Re-examining cross-cultural similarity judgments using language statistics

Khuyen N. Le (knl005@ucsd.edu)[1], **Shan Gao (shangaocog@gmail.com)**[2],
Michael C. Frank (mcfrank@stanford.edu)[3], **Alexandra Carstensen (abc@ucsd.edu)**[1],

[1]Department of Psychology, University of California, San Diego, [2]Department of Psychology, University of Chicago,
[3]Department of Psychology, Stanford University

Abstract

Is "cow" more closely related to "grass" or to "chicken"? Speakers of different languages judge similarity in this context differently, but why? One possibility is that cultures covarying with these languages induce differences in conceptualizations of similarity. Specifically, East Asian cultures may promote reasoning about thematic similarity, by which cow and grass are more related, whereas Western cultures may bias judgments toward taxonomic relations, like cow and chicken. We measure similarity judgments across the US, China, and Vietnam and replicate US-China differences, but do not find that responding in Vietnam patterns with China. Instead, similarity judgments in Vietnam are intermediate between the US and China. We also show that word embedding models (fastText models for each language) are related to judgments within each country, suggesting a possible alternative interpretation of cross cultural differences. Perhaps notions of similarity are similar across contexts, but the statistics of the linguistic environment vary.

Keywords: similarity; culture; language; semantics; lexical co-occurrence; variation; US; China; Vietnam

Introduction

Many cognitive processes rely on similarity, from inference and generalization to analogy, mathematics, and science. There is substantial consistency in human similarity reasoning, but also systematic variation across cultural and linguistic contexts. In particular, there is considerable evidence showing that similarity reasoning in East Asian cultural contexts differs from that in Western cultures (e.g. Nisbett & Masada, 2003). For example, in a triad task comparing preferences for taxonomic and thematic similarity (choose two out of three words that are most related to one another), Ji, Zhang, & Nisbett (2004) found that Chinese participants preferred thematic matching to a greater extent than European Americans. In an image version of this task, Chinese children (9-10 years old) are also more likely to choose thematic matches compared to their American counterparts (Chiu, 1972). This cross-cultural difference is also observed in novel object categorization, with Chinese participants preferring to group by family resemblance across multiple features and Americans preferring a single-feature rule (Norenzayan, Smith, Kim, & Nisbett, 2002). Across tasks, East Asian participants show a preference for thematic similarity based on causal, spatial, and temporal relationships while Western participants are more likely to make taxonomic matches based on the similarity of attributes, like shared color or shape, among objects (see Markman & Hutchinson, 1984 for a discussion of these types of similarity).

An influential perspective within cultural psychology links these differences in similarity judgment to tendencies toward analytic processing in Western cultures, and holistic processing in East Asian cultures (Nisbett, 2003). Analytic processing emphasizes rule-like relationships predicated on objects and their properties and, correspondingly, taxonomic similarity, while holistic processing emphasizes relations between objects and their context and therefor, thematic similarity. In related work, East Asian participants show a higher level of sensitivity to context than their Western counterparts when reproducing drawings from memory (Ji, Peng, & Nisbett, 2000); visually exploring naturalistic scenes (Chua, Boland, & Nisbett, 2005); describing scenes (Masuda & Nisbett, 2001); and in explaining the causes of ambiguous behaviors (Choi, Nisbett, & Norenzayan, 1999). The consensus interpretation in this literature ascribes cross-cultural differences in similarity judgment to variation in the *conceptualization* of similarity – with people from East Asian cultures relying on a more thematic notion of similarity than Westerners.

Alternatively, these judgments could be shaped by cross-cultural differences in the *input* to similarity judgment, that is, the statistics of the linguistic and/or physical environment as manifest in everyday experiences. Perhaps when faced with the triad task, participants from all cultures follow the same process for conceptualizing similarity, but rely on language or culture-specific input to this process. If we observe a difference in categorization between East Asian and Western participants, it could be that members of both groups use the same procedure (considering similarity that is influenced by both taxonomic and thematic relations), but that the input to this procedure differ between cultures, with East Asian participants exposed to more support for thematic similarity in their experience in comparison to their Western counterparts. Because this is a question about naturally-occurring differences in real-world reasoning contexts, it is difficult to address experimentally. In a controlled study, we could manipulate participants' input and determine whether differing inputs influence similarity judgments (e.g. McDonald & Ramscar, 2001), but this would not address the actual sources of cross-cultural variation in similarity judgments. Here, we draw on real-world data to estimate variation across cultures.

