Re-examining cross-cultural similarity judgements using lexical co-occurrence

Anonymous CogSci submission

Abstract

Is "cow" more closely related to "grass" or "chicken"? Speakers of different languages judge similarity in this context differently, but why? One possibility is that cultures co-varying 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 judgements toward taxonomic relations, like cow-chicken. This difference in notions of similarity is the consensus interpretation for cross-cultural variation in this paradigm. We consider, and provide evidence for, an alternative possibility, by which notions of similarity are similar across contexts, but the statistics of the environment vary. On this account, similarity judgements are guided by co-occurrence in experience, and observing or hearing about cows and grass or cows and chickens more often could induce preferences for these groupings, and account in part for apparent differences in notions of similarity across contexts.

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

Introduction

By virtue of discrete words and grammatical features, language provides a categorical partition of our continuous experiences. By measuring the similarity between words (as commonly done with lexical co-occurrence models), it may be possible to model the shape of similarity space in order to capture cross-linguistic and cross-cultural variation and provide a more general answer to the question of how language influences our conception of the world.

Taxonomic and thematic similarity provide a convenient entry point to broader debates about cross-linguistic and cross-cultural variation in notions of similarity. Taxonomic categorization is based on the similarity of attributes, for example, similar perceptual properties, like shared color or shape, among objects. In contrast, thematic categorization is based on causal, spatial, and temporal relationships among objects (Markman & Hutchinson, 1984).

Cross-cultural variation in similarity

Preferences for taxonomic and thematic similarity vary across cultures. Chiu (1972) found that Chinese children (9-10 years old) are more likely to choose thematic matches in a picture triad task (shown, e.g., images of an ant, a bee, and honey) than their American counterparts. These cross-cultural differences are 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). The authors link these differences to tendencies toward analytic processing in Western cultures, which emphasizes objects and their properties, and holistic processing in East Asian cultures, which emphasizes relationships between objects and their context (see also Nisbett (2003)). 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).

Ji, Zhang, & Nisbett (2004) asked whether differences in analytic and holistic processing are driven by language, culture, or a combination of these. They presented Chinese and European American adults with a word triad task, and found a preference for thematic matching among Chinese participants compared to European Americans. They found that test language contributed to this difference, but did not fully explain it: Chinese participants showed a stronger thematic preference when tested in Mandarin, but still showed a thematic preference when tested in English. Ji et al. conclude that culture (independent of language) leads to different styles of reasoning about similarity, whereas language serves as a "tuning" mechanism operating within a culturally-specific style, by activating representations corresponding to the language being used.[1]

This view ascribes cross-cultural differences in similarity judgment to variation in the conceptualization of similarity itself. Alternatively, these judgments could be shaped by cross-cultural differences in the statistics of the environment, and accordingly the content of everyday experiences. Perhaps when confronted with the triad task, participants from all cultures follow the same process for reasoning about similarity, but rely on language or culture-specific inputs 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

[1] But note that this work did not include a manipulation to prevent participants from engaging in e.g., covert naming in a language other than the test language.

that is influenced by both taxonomic and thematic relations), but the inputs to this procedure differ between cultures, with East Asian participants exposed to more instances of thematic similarity than their Western counterparts.

Estimating variation in experience

While co-occurrences in experience are difficult to measure, co-occurrence in language can provide a rough proxy–and indeed, language may afford many of the "experiences" that people have with infrequently encountered items, like cows or helicopters. In this study, we use co-occurrence in language as a proxy of cultural experience. This is a generalizing assumption, and while no linguistic corpus can be expected to fully capture a culture, using this proxy as a model of culture provides a conservative test of the hypothesis that differences in the input, rather than the process, of reasoning produces cultural variation in this task. This means that if such a model does work – if linguistic co-occurrence can predict cultural variation in this similarity task without building differences in the decision process into the model – it is relatively strong evidence that language alone (perhaps as a more easily observed source of information on broader cultural and ecological variation) may account for these cross-cultural differences in reasoning.

