A Cross-linguistic Study into the Contribution of Affective Connotation in the Lexico-semantic Representation of Concrete and Abstract Concepts

Simon De Deyne (simon.dedeyne@unimelb.edu.au)[a]
Álvaro Cabana (acabana@psico.edu.uy)[b]
Bing Li (52163200020@stu.ecnu.edu.cn)[c]
Qing Cai (qcai@psy.ecnu.edu.cn)[c]
Meredith McKague (mckaguem@unimelb.edu.au)[a]
[a] School of Psychological Sciences, University of Melbourne, Melbourne, VIC, Australia
[b] Facultad de Psicología, Universidad de la República, Montevideo, Uruguay
[c] Speech, Language and Neuroscience Group, NYU Shanghai, Shanghai, China

Abstract

Words carry affective connotations, but the role of these connotations in the representation of meaning is not well understood. Like other aspects of meaning, connotation might be culture or language-specific. This study uses a large-scale relatedness judgment task to determine the role of affective connotations in concrete and abstract words in English, Rioplatense Spanish, and Mandarin Chinese. Across languages, word valence, or how positive or negative a word is, was one of the main organizing factors in both concrete and abstract concepts. Moreover, predicted culture-specific affective connotations were reliably found in the similarity space of abstract concepts. A follow-up analysis was conducted to investigate whether distributional semantic representations derived from language similarly encodes these connotations using word embeddings. The language models did only partly captured the overall similarity structure and the affective connotations shaping it.

Keywords: affective connotation; cross-cultural meaning; relatedness; word embeddings

Introduction

Previous work on the representation of natural language concepts in cognitive psychology focuses on concrete concepts like *rose* or *dog*, their core features (*rose – has thorns*), taxonomic relations (*dog – mammal*), and thematic relations (*dog – bone*). The fact that these concrete words have strong affective connotations (*rose – romantic*, *dictator – evil*), is sometimes ignored as it is not clear whether such information constitutes a core property required to understand the meaning of a word. For example, in feature listing studies, introspective features reflecting attitudinal or emotional connotations are omitted in instructions. This practice might reflect the assumption that connotations are highly subjective (*vegetable – disgusting*), and peripheral to understanding the meaning of a word. In contrast to cognitive psychology, attitudinal components of meaning are well-accepted in social psychology ever since the work of Charles Osgood. Over an extensive research program, Osgood and colleagues identified three main factors that contribute to word meaning: valence (*unhappy - happy*), arousal (*calm - exciting*), and dominance or potency (*weak - strong*) (Osgood, Suci, & Tannenbaum, 1957). Subsequent cross-cultural work also showed that these three factors were universally shared across a wide range of languages (Osgood, May, Miron, & Miron, 1975). However, this universality only implies that these factors are important across languages but does not mean that words across different languages are equally positive or arousing. For example, the word *fat* in Chinese does not have the same negative connotation it has in English (Bozzeti-Engstrom, 2002). Previous research using explicit ratings of valence and arousal suggests that systematic differences are expected between Indo-European and East-Asian languages, especially in terms of how arousing words are (Lim, 2016). This affective connotation, defined in terms of valence and arousal, will be the focus of this study. If connotation (defined in terms valence and arousal) is central to meaning, the mental representations underlying meaning should reflect both universal and reliable culture-specific connotations. Furthermore, connotations should also affect human judgments in tasks that do not explicitly probe how positive or arousing words are.

To determine the role of affective connotations within and across languages, relatedness judgments will be collected in three world languages: English, Spanish, and Chinese. Theoretically, this choice allows us to investigate two Indo-European languages and Mandarin Chinese, a language and culture that is more distinct, especially in terms of affective connotations (Lim, 2016). Practically, these languages were chosen because relevant lexico-semantic norms (valence, arousal, and concreteness ratings; translation probability, word associations) are available in each of these languages. The current comparison focuses mainly on differences between English and Chinese, with Spanish added as a baseline to verify and contextualize the findings. Finally, to determine whether affective connotations generalize to both concrete and abstract words, two separate sets of stimuli that vary in concreteness will be used. This allows us to establish affective connotations in a wide range of words, including concrete emotion-laden ones (cf. Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011).