Estimating variation in experience via language statistics

To determine whether varied input to similarity judgments can in part explain cross-cultural differences, we need a way to operationalize variation in exposure to thematic and taxonomic relations. While exposure to types of similarity in everyday experience is difficult to measure, the statistics of language can provide a rough proxy. Language statistics are useful in that they are part of the input to everyday experience – and indeed, may afford many of the "experiences" that people have with infrequently encountered items, like cows or helicopters – and they provide an accessible measure. Previous work also suggests that language statistics, such as lexical co-occurrence or cosine distance of word embeddings, can be good predictors of similarity reasoning. Semantic models that are constructed using lexical co-occurrence (in comparison to annotated relations) have been shown to perform well on predicting human judgments about similarity between word pairs that are thematically or taxonomically related (Rohde, Gonnerman, & Plaut, 2006). Relatedly, a model trained on word-document co-occurrence can predict word association and the effects of semantic association on a variety of linguistic tasks (Griffiths, Steyvers, & Tenenbaum, 2007). Word embeddings like word2vec, gloVe, and fastText have also been shown to be good predictors for similarity judgments (Asr, Zinkov, & Jones, 2018; Jatnika, Bijaksana, & Suryani, 2019; Liu, Feng, Wu, Chan, & Fulton, 2019).

Our study uses cosine distances of fastText word vectors as a measure of lexical co-occurrence.[1] fastText is a system that uses lexical co-occurrence information to generate a vector representing each word in its lexicon (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). Practically speaking, fastText vectors are available for many languages, and fastText has also been shown to be sensitive to cultural differences in word meanings: the distribution of semantic meaning clusters generated by language-specific fastText vectors correlates with the cultural, historical, and geographical similarities of these languages (Thompson & Lupyan, 2020).

Language statistics may incorporate both cultural-specific environmental statistics, that is, the experiential *input* to similarity judgments, and culture-specific senses of similarity, the *conceptualization* of similarity. For example, cultural features (like farming) can lead to differences in environmental statistics (seeing cows and grass) and these can influence language (talking more about cows and grass). But other cultural features (like conceiving of similarity thematically) could also cause individuals to talk differently about the same experiences (mentioning what cows eat rather than what other animals cows are like). Accordingly, our approach examines the extent to which language statistics can predict cross-cultural differences in similarity judgments with the understanding that language statistics are likely a proxy for both *input* to and *conceptualization* of similarity.

[1] We also carried out our analysis using raw lexical cooccurrences and obtained similar results.

The present study

We measured taxonomic versus thematic similarity matching in a forced-choice word triad task in three populations. Following Ji et al. (2004), we measured preferences in the US and China. In addition, we collected data in a novel context: Vietnam. Vietnam is a Southeast Asian country that borders China and has historically been greatly influenced by Chinese culture (Hui, 2002). We were unable to find cultural indices specifically on thematic/taxonomic reasoning, but Vietnam shows greater similarity to China than the US across the Hofstede dimensions of cultural comparison (2018) and is often cited for its similarity to more canonical exemplars of East Asian culture (Hirschman & Loi, 1996; Whitmore, 1984; Woodside, 1998), for example, patterning with East Asian countries rather than Southeast Asian countries on a measure of family values (Minkov & Hofstede, 2012). Therefore, Vietnam serves as a suitable cultural context to investigate whether the claim made by Ji et al. (2004) and previous studies – that Eastern and Western cultures have different notions of similarity – extends beyond mainland China. In addition to these replication and extension questions, we tested whether fastText vectors from corpora corresponding to each language context (English, Mandarin, and Vietnamese) are good predictors for similarity judgments in each population.

This study is correlational and cannot evaluate causal relationships between environmental statistics and similarity judgments. However, this work can inform potential mechanisms by examining whether similarity judgments co-vary with environmental statistics that differ across these contexts.

Our specific research questions are as follows:

1. Do we replicate cross-cultural differences in similarity judgments between East Asian and Western cultures?
2. Are these cross-cultural differences related to differences in language?
3. Is our language model specific to items with a taxonomic-thematic contrast, or can it predict similarity more generally?