Indeed, previous work suggests that using lexical co-occurrence as a proxy can be useful in thinking about similarity reasoning. Natural language processing tasks have found that lexical co-occurrence is a good predictor for performance in similarity judgment. Griffiths, Steyvers, & Tenenbaum (2007) showed that a model that takes word-document co-occurrence as input can be used to predict word association and the effects of semantic association on a variety of linguistic tasks. Additionally, lexical semantic models that are constructed using lexical co-occurrence (as opposed 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).

Our study uses cosine distances of fastText word vectors as a measure of lexical co-occurrence[2]. 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).

Systems like fastText and word2vec have been demonstrated to be good predictors for similarity judgments. Liu, Feng, Wu, Chan, & Fulton (2019) showed that fastText performed well as the data for a system matching responses in a knowledge base by how similar they are to a customer's query. A study by Jatnika, Bijaksana, & Suryani (2019) tested a word2vec model trained on the English wikipedia corpus against 353 word pairs labeled with similarity values based on human judgment. The word2vec model correlates moderately with human judgment. fastText has also been shown to be sensitive towards cultural effects on word meanings: Thompson & Lupyan (2020) applied fastText to language-specific Wikipedia corpus (among others) to generate a "semantic neighborhood" of 1010 meanings in different languages. The study showed that languages that are spoken by more similar cultures, more geographically proximate and/or historically related had more semantic neighborhoods that aligned. This is a proof of concept for our use of fastText in different languages with varying levels of relatedness.

[2]We also carried out our analysis using raw lexical co-occurrences and obtained similar results.

The present study

Ji et al. (2004) suggest that the notion of similarity is what varies across cultural groups, but it is possible that similarity is regarded in the same way and it is actually the input to this similarity judgment that varies across cultural contexts. It may be possible to gain traction on potential mechanisms by examining whether variation in similarity judgments covary with environmental statistics that differ across cultural and linguistic contexts.

In this study, we ask whether variation in lexical co-occurrence can account for the cross-context (culture and language) differences observed in how people evaluate similarity. First, we attempt to replicate Ji et al. (2004) by running a study comparing English speakers in the US and Mandarin speakers in mainland China. Second, we evaluate whether the differences observed between English and Mandarin speakers' similarity judgments also extend to comparisons between English and Vietnamese, by running a similar study between English speakers in the US and Vietnamese speakers in Vietnam. Vietnam is a Southeast Asian country that borders China and is historically greatly influenced by Chinese culture (Hui, 2002). Therefore, it 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.

We then ask to what extent similarity judgments show cross-context differences that align with co-occurrence patterns in participants' languages. We would only see a main effect of cultural context if differing notions of similarity drive these cross-cultural differences. Meanwhile, if we do see variation in responding trial-to-trial that tracks with lexical co-occurrence statistics (as a rough proxy combining both linguistic and cultural context), this suggests that cross-cultural notions of similarity might be similar, and it is environmental statistics that guide the differences in similarity judgments.

To preview our results, we find that, while we do not see overall group differences (between the East Asian contexts and the US) in similarity, we do find language-specific co-occurrence can indeed explain cross-context differences in similarity judgments. In contrast to previous accounts, by which culture induces differing conceptions of similarity (varying between taxonomic and thematic), these findings provide an alternative, and more specific, account by which language may explain these cross-cultural differences without invoking variation in notions of similarity.

There are several 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. 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.

Language, culture, cognition, and individual experiences are intertwined in complex causal relationships: cultural features (farming) can lead to differences in individuals' experiences (seeing cows and grass) and in turn these can influence language (talking more about cows and grass). But cultural features (a thematic orientation) could also more directly cause individuals to talk differently about the same experiences (mentioning what cows eat rather than what other animals cows are like). In this study, we measure language and its relation to cross-cultural differences in categorization, but these relations test only the plausibility of a language-based account; they cannot establish the direction of causality.