Relatedness Judgments

The role of connotation will be determined using relatedness judgments. In contrast to similarity judgments, relatedness judgments allow participants to consider commonalities between antonyms such as *bright* and *dark*, or *nice* and *awful.* Asking participant to judge similarity instead might introduce a bias by drawing attention to a smaller subset of semantic relations, which in this study might overestimate the role of valence in the representation. Moreover, previous work has shown that in contrast to relatedness judgments, similarity is difficult to distinguish from relatedness (e.g., *coffee* and *milk*), and consequently result in less reliable ratings (e.g., Hill, Reichart, & Korhonen, 2016). From a cross-linguistic perspective, focusing on relatedness rather than similarity might be of particular importance as well. Different languages have different semantic-syntactic interfaces and words belong to classes such noun or adjective based on distinct criteria (Haspelmath, 2012). As a consequence, whether two words can be considered related, but not similar, might be language-dependent.

In what follows the term similarity will be used in a technical sense to indicate the opposite of distance or a measure of distributional overlap of semantic vectors to construct similarity matrices.

Multivariate techniques such as multidimensional scaling (MDS) are ideally suited to investigate which factors contribute to the organization of word meaning. However, MDS requires a full similarity matrix for all the concepts under consideration. This presents a challenge when using direct pairwise judgments since the number of judgments quickly becomes prohibitively large. In the current study consisting of two sets of 81 words, this means that if 20 judgments per pair are required, a total of $20 \times (81 \times 80)/2 = 64{,}800$ judgments are needed per set. To address this issue, a partial relatedness ranking task was used in which participants had to pick the three most related items out of a list of response options.

Method

Participants A total of 24 English (17 female), 21 Spanish (14 female) and 23 Chinese (20 females) completed the abstract task, whereas 20 (15 female) English, 21 (13 female) Spanish and 20 (16 female) Chinese completed the task with concrete words. All participants were compensated with a gift voucher, except for 38 English participants who received course credits. The participants completed a language background and history questionnaire. Participants who did not speak English, Mandarin Chinese, or Rioplatense Spanish were not included in the study. This study was approved by the University of Melbourne Ethical Committee.

Stimuli For both abstract and concrete words, a list of nouns was compiled that allowed for some variation in affective connotation and where the Chinese and Spanish forms represented a plausible translation from English by a majority of Spanish and Chinese speakers. The lists were constructed from two ongoing translation studies in Rioplatense Spanish and Mandarin Chinese and two existing studies (Prior, MacWhinney, & Kroll, 2007; Wen & van Heuven, 2017). The combined norms had at least 20 observations per word and were used to select stimuli for which English to Chinese or Spanish translation agreement was larger than 60%. Based on the Brysbaert, Warriner, and Kuperman (2014) concreteness norms, words with a rating > 3.5 on a 5-point concreteness scale were considered concrete and the others abstract. Next, English and Spanish valence and arousal ratings were sourced from respectively Mohammad (2018) and Stadthagen-Gonzalez, Imbault, Sánchez, and Brysbaert (2017). For Chinese, a large-scale dataset was not available, and multiple resources were combined. These consisted of the data from Yu et al. (2016) and Yao, Wu, Zhang, and Wang (2017). As the majority of words covered in these studied were abstract, data from an unpublished dataset with norms for 2,418 words in Cantonese were included as well. To be able to investigate culture-specific affective connotation, the difference between the z-transformed values for either valence or arousal was calculated between English and Chinese words. The final set of concrete of abstract words contained 27 words where the difference between affective connotation (valence or arousal) was at least 1.5. For these 27 incongruent English-Chinese word pairs, most differed in terms of arousal (24 abstract words, 27 concrete words). One word (myself, see Osgood et al., 1975) was added to balance the number of alternatives per trial but was not included in further analyses.