To preview our results, we replicate US-China differences and find that Vietnamese judgments are intermediate between these two. We find that language-specific statistics provide good predictions for cross-cultural differences in similarity judgments, and additionally, for more general similarity judgments that do not contrast taxonomic and thematic matches. Critically, the stimuli used to assess this more general case of similarity reasoning were constructed to be outside the scope of explanation for taxonomic-thematic accounts (for use as filler items). While the language model may succeed in taxonomic-thematic similarity predictions by picking up on differences in the conceptualization of similarity that are reflected in language data, this conceptualization account provides no prediction for these filler stimuli.

Taken together, our findings provide support for an alternative to previous accounts on which differing cultures induce

differing conceptions of similarity. On this alternative, language explains for culture-specific variation in similarity reasoning, both for taxonomic-thematic judgments and for judgments not attributable to these senses of similarity. While these findings do not rule out other cross-cultural factors, they show it is possible to explain some cross-cultural variation in similarity judgments purely from language statistics.

Methods

Participants

Data collection and analyses for this study were preregistered, and the pre-registration is available at `https://osf.io/g76sc`. We recruited 200 participants from the US, 200 from mainland China, and 199 from Vietnam. All participants were recruited through snowball sampling, which, in the US was seeded with student email lists at a large university, in China (CN) with student groups on WeChat, and in Vietnam (VN) with Vietnam-based student groups on Facebook. US participants were compensated with $5 gift certificates (USD), CN participants received 25CNY through WeChat credit transfer, and VN participants received 50,000₫ (VND) in phone credit,

We excluded 8 US participants (4%), 16 CN participants (8%), and 62 VN participants (31.16%[2]), who missed three or more of the 10 attention check questions. In doing so, we deviated from our preregistered exclusion criterion, which used a more stringent policy, excluding participants who missed any of the 10 attention checks. This approach would have substantially reduced our sample size, disproportionately affecting the VN sample (resulting in a final sample of 109 US, 132 CN, and 57 VN participants). We report results with the less stringent exclusion criterion, but we performed analyses with both samples and found broadly similar results. Any differences are noted with a footnote in the Results & Discussion section.

We also applied four demographic exclusion criteria designed to reduce cross-cultural influences across our samples: we excluded participants who (1) were not native speakers of the test language (English, Vietnamese, or Mandarin Chinese), (2) were fluent in another one of the study languages, (3) had lived outside of the test country for more than two years, or (4) had significant international experience (more than 6 international experiences of 2 days or longer). Following our pre-registration, we dropped any exclusion criterion that would exclude 25% or more of any one population, but only for that population. Specifically, we did not exclude VN and CN participants who spoke English, or US participants with significant international experience. As a result of these demographic exclusions, we excluded 73 US, 35 CN, and 27 VN participants. The final US sample included 119 participants (30M, 84F, 3 non-binary, 2 other), with mean age = 22.2 (SD = 8.15). The CN sample included 149 participants (61M, 87F, 1 other), with mean age = 23.1 (SD = 3.65). And the VN sample included 110 participants (34M, 71F, 5 other), with mean age = 22.21 (SD = 5.81).

[2]The high rate of exclusion within our VN sample may reflect some unfamiliarity with attention checks, leading to exploratory responding. However, because exploratory responding to attention checks is indistinguishable from chance responding (from bots or inattentive participants), we did not analyze this data further.

Materials

All stimuli and scripts for the experiment and analysis are available at `https://github.com/khuyen-le/coglink-cogsci23/`. We collected and adapted stimuli from previous studies to create test triads consisting of a cue with one thematic and one taxonomic match option. For example, “cow,” “grass,” and “chicken,” where “cow” is the cue, “grass” is the thematic match, and “chicken” the taxonomic match. We included 105 such triads, a superset including triads pulled from supplemental information and in-text examples across the literature, and others that we adapted or created. We selected triads on the basis of cultural familiarity within the US, Vietnam, and China. The triads were translated from English to Vietnamese and Mandarin by fluent bilingual speakers of each language, and back-translated to English by another fluent bilingual who was naive to the original English version. The back-translations were checked against the original English to identify ambiguities in the translation, which were resolved through discussion and selection of alternative terms when relevant.