Methods

Participants

We recruited 200 participants from the US, 199 participants from Vietnam, and 200 participants from mainland China. US participants were recruited through snowball sampling seeded with Stanford student email lists, Vietnam participants were recruited through snowball sampling seeded with Vietnam-based student groups on Facebook, and mainland China participants were recruited through snowball sampling seeded with group chats on WeChat. US participants were compensated with $5 gift certificates (USD), VN participants received 50,000₫ (VND) in phone credit, and mainland China participants received 25CNY through WeChat credit transfer.

We excluded 26 US participants, 123 Vietnam participants, and 34 China participants who did not answer all attention checks correctly. We followed 4 exclusion criteria that aim to retain only participants who are influenced by one culture: (1) non-native speakers of English and Vietnamese, respectively, (2) fluent in at least one of the other two study languages (Vietnamese for US participants, English for Vietnamese participants and Chinese participants), (3) have lived outside of the test country (US, Vietnam, or China) for more than two years, and (4) have significant international experience (more than 6 international experiences of 2 days or longer.) We did not use a particular criterion for a language if it would exclude 25% or more of any one sample. In this round of exclusion, we excluded 65 US, 19 Vietnam participants, and 34 China participants. After these exclusions, the US sample included 109 participants (27M, 78F, 3 non-binary, 1 other), with mean age = 22.4 (SD = 8.49) and median age = 20. The Vietnam sample included 57 participants (17M, 40F, other), with mean age = 22.21 (SD = 6.12) and median age = 21. The China sample included 132 participants (51M, 81F, other), with mean age = 23.27 (SD = 3.75) and median age = 23.[3]

Notably, we lost a majority of our Vietnam participants (more than 60%) in the attention check exclusion. One reason we suspect why this might have happened is because our Vietnam participants are less familiar with research surveys and attention check questions, and thus might have thought too much about the attention check questions.[4]

Stimuli

We adapted stimuli from previous studies to create a set of 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 in the US, Vietnam, and China. The triads were originally in English; they were translated to Vietnamese and Mandarin by a fluent bilingual speaker in each language. The translations were then checked for accuracy after backtranslation to English by another fluent bilingual in each language who was naive to the original English versions. –>

Each participant completed all 105 triads in sets 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 ("Which thing is most closely related to the bolded item?" or "Thứ nào liên quan nhất với thứ được in đậm?"). 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, e.g., cue "bird" with match options "lizard" and "toad." Additionally, we included 10 attention check trials, which were formatted like 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 and 10 attention check questions, with triads presented in randomized orders that varied between subjects.

Corpus model

To give an intuition for our model, consider again the cow-grass-chicken triad: we retrieved word vectors for "cow" and

[3] A table summarizing number of participants lost at each round of exclusions is included in the Appendix.

[4] We carried out an exploratory analysis where we used a less stringent attention check exclusion (covered in the final part of the Results & Analysis section). To preview our findings for this analysis, we did not find any difference in significant results when compared with the main analysis.

"grass", and calculate the cosine distance between these vectors. Similarly, we retrieved vectors for "cow" and "chicken" and calculate the cosine distance between them. Our similarity prediction is then inversely proportional to the cosine distance of these pairs. 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, then our model predicts, correspondingly, that 30% of responses to the triad will be grass, and the other 70% chicken. We then use a mixed-effects regression to evaluate how well each corpus model predicts participants' similarity judgments, across triads and languages.

Word vector retrieval We use the fastText pre-trained models of English, Mandarin, and Vietnamese in Grave, Bojanowski, Gupta, Joulin, & Mikolov (2018). These models are trained on Common Crawl and Wikipedia using We distribute pre-trained word vectors for 157 languages, trained on Common Crawl and Wikipedia using fastText. These models were trained using a Continuous Bag of Words (CBOW) with position-weights and a window of size 5. The models use character n-grams of length 5 and 10 negative examples.From the aforementioned models, we retrieve the word vectors (dimension 300) for each word we are interested in.

Similarity model From the word vectors, we calculated the cosine distance between each cue-thematic match (thematic cosine distance) and cue-taxonomic match (taxonomic cosine distance), using the spatial.distance.cosine function from the SciPy package (Virtanen et al., 2020). We then calculated the thematic cosine distance proportion as thematic cosine distance over the sum of taxonomic cosine distance and thematic cosine distance. We did this for all three corpora. We were able to obtain predictions for all triads in all languages.