Procedure The study was conducted online and consisted of a series of standard questionnaires (cf. supra) after which the participants were directed to the partial ranking task. Each participant was randomly assigned to the abstract or concrete condition and received instructions in their native language.

The cue words were shown in random order on top of the screen followed with three response boxes into which participants drag and rank the three most related words from a list of alternatives showing beneath the form. Instead of showing all 81 response alternatives at once, they were split into non-overlapping random and individually unique sets of 27 alternatives. This way, each participant ranked a total of 9 responses for each cue word. The participants were also instructed that cues would be repeated with different response alternatives combinations.

Results

For each individual, a similarity matrix was calculated which was used to determine the reliability of the ratings and contributed towards an aggregate similarity matrix used in later analyses. The responses for each individual were tabulated in a matrix with 81 rows and columns. Each row of this matrix consisted of summed counts for a specific cue that were weighted by adding the cue word itself (4/4), the first choice (3/4), the second (2/4), and the third choice (1/4). Next, the rows were normalized to sum to one. At this stage, the vectors are extremely sparse, and a similarity matrix derived from these data will consist of mostly zeros, which could lead to degenerate results when applying MDS. To address this issue, the ranked choice matrix was smoothed by considering this matrix as a weighted adjacency matrix of a graph in which each word is connected to its most similar neighbors. To increase the density of this graph, weighted indirect paths connecting word pairs were added similar to (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019). Using this graph, a similarity matrix was calculated between the vectors corresponding to the distributions of the weighted sum of direct and indirect paths. Across both experiments the decay parameter α, which determines the weight or contribution of longer over shorter paths was fixed at 0.5.

To calculate the reliability of the obtained space, a similarity matrix was extracted for each individual. Next, the upper triangle from the similarity matrix of each individual was compared to the data averaged over all remaining individuals in one of the three languages. Six English and 5 Spanish participants were removed because their similarity matrices correlations were smaller than 1.5 SD of the mean correlations of all other participants. The resulting Spearman-Brown corrected split-half reliability for each dataset in all three languages was larger than .90. The final sample of participants indicated that they were exposed to English, Spanish, or Chinese respectively 97%, 87% and 85% of the time.

Monolingual affective connotation

The first question is whether affective connotation is consequential in a relatedness ranking task where no explicit mention is made about affective connotations. To address this question, MDS will be used to explore and confirm the nature of the underlying factors that capture the similarity space.

Procedure

The similarity matrices were converted to distances, and interval multidimensional scaling was applied using SMACOF (de Leeuw & Mair, 2013). A property fitting approach was used to confirm the visual interpretation of the obtained configurations, aiming to identify dimensions corresponding to affective connotation (valence or arousal). To provide a baseline for this comparison, two additional lexico-semantic variables were also included: concreteness (Brysbaert et al., 2014, only for English), and log-transformed word frequency in English (Brysbaert & New, 2009), Mandarin Chinese (Cai & Brysbaert, 2010) and Spanish (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2012).

Results

To determine whether affective connotation plays an important role in organizing meaning, the present analysis focuses on the most important dimensions to achieve a reasonable MDS fit instead of the trivial case where a high dimensional solution leads to a perfect fit. Based on stress-plots for each dataset, a solution with 8 dimensions resulting in Stress-1 values < .08 was selected. Figure 1 shows configuration plots for dimensions D1 and D2 annotated by the arousal and valence of a word. Visual inspection suggests that meaning is organized primarily in terms of valence for both concrete and abstract concepts, whereas an organization by arousal features less prominently. Next, a series of property fitting analyses was as a follow-up to confirm this observation. Figure 2 shows the property fitting as the Pearson correlation of each of the external predictors (valence, arousal, concreteness and word frequency) with each of the dimensions. In all languages, valence was strongly correlated with the first dimension (abstract words) and moderately correlated with the second dimension (concrete words). Significant correlations were also found for concreteness and arousal as well, but the correlations were considerably smaller. As expected, the influence of word frequency was limited: Significant effects were only found for one out of 8 comparisons (Spanish concrete words), where it correlated with D7. Given that the dimensions are ordered in terms of decreasing stress, word frequency only plays a minor role in organizing meaning.