Procedure

Each participant completed all 105 triads in blocks of 21 trials at a time (10 test triads, 10 filler triads, and 1 attention check per page), by selecting the match most related to the cue. We elicited similarity match responses by asking "Which thing is most closely related to the bolded [cue] item?" In Mandarin Chinese, we used the phrasing for "most closely related" reported by Ji et al. (2004): "关联". And in Vietnamese, we translated this as "liên quan nhất". The test triads were presented with 105 filler triads mixed in, to obscure the taxonomic-thematic two-answer forced choice structure of the test stimuli and reduce the likelihood that participants would become aware of the design. The filler triads were groups of three semantically related words, but where the match options were not distinguished by thematic vs. taxonomic similarity, for example, the cue "bird" with match options "lizard" and "toad." Additionally, we included 10 attention check trials, which were formatted similarly to the test and filler triads but included an instruction instead of a cue item, e.g., "Choose wife" with match options "wife" and "husband." In total, each participant completed 210 similarity judgments (105 test triads and 105 fillers) and 10 attention check questions, with all items presented in randomized orders that varied between subjects.

Corpus model

Our general approach is to predict behavioral preferences in similarity judgment using relative similarity between fastText word embeddings.

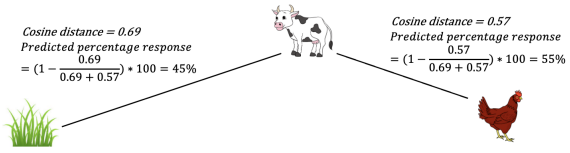

Figure 1: Thematic response prediction based on cosine distance for the triad cow: grass/chicken in English.

Word vector retrieval We use the fastText pre-trained models of English, Mandarin, and Vietnamese in Grave, Bojanowski, Gupta, Joulin, & Mikolov (2018). These models were trained on Common Crawl and Wikipedia using fastText and use character n-grams of length 5, and 10 negative examples. The training used a Continuous Bag of Words with position-weights and a window of size 5. From these models, we retrieve the word vectors (dimension 300) for each word in our triad and filler stimuli.

Similarity model To give an intuition for our model, consider again the cow: grass/chicken triad. We retrieved word vectors for "cow" and "grass", and calculated the cosine distance between these vectors. We did the same for "cow" and "chicken." Our similarity prediction is inversely proportional to the ratio of cosine distance of these pairs (Figure 1). This is because a larger cosine distance means the word vectors are further apart, and thus the words are less similar. For example, if the cosine distance of thematic cow-grass is 0.7 and the cosine distance of taxonomic cow-chicken is 0.3 (the more similar of the two), then our model predicts, correspondingly, that 30% of responses to the triad will be grass, and the other 70% chicken.

We calculated the cosine distance between each cue-thematic match (thematic cosine distance) and cue-taxonomic match (taxonomic cosine distance). We then calculated the thematic cosine distance proportion as thematic cosine distance over the sum of taxonomic cosine distance and thematic cosine distance. We followed this process for all three corpora, and were able to obtain predictions for all triads in all languages. Then, we used a mixed-effects regression to evaluate how well each corpus model predicts participants' similarity judgments, across triads and cultural contexts.

Results & Discussion

1. Replication of previous work and extension to a Vietnamese sample

In keeping with Ji et al. (2004), we would expect participants from mainland China to prefer thematic matches more than US participants (e.g., to prefer the cow-grass match over cow-chicken to a greater extent than US participants). On the basis of the previous literate more broadly, we would also expect participants from Vietnam to pattern with China, showing a stronger preference for thematic matches than US participants.

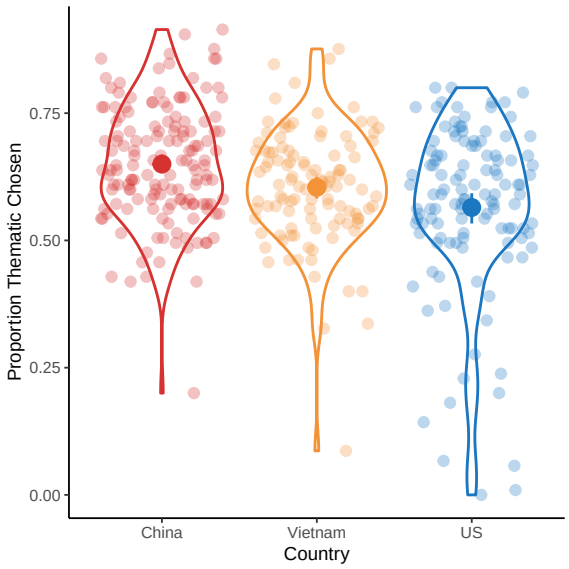

Figure 2: Proportion of thematic responses by country.