Results

1. Do we extend previous work reporting a preference for taxonomic matching in the US and thematic matching in Asia?

The group means of proportion of thematic response in mainland China is the highest (M = 0.65, SD = 0.12), followed by the groups means in Vietnam (M = 0.6, SD = 0.13), which is slightly higher than that of the US (M = 0.56, SD = 0.17) (Figure 1).

To test for cross-context differences in similarity judgments between the countries, we ran a mixed-effects logistic regression predicting triad responding (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. In R syntax, the model is: response ~ country + (1 | subject) + (country | triad).

Overall, there is a significant effect of country on proportion of thematic responses ($\chi^2(2) = 15.22$, $p < .001$). However, this effect is driven by the difference between US and China responding ($\beta = -0.53$, $p < .001$). There is no statistical difference between the Vietnam and China responding ($\beta = -0.25$, $p = 0.137$), and the US and Vietnam responding ($\beta = 0.28$, $p = 0.142$).

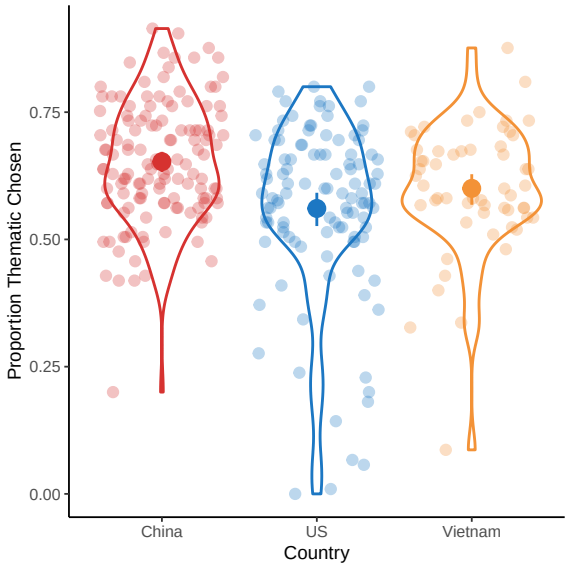

Figure 1: Proportion of thematic responses by country.

On this analysis, we do not find support that the US-China tendencies toward taxonomic and thematic responding (respectively) extend to the US-Vietnam comparison. Accordingly, we cannot speak to overall biases toward thematic responding across Asian cultural contexts broadly, but we do replicate the differences documented by Ji et al. (2004) between the US and China. However, in our corpus model comparison, we do find evidence for different, more fine-grained variation in similarity judgments between the US and Vietnam.

2. Can the differences in similarity judgments between English, Mandarin and Vietnamese speakers be explained by variation in lexical statistics?

Is responding in each cultural context predicted by lexical statistics? To test whether variation in lexical statistics can explain differences in similarity judgments between US and Vietnam participants, we compare logistic mixed-effects regression models fit to the thematic responding data from each country separately. We first ask how well each corpus model (English, Vietnamese, or Mandarin) predicts similarity judgments by speakers of the corresponding language (US, Vietnam, or China). To do this, we use a mixed-effects logistic regression to predict triad responses (0=taxonomic or 1=thematic) with corpus prediction (proportion of cosine distance) as a fixed effect and participant and triad as random effects.

We found that all corpora are significant predictors of all cultural context responding, with $p<0.05$ and β from -9.05 to

-2.44. (For a full report, see Appendix.)

Is responding in each cultural context best predicted by the corresponding corpus's lexical statistics, as opposed to the other two corpora? Next, we directly compare the corpus models by including both as fixed effects in three mixed-effect regressions (predicting US, Vietnam and China responding) with the same random effects as above. In R syntax, the model is: response ~ corpus_English + corpus_Vietnamese + corpus_Mandarin + (1|triad) + (1|subject).

For US responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are all significant predictors. EN corpus: β = -7.14, $\chi^2(1) = 16.78$, $p < .001$. VI corpus: β = -2.28, $\chi^2(1) = 3.74$, $p = 0.053$. ZH corpus: β = -3.53, $\chi^2(1) = 3.99$, $p = 0.046$.