Cross-linguistic connotation differences

The similarity matrices across all three languages were highly correlated: For abstract and concrete words the results using the 3240 word combinations in the upper triangle were respectively: English-Mandarin $r = .75$, $CI_{95}[.73, .76]$; $r = .86$, $CI_{95}[.85, .87]$; English-Spanish $r = .80$, $CI_{95}[.79, .81]$; $r = .90$, $CI_{95}[.89, .90]$; Spanish-Mandarin: $r = .72$, $CI_{95}[.70, .73]$; $r = .86$, $CI_{95}[.85, .87]$. Unsurprisingly, the correlations were higher between the more related languages (English and Spanish) and the concrete concepts, but they are still not perfect. As such, differences in affective connotation may contribute to the strength of correlation. Two separate analyses were used to compare whether differences in meaning *across* languages can be explained by affective connotation. The first analysis uses the unscaled similarity matrix, whereas the second focuses on culture-specific affective connotations on the valence dimension.

Semantic Correspondence Analysis

In this analysis, the full similarity matrix was used to determine a direct measure of how similar the meanings across two languages are. This measure of cross-linguistic *semantic correspondence* was derived by calculating a second-order similarity measure. Each word corresponding to a row in the monolingual 81 by 81 similarity matrices was correlated with the corresponding word (row) in the other two languages. For example, for concrete English-Mandarin pairs, examples of words low correspondence scores *colony* and *acid*, whereas *onion* and *animal* had high correspondence scores. Next, a linear regression model was used to predict whether differences in affective connotation can explain these semantic correspondence scores between English and Spanish and English and Mandarin. The model predictors consisted of difference scores between the human ratings of valence and arousal (see Stimuli section). Previous research also suggests that the agreement between languages should be larger for concrete words (e.g. Van Hell & De Groot, 1998) and forward translation probability might also explain the degree of correspondence between words. Therefore, both variables were added to the set of predictors as well.

A significant effect of valence was found for the abstract word correspondence scores in English-Mandarin (b = -0.096, $t(76)$ = -2.80, p = .007) and English-Spanish (b = -0.058, $t(76)$ = -2.04, p = 0.045), but not for concrete words. No significant effects for arousal were found in any comparison. Additionally, significant effects for English-Mandarin Chinese concreteness were present for both abstract (b =