We observed the strongest preference for thematic matching in China (M = 0.65, SD = 0.11), followed by Vietnam (M = 0.6, SD = 0.11), and then the US (M = 0.56, SD = 0.17; see Figure 2).

To test for differences in similarity judgment between the countries, we ran a mixed-effects logistic regression predicting triad match (taxonomic or thematic) with country (US, China, or Vietnam) as a fixed effect. As random effects, we included an intercept per subject and one per triad, as well as by-triad random slopes for country to account for variation in the country effect across triads.

We found a significant effect of country ($\chi^2(2) = 15.37$, $p < .001$), but this effect is driven by the difference between US and CN responding ($\beta_{US} = -0.48$, $p < .001$). There was no statistical difference between matching in Vietnam and China ($\beta_{Vietnam} = -0.22$, $p = 0.09$), or the US and Vietnam ($\beta_{Vietnam} = 0.26$, $p = 0.086$).

In sum, we replicated the differences documented by Ji et al. (2004) between the US and China. However, we did not find that Vietnamese participants grouped particularly with Chinese participants. Instead they were intermediate between the two. Accordingly, we did not find strong support for the hypothesis that there are overall biases toward thematic responding across Asian cultural contexts broadly.

2. Language statistics as a predictor for cross-cultural variation in similarity judgments

In our next analysis, we tested whether triad-level differences in judgments across countries could be predicted from language statistics.

Single corpus model First, to test whether variation in language statistics can explain differences in similarity judg-

ments, we use a mixed-effects logistic regression fit individually for each country. Our models predicted responses by individual participants to particular triad (0=taxonomic or 1=thematic) with fastText predictions (proportion of cosine distance) as a fixed effect and participant and triad as random effects. If language statistics contribute to the differences in similarity judgments, we would expect each language corpus to be a good predictor for similarity judgments in the corresponding population.

All corpora were significant predictors of all cultural context responding, with $p < 0.05$ and β from -8.69 to -2.4.

Multi-corpus model If language statistics are able to predict meaningful culture-specific variation in similarity judgments (rather than just consistency across cultures), we would expect each corpus to be the best predictor of its corresponding culture compared to the other two corpora. We directly compared the corpus models by including all three corpus predictors as fixed effects in three mixed-effect regressions (predicting US, VN, and CN responding) with the same random effects as above.

For US responding: only the English (EN) corpus was a significant predictor[3]. EN corpus: β = -6.96, $\chi^2(1) = 16.57$, $p < .001$. VI corpus: β = -2.22, $\chi^2(1) = 3.73$, $p = 0.054$. ZH corpus: β = -3.18, $\chi^2(1) = 3.37$, $p = 0.066$.

For Vietnamese participants' responses, only the Vietnamese (VI) and Mandarin (ZH) corpora were significant predictors[4]. EN corpus: β = -2.98, $\chi^2(1) = 2.58$, $p = 0.108$. VI corpus: β = -2.75, $\chi^2(1) = 4.84$, $p = 0.028$. ZH corpus: β = -4.39, $\chi^2(1) = 5.49$, $p = 0.019$.

For responses by Chinese participants, only the Mandarin (ZH) and English (EN) corpus were significant predictors. EN corpus: β = -3.32, $\chi^2(1) = 7.18$, $p = 0.007$. VI corpus: β = -1.59, $\chi^2(1) = 3.63$, $p = 0.057$. ZH corpus: β = -5.31, $\chi^2(1) = 17.69$, $p < .001$.

We observed some language specificity in this analysis (Figure 3). The English corpus was the best predictor for US responding, and the Mandarin corpus was the best predictor for Chinese responding. While this is not the case with the Vietnamese corpus and Vietnamese participants' responding, the Vietnamese corpus was still a significant predictor for this responding. These results support our hypothesis that specific language statistics can predict cross-cultural variation in similarity judgment.