For Vietnam responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are all significant predictors. EN corpus: β = -4.25, $\chi^2(1) = 3.87$, $p = 0.049$. VI corpus: β = -2.98, $\chi^2(1) = 4.17$, $p = 0.041$. ZH corpus: β = -4.86, $\chi^2(1) = 4.94$, $p = 0.026$.

For China responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are all significant predictors. EN corpus: β = -3.52, $\chi^2(1) = 7.85$, $p = 0.005$. VI corpus: β = -1.62, $\chi^2(1) = 3.57$, $p = 0.059$. ZH corpus: β = -5.35, $\chi^2(1) = 17.19$, $p < .001$.

We observed some level of language specificity from this analysis. The English corpus is the best predictor for US responding, and the Mandarin corpus is the best predictor for China response. While this is not the case with the Vietnamese corpus and the Vietnam responding, the Vietnamese corpus is still a significant predictor for the Vietnam responding (Figure 2).

However, we found that language specificity alone does not explain our results. We used an ANOVA to compare the model with only the corresponding corpus, and the model with all 3 corpora for each cultural context. We found that in all three cases (US, China, Vietnam), adding the other two corpora produces a significantly better fit than the identical model without the additional corpora, and only the corresponding corpus included as a predictor (US response: $\chi^2(2) = 8.42$, $p = 0.015$; Vietnam response: $\chi^2(2) = 15.21$, $p < .001$; China response: $\chi^2(2) = 10.83$, p NA).

3. Does similarity reasoning differ across cultures, only the input to it, or both?

Our analysis shows that lexical statistics of each corpus is a significant predictor for similarity judgement in the corresponding cultural context, even without the addition of the other two corpora. Assuming that lexical statistics is a good proxy for environmental statistics, we take the above result as evidence that cross-cultural differences in input to the process of similarity judgement at least partially drives cross-cultural similarity judgement differences. However, it is still an open question whether the nature of similarity judgement (for e.g., the notion of similarity) itself differs across cultures and contributes to cross-cultural differences in similarity judgement responding. In an alternative hypothesis, the cross-cultural differences in similarity judgement observed is driven by not only the input, but also an interaction between cultural-specific nature of similarity judgement and input. To test this hypothesis, we operationalize cultural-specific nature of similarity judgement as country context, and input to similarity judgement as the corresponding corpus statistics. We then compare a model that includes country, corresponding corpus statistics and their interaction as fixed effects (R syntax: responses ~ corresponding_corpus * country + (1|triad) + (1|subject)) to a model with only the corresponding corpus statistics as the fixed effect (R syntax: responses ~ corresponding_corpus + (1|triad) + (1|subject))

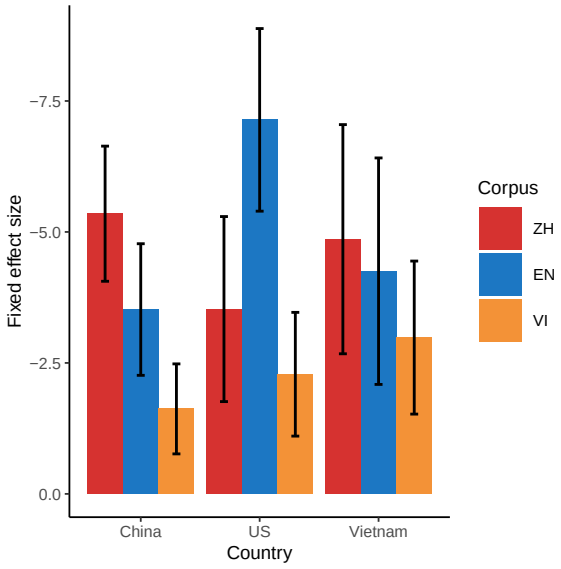

Figure 2: Fixed effect sizes of each corpus lexical statistics (cosine distance proportion) when included as a predictor for China, US, and Vietnam responding, respectively. The English corpus is the best predictor for US response, and the Mandarin corpus is the best predictor for China response.