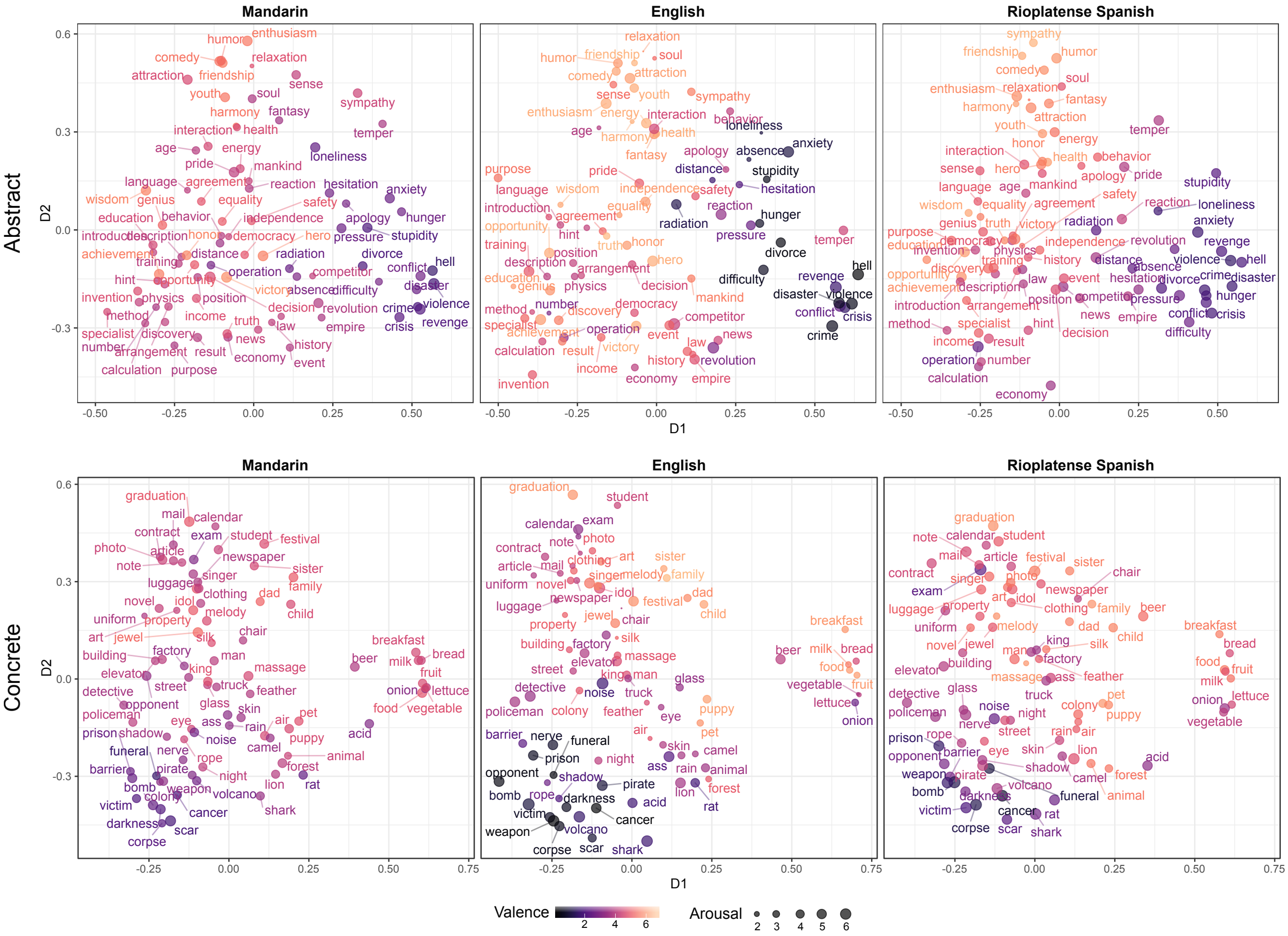

Figure 1: MDS configuration plots for the first two dimensions of abstract words (top panel) and concrete words (bottom panel). Mandarin and Spanish solutions are Procrustes rotation to English dimensions. Words are annotated by language-specific valence and arousal ratings.

0.058. $t(76) = 2.04$. $p = .044$) and concrete concepts ($b = 0.099$, $t(76) = 3.31$, $p = 0.001$) but not in English-Spanish. The role of concreteness partly confirms previous findings (Van Hell & De Groot, 1998) and the cross-linguistic correlations reported in the previous section, showing larger agreement between concrete than abstract concepts. Finally, forward translation probability also captured some variance in one case: English-Spanish ($b = .004$, $t(76) = 3.60$, $p = 0.006$) abstract words.

To increase the robustness of the comparison and determine the relative importance of the regressors, a follow-up analysis was conducted using the *relaimpo* package in R (Groemping, 2007). The *lmg* method was chosen to quantify the contribution of the predictors across models of different sizes in all possible orders. The bootstrapped 95% confidence intervals and the proportional effect-sizes (normalized to sum to 100%) shown in Figure 3. Overall, the results confirm earlier results showing that English-Mandarin valence differences contribute primarily in abstract words.