However, in all cultural contexts, adding the other two corpora produced a significantly better fit than the identical model with only the corresponding corpus included as a predictor (US responses: $\chi^2(2) = 7.93$, $p = 0.019$; VN responses: $\chi^2(2) = 13.72$, $p = 0.001$; CN responses: $\chi^2(2) = 10.56$, $p = 0.005$). This analysis suggests that culture-specific input to similarity judgments (as proxied by language statistics) do not fully explain cross-cultural differences in similarity judgment. This could be due to noise or sparsity in the estimates of language statistics, so that adding additional corpora simply improves fit by expanding the effective size of the model corpus. Alternatively, it could be due to true additional cultural-specific variance picked up by embeddings from other corpora – this is a question for future work.

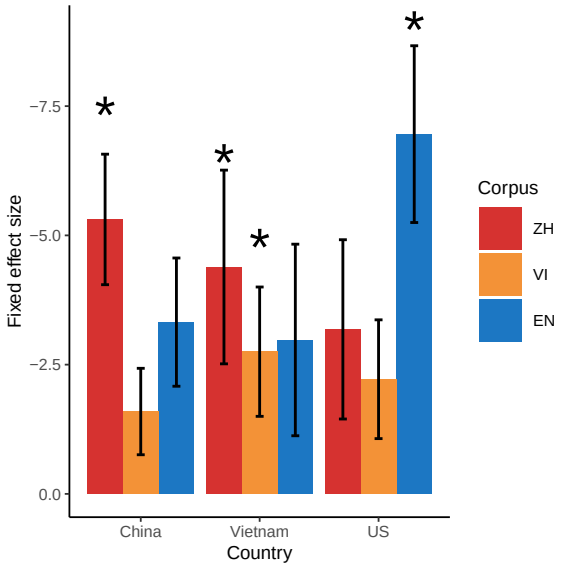

Figure 3: Fixed effect sizes of each corpus lexical statistics (cosine distance proportion) when included as a predictor for China, Vietnam, and US responding, respectively. Asterisks indicate corpora that are significant predictors. The English corpus is the best predictor for US response, and the Mandarin corpus is the best predictor for China response.

3. Language statistics as a predictor for non-structured filler items

One possible concern with our approach is that our models are only able to account for variation in similarity judgments because the word embedding models pick up on cross-cultural differences in the conceptualization of similarity. Perhaps some aspect of the geometry of the embedding space specifically reflects these differences rather than capturing more item-specific patterns of co-occurrence.

If this is the case, the embedding models should be less accurate at predicting responding for filler items (such as whether "tomorrow" or "yesterday" is more related to "today") because these items do not have a thematic/taxonomic structure. Additionally, even when one option might be more related to the cue, that relationship is not systematic throughout the set of fillers. Alternatively, if the embedding models capture similarity judgments more generally (beyond thematic/taxonomic preferences), they should predict judgments for the unstructured filler items as well as they do for our experimental items.

For each filler triad, we randomly assigned one of the responding options as 'Word1'. Running a mixed-effects logis-

[3]With the preregistered exclusion criterion, only the English (EN) and Mandarin (ZH) corpus were significant predictors.

[4]With the preregistered exclusion criteria, all three corpora were significant predictors.

tic regression to predict responding (Word1 or Word2) with country as a fixed effect and a random effect structure as above in Question 1, we found no effect of country on filler responding ($\chi^2(2) = 0.74$, $p = 0.692$).

Using the same mixed-effects logistic regression structure as the single corpus models in Question 2, we predicted responses (1=word1 or 0=word2) in each cultural context with the corresponding corpus predictor as a fixed effect, and participant and triad as random effects. In all cultural contexts, the corresponding corpus was a significant predictor of responding, with $p < 0.05$ and β from -9.59 to -3.14. These results show that the word embedding models predict general similarity in addition to structured thematic/taxonomic similarity.