In the model containing only the corresponding corpus statistics, the corpus statistics is a significant predictor (β = -1.75, $\chi^2(1) = 48.9$, $p < .001$). When adding country and interaction between country and corpus, both the corpus statistics (β = -1.51, $\chi^2(1) = 13.96$, $p < .001$) and interaction between country and corpus (β = NA, $\chi^2(2) = 9.1$, $p = 0.011$) are significant predictors. This suggests that there are cross-cultural differences in both input and nature of similarity judgement that drive differences in similarity judgement responding.

4. How general are the cross-language differences in similarity judgments we observe? Are they limited to taxonomic-thematic contrasts or do they extend to other cases?

Previous work using the triad task such as Ji et al. (2004) has focused on contrasting taxonomic-thematic match preferences, driven by a theory of cross-cultural differences in no-

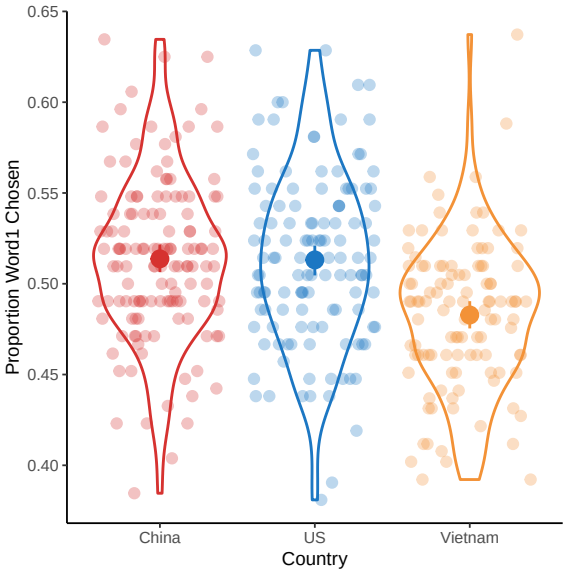

Figure 3: Proportion of thematic responses by country.

tions of similarity. Meanwhile, our approach has shown that notions of similarity only has an effect on cross-cultural differences in similarity judgement when modulated by input, and that input as proxied by lexical statistics of the corresponding corpus by itself is a significant predictor for similarity judgement responding. Our approach using lexical statistics do not assume a taxonomic/thematic structure in the triads. Therefore, it should be possible to predict similarity judgements to triads that do not follow a taxonomic/thematic structure, such as our filler triads. If our approach is tractable, we should be able to repeat the above analyses to the filler triads, and find that lexical statistics of each corpus is a significant predictor for similarity judgements in filler triads of the corresponding cultural context.

For each filler triad, we randomly assigned one of the responding optiosn as 'Word1'. Running a mixed-effects logistic regression predict responding (Word1 or Word2) with structure equivalent to Analysis 1, we found no effect of country on filler responding (Figure 3).

Does lexical statistics for fillers predict responding in the corresponding cultural context? Using the same mixed-effects logistic regression structure as Analysis 2, we predict responses (1=word1 or 0=word2) with corpus prediction (proportion of cosine distance) as a fixed effect and participant and triad as random effects. We found that in all cultural contexts, the correpsonding corpora is a significant predictor of responding, with $p<0.05$ and β from -9.59 to -3.05. (For a full report, see Appendix.)

Discussion

In this paper, we consider whether statistics of the language environment can account for cross-cultural differences in a classic similarity judgment paradigm, as an alternative to the view that members of different cultures vary in their conception of similarity.

We first tested the generality of a cultural account which holds that people from Western and East Asian cultures tend to conceive of similarity in more taxonomic and thematic ways, respectively, and respond accordingly in categorization tasks such as ours. While we managed to replicate the previously documented contrast between English speakers in the US and Mandarin Chinese speakers from East Asia (mainland China, Taiwan, Hong Kong, and Singapore), we do not extend this contrast to our sample of Vietnamese speakers in Vietnam and English speakers in the US. This finding suggests some limitations on the generality of this cultural account.

