1 English speakers ~ Multilingual model
==
L2 English speakers ~ Multilingual model

Method

53 English monolinguals (L1 English speakers) , 33 Mandarin-English bilinguals (L2 English speakers)

Participants rated similarity of meaning between pairs of two words (28 words, 406 pairs)

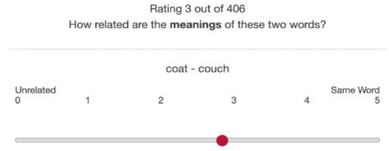
Human ratings of semantic similarity

Procedure:

1. Participants self-reported proficiency and history for each language they speak or understand
2. Participants rated similarity of meaning between two words (28 total words, 406 combinations, plus 28 "same word" catch trials)

Analyses:

- Subtract by 5, took mean score of each word pair
- Applied hyperbolic arctan transformation
- Construct distance matrices from word ratings



Language models of semantics

Model	*Model Type*	*Dimension*	*Languages*
Word2Vec	Embedding	300d	English, Mandarin
GloVe	Embedding	300d	English
Fasttext	Embedding	300d	English, Mandarin
BERT	Transformers	768d (L0, pooled)	English, Mandarin, Multilingual (104 langs)

Procedure:

1. Query English models with original word pairs in English
2. Query Mandarin models with translated Mandarin word pairs

Analyses:

- Extract word representations from each model
- Calculate cosine distance between each word pair
- Applied hyperbolic arctan, construct distance matrices

Pairwise correlations of human ratings to the models

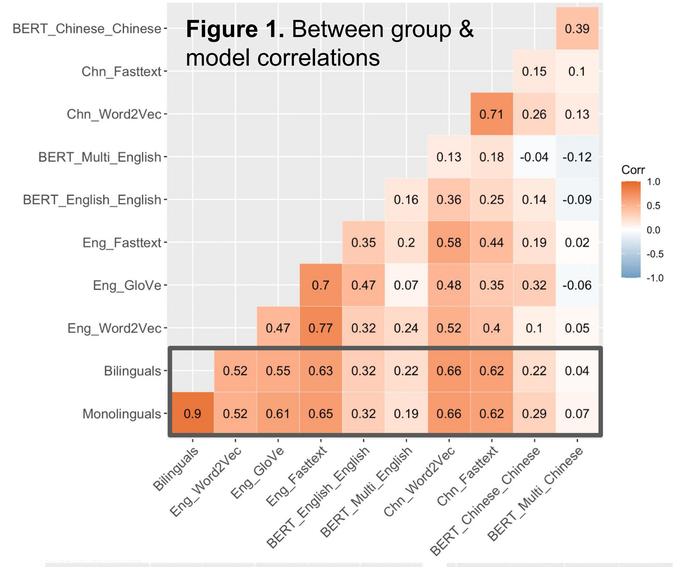


Figure 1. Between group & model correlations

Result

- All correlations between one model and one human group are significant ($r > 0$, $p < 0.001$; in gray box on Figure 1) except for Multilingual BERT of Mandarin word pairs
- English GloVe correlated more with L1 than L2 English speakers ($p<0.001$, $n = 375$)
- Mandarin Word2Vec correlated with L2 **and L1** English speakers more than English Word2Vec with L2 **and L1** English speakers ($p<0.001$)
- No Mandarin model correlated more with L2 English speakers than with L1 English speakers
- Multilingual BERT model of English word pairs did not correlate significantly more (or less) with L1 than L2 English speakers ($p=0.2$)
- No other models correlated differently between L1 and L2 English speakers

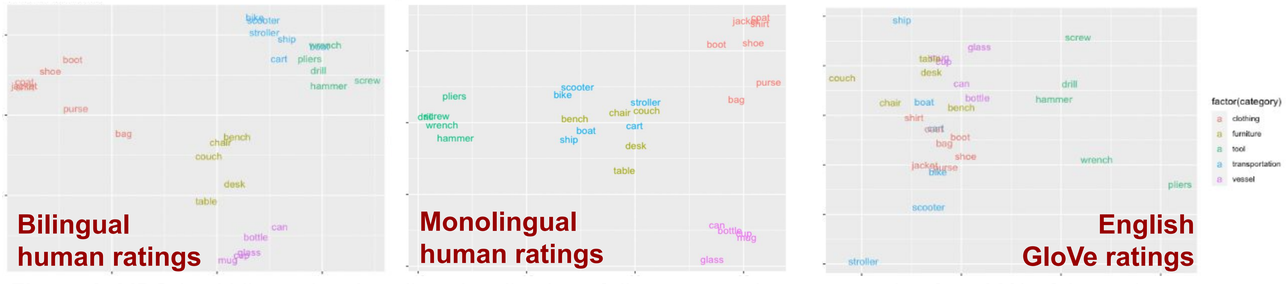


Figure 2. MDS (multidimensional scaling visualization of distance matrices, accounting for ~60% of the variance

Discussion

Language models show cross-language similarities rather than differences of L1 and L2 English speakers

- GloVe model **correlated moderately with L2, but less than L1.** This trend is consistent with previous findings about translation inequivalence [6], but a smaller effect than expected.
- All models except for English GloVe correlated similarly with L1 and L2. In general, language models exhibit cross-language similarities than differences
- Between-group correlation of L1 and L2 speakers was **stronger than expected** & by far the **closest match between any of the comparisons**

Static embedding models outperform transformers models in noun meaning representations

Older static embedding models still correlate more to behavioral noun-noun similarity ratings than state-of-the-art transformers models.

Language models are not reliable references for bilingual lexical semantics

- L2 adaptation takes years to fine-tune, and it's hard to track individuals in longitudinal studies
- This shows that language models are not yet reliable references to track language learning & word meaning acquisition

Acknowledgements

We are grateful to Gaby Ma for helping us with the initial data collection, Wendy Wen for word pair translations & helpful feedback on our study, and the students of the PSYC 001 course in Fall 2020 and Spring 2021 semesters for participating in this project.

References

1. Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of Memory and Language, 53*(1), 60-80.
2. Dong, Y., Gui, S., & MacWhinney, B. (2005). Shared and separate meanings in the bilingual mental lexicon. *Bilingualism, 8*(3), 221.
3. Yang, Y., Wang, J., Bailer, C., Cherkassky, V., & Just, M. A. (2017). Commonality of neural representations of sentences across languages: predicting brain activation during Portuguese sentence comprehension using an English-based model of brain function. *NeuroImage, 146*, 658-666.
4. Zinszer, B., Anderson, A. J., & Raizada, R. D. (2016). Chinese and English speakers' neural representations of word meaning offer a different picture of cross-language semantics than corpus and behavioral measures. In *Proceedings of the Cognitive Science Society*.
5. Zinszer, B. D., Anderson, A. J., Kang, O., Wheatley, T., & Raizada, R. D. (2016). Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. *Journal of Cognitive Neuroscience, 28*(11), 1749-1759.
6. Zinszer, B. D., Malt, B. C., Ameel, E., & Li, P. (2014). Native-likeness in second language lexical categorization reflects individual language history and linguistic community norms. *Frontiers in Psychology, 5*, 1203.