To investigate whether language statistics predict taxonomic/thematic triads and filler items differently, we compared the variance accounted for by the models of the two different types of items. We ran two mixed-effects logistic regression models that predict responding across cultural contexts (1=thematic or word1, 0=taxonomic or word2), with the corresponding corpus predictor as a fixed effect and participant and item as random effects. Consistent with our results above, the corresponding corpus was a significant predictor for responding to both triads and filler items (triads: $\beta = -1.79$, $\chi^2(1) = 75.58$, $p < .001$, fillers: $\beta = -2$, $\chi^2(1) = 70.89$, $p < .001$). Importantly, we found comparable conditional R^2 values when the corresponding corpus was used to predict only triad items ($R^2 = 0.3$) and only filler items ($R^2 = 0.26$). This result provides evidence for our view that word embeddings index general tendencies in similarity judgment, beyond a specific dimension of thematic/taxonomic relations.

General Discussion

Do members of different cultures vary in their conceptualization of similarity? Cross-cultural differences in similarity judgment would seem to suggest this conclusion. In this paper, we consider whether statistics of the environment (as indexed by language statistics) can account for cross-cultural differences in a classic similarity judgment paradigm. We replicated the previously documented contrast between English speakers in the US and Mandarin Chinese speakers from East Asia (mainland China, Taiwan, Hong Kong, and Singapore): Chinese participants preferred thematic relations to a greater extent compared to US participants. Our sample of Vietnamese participants showed an intermediate preference but were not significantly different from either Chinese or US participants. This finding suggests some limitations on the generality of the cultural account, which proposes that thematic/taxonomic similarity preferences align with East Asian and Western tendencies toward holistic and analytic processing, respectively.

In contrast, we found some support for the environmental statistics account: word embeddings from language-specific corpora were a good predictor for the corresponding country's similarity judgments, even when other corpus statistics were included, and even for filler triads without a thematic/taxonomic structure. This evidence is correlational and cannot evaluate causes of variation in similarity reasoning; indeed, it may be that cultural variation in similarity reasoning shapes the linguistic variation we measure, but we note that this account does not explain the fit to filler triads. Overall, our results provide evidence that cross-cultural differences in similarity judgments are related to patterns in linguistic statistics (which also vary across cultures).

There are some important limitations of our approach. While we discuss cross-cultural variability at the level of countries or larger world areas, these are not cultural monoliths. For convenience, we operationalize culture at the level of country, based on where participants were raised and had lived for the large majority of their lives. It is an open question whether performance in our participant populations (of relatively young and well-educated adults) is representative of the broader country. This is especially true for societies with substantial ethnic and cultural variation such as the US. We expect that our data is likely to underestimate variation both within and between the countries we sample from.

Additionally, language, culture, cognition, and individual experiences are intertwined in complex causal relationships. In this study, we measure language and its relation to cross-cultural differences in similarity judgments, but these relations test only the plausibility of a language-based account; they cannot establish the direction of causality.

Ji et al. (2004) established that culture-aligned differences in this paradigm exist, even when the test language is held constant, concluding that "it is culture (independent of the testing language) that led to different grouping styles". Our data provide a cautionary note to this conclusion, suggesting that while cultural differences in similarity judgments exist, we might be able to better model such variation through environmental statistics such as lexical co-occurrence.

There are still many open questions for this account, however. Our operationalization of environmental statistics with word embeddings likely captures both language inputs and culture-specific conceptualization of similarity. Further investigation of the semantic spaces captured by cross-cultural word embedding models is required. Future work should also aim to provide a more specific computational account of how lexical co-occurrence might guide categorization preference beyond the simple proportion-of-similarity model tested here. This work should also investigate person- and item-specific factors, and characterize responses that are representative of each group compared to the others.

Despite these caveats, our findings here demonstrate the plausibility of an alternative perspective on cross-cultural accounts of similarity in the case of taxonomic and thematic reasoning. It may be the input to similarity judgments, rather than the evaluative process or the conceptualization of similarity that produces variation in similarity reasoning across cultural and linguistic contexts. We hope this work provides a foundation for further research probing this question.

Acknowledgements

We are very grateful to Anjie Cao, Mai Nguyen, Victoria Yang, and Van Luong for their help with stimulus design and international participant recruitment. We also thank members of the Language and Cognition Lab for their feedback and helpful comments. This work was funded in part by awards from NSF under grant SBE-2047581, the McDonnell Foundation, the Center for the Study of Language and Information at Stanford University supporting AC, and by the Vice Provost for Undergraduate Education Small Research Grant at Stanford University to KNL.