We did find some signatures of language specificity in our analysis, such as the large positive correlation between similarity judgments of each country and the respective corpus statistics, and how each corpus statistics are good predictors for corresponding country's similarity judgments. However, this is potentially due to the high correlation between corpus statistics of English, Vietnamese and Mandarin. We find even stronger evidence for consistency across the three groups, with substantive overlapping predictions across the corpus models, highly similar responding across the experiments, and a correspondingly high fit in cross-language comparisons between models and data.

Our findings raise additional questions for future work: if not differences in taxonomic vs. thematic responding, then what differences drive the relativity effects previous studies have observed? To what extent are the relativity effects driven by language, and to what extent by culture? 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" in their study. Our data provide a cautionary note to this conclusion, suggesting that semantic representations in bilinguals (see Francis (2005) for a review) may have the potential to provide an offline account for cross-context differences in similarity judgments, independent of test language. However, there are still many open questions for this account. How do semantic associations guide categorization? Can they explain taxonomic-thematic differences of the type reported by Ji et al. (2004) and others? Can we provide a more specific computational account than the simple frequency model tested here?

Despite these caveats, our findings here demonstrate the plausibility of an alternative perspective on cross-cultural accounts of language, thought, and similarity in the case of taxonomic and thematic reasoning: that 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.

Appendix

1. ICC of item variability and participant variability

Sample	Intra-rater ICC	Inter-rater ICC
US	0.32	0.97
China	0.20	0.96
Vietnam	0.32	0.95

Table 1: ICCs for reliability within participants (intra-rater) and stimuli (inter-rater) for all cultural contexts.

2. Supplemental for whether each corpus predicts similarity judgement in each cultural context

Triads For US responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are significant predictors of US responding (EN-US model: β = -9.05, $\chi^2(1)$ = 31.28, p < .001; VI-US model: β = -3.13, $\chi^2(1)$ = 5.1, p = 0.024; ZH-US model: β = -7.42, $\chi^2(1)$ = 18.48, p < .001).

For Vietnam (VN) responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are significant predictors of VN responding (EN-VN model: β = -6.86, $\chi^2(1)$ = 11.71, p < .001; VI-VN model: β = -3.83, $\chi^2(1)$ = 6.02, p = 0.014; ZH-VN model: β = -7.49, $\chi^2(1)$ = 13.84, p < .001).

For China (CN) responding: English (EN), Vietnamese (VI), and Mandarin (ZH) corpus are significant predictors of CN responding (EN-CN model: β = -6.18, $\chi^2(1)$ = 25.39, p < .001; VI-CN model: β = -2.44, $\chi^2(1)$ = 5.75, p = 0.017; ZH-CN model: β = -7.36, $\chi^2(1)$ = 37.72, p < .001).

Fillers English (EN) is a significant predictors of US responding (β = -7.93, $\chi^2(1)$ = 35.87, p < .001).

Vietnamese (VI) is a significant predictors of Vietnam responding (β = -3.05, $\chi^2(1)$ = 5.88, p = 0.015).

Mandarin (ZH) is a significant predictors of China responding (β = -9.59, $\chi^2(1)$ = 36.65, p < .001).

3. Demographic Exclusion Statistics

	Country		
	US	China	Vietnam
Ppts Finished	200	200	199
Attention Check Exclusions			
Excluded	26	34	123
After Exclusion	174	166	76
Demographics Exclusions			
Non-Native Speakers	18 *	3 *	2 *
EN Speakers	N/A	79	127
VI Speakers	4 *	1 *	N/A
ZH Speakers	30 *	N/A	9 *
Lived Abroad	29 *	40 *	18 *
Overseas Experience	77	12 *	20 *
After Exclusions	109	132	57

Acknowledgements

Place acknowledgments (including funding information) in a section at the end of the paper.

References

10 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

Francis, W. S. (2005). *Bilingual semantic and conceptual representation*. Oxford University Press.

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

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

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)*.

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

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

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. http://doi.org/10.1038/s41592-019-0686-2