References

Asr, F. T., Zinkov, R., & Jones, M. (2018). Querying word embeddings for similarity and relatedness. In *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 675–684).

Chiu, L. (1972). A cross-cultural comparison of cognitive styles in chinese and american children. *International Journal of Psychology*, *7*(4), 235–242. http://doi.org/doi:10.1080/00207597208246604

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *International Journal of Psychology*, *125*(1), 47–63. http://doi.org/doi:10.1037/0033-2909.125.1.47

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences*, *102*(35), 12629–12633. http://doi.org/10.1073/pnas.0506162102

Country Comparison - Hofstede Insights. (2018). *Hofstede Insights*. Retrieved from https://www.hofstede-insights.com/country-comparison-tool?countries=china%2Cunited%2Bstates%2Cvietnam

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (LREC 2018)*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. http://doi.org/doi:10.1037/0033-295x.114.2.211

Hirschman, C., & Loi, V. M. (1996). Family and household structure in vietnam: Some glimpses from a recent survey. *Pacific Affairs*, *69*(2), 229–249. Retrieved from http://www.jstor.org/stable/2760726

Hui, W. (2002). Modernity and 'asia' in the study of chinese history. In E. Fuchs & B. Stuchteyi (Eds.), *Across cultural borders: Historiography in global perspective*. Rowman & Littlefield.

Jatnika, D., Bijaksana, M. A., & Suryani, A. A. (2019). Word2Vec model analysis for semantic similarities in english words. *Procedia Computer Science*, *157*, 160–167. http://doi.org/https://doi.org/10.1016/j.procs.2019.08.153

Ji, L., Peng, K., & Nisbett, R. E. (2000). Culture, control, and perception of relationships in the environment. *Journal of Personality and Social Psychology*, *78*(5), 943–955. http://doi.org/doi:10.1037/0022-3514.78.5.943

Ji, L., Zhang, Z., & Nisbett, R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology*, *87*(1), 57–65. http://doi.org/doi:10.1037/0022-3514.87.1.57

Liu, N., Feng, C., Wu, S., Chan, A., & Fulton, J. (2019). Automate RFP response generation process using FastText word embeddings and soft cosine measure. In *Proceedings of the 2019 international conference on artificial intelligence and computer science* (pp. 12–17). http://doi.org/10.1145/3349341.3349362

Markman, E. M., & Hutchinson, J. E. (1984). Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive Psychology*, *16*(1), 1–27. http://doi.org/doi:10.1016/0010-0285(84)90002-1

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of japanese and americans. *Journal of Personality and Social Psychology*, *81*(5), 922–934. http://doi.org/doi:10.1037/0022-3514.81.5.922

McDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (LREC 2018)*.

Minkov, M., & Hofstede, G. (2012). Is national culture a meaningful concept?: Cultural values delineate homogeneous national clusters of in-country regions. *Cross-Cultural Research*, *46*(2), 133–159. http://doi.org/10.1177/1069397111427262

Nisbett, R. E. (2003). *The geography of thought: How asians and westerners think differently ... and why*. New York: Free Press.

Nisbett, R. E., & Masada, T. (2003). Culture and point of view. *Proceedings of the National Academy of Sciences*, *100*(19), 11163–11170. http://doi.org/10.1073/pnas.1934527100

Norenzayan, A., Smith, E. E., Kim, B. J., & Nisbett, R. E. (2002). Cultural preferences for formal versus intuitive reasoning. *Cognitive Science*, *26*(5), 653–684. http://doi.org/doi:10.1207/s15516709cog2605_4

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2006). An improved model of semantic similarity based on lexical co-occurrence. *Communications of the ACM*, *8*, 627–633.

Thompson, R., B., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour Volume*, *4*, 1029–1038. http://doi.org/https://doi.org/10

.1038/s41562-020-0924-8
Whitmore, J. K. (1984). Social organization and confucian thought in vietnam. *Journal of Southeast Asian Studies*, *15*(2), 296–306. Retrieved from http://www.jstor.org/stable/20070597
Woodside, A. (1998). Territorial order and collective-identity tensions in confucian asia: China, vietnam, korea. *Daedalus*, *127*(3), 191–220. Retrieved from http://www.jstor.org/stable/20027512