Predicting culture-specific valence

As illustrated by Figure 1, the coordinates for the abstract words on D1 overlap strongly between English and Spanish, $r = .95$, $CI_{95}[.93, .97]$, and Chinese $r = .93$, $CI_{95}[.90, .96]$. However, some words are outliers. For example, in Figure 1, *temper* (脾气) seems more negative in English than in Mandarin Chinese.

In contrast to the previous analysis, the next analysis aims at determining whether cross-cultural differences in the single best-fitting affective dimension can be explained in terms of differences in direct judgments of this affective dimension from existing ratings for valence or arousal. Here, the focus is on valence, as arousal did not correlate as strongly with any dimension. Valence correlated strongly with the first dimension in abstract concepts and the second dimension in concrete concepts (see Figure 2). To investigate what aspect of culture-specific meaning is due to connotation differences on the same dimension (D1 or D2), semi-partial correlations were calculated. These correlations were derived using the difference in affective norms (valence or arousal) and the coordinates of Mandarin Chinese or Spanish on the first or second dimension after removing the influence of English. This amounts to regressing the English coordinates out of the Spanish or Mandarin Chinese coordinates and correlating the residuals with the valence differences.

The results showed significant semi-partial correlations for

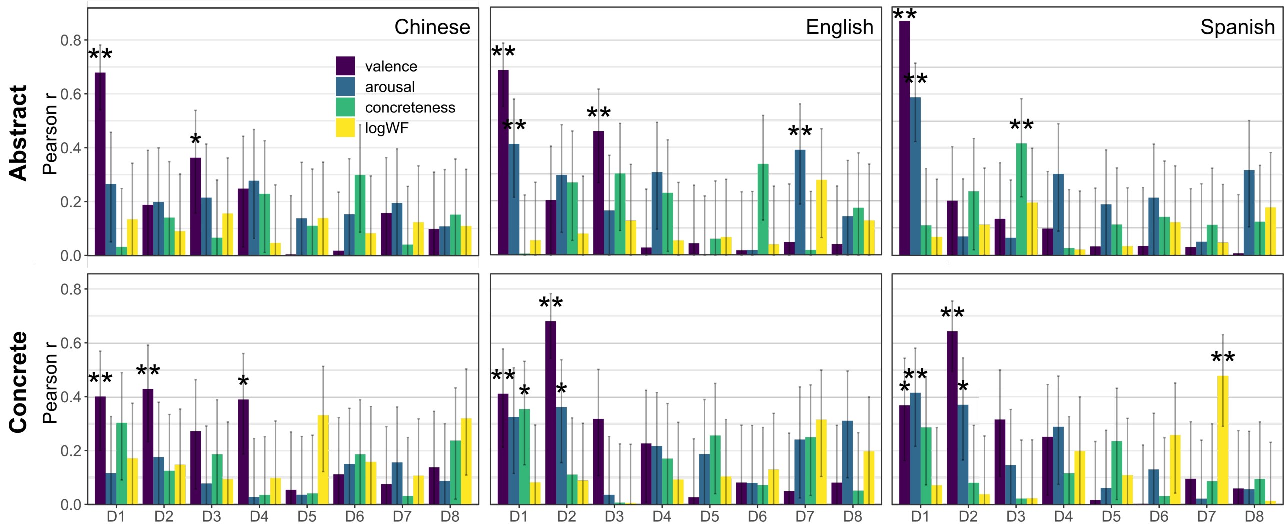

Figure 2: Correlations and 95% confidence intervals between the first Dimensions 1-8 and lexico-semantic variables (valence, arousal, concreteness) for abstract (top) and concrete (bottom) words.

valence norm differences and the D1 coordinates for 81 Mandarin Chinese abstract concepts when controlling English D1, $r = -.37$, $CI_{95}[-.54, -.16]$. For Spanish, valence norm differences did not predict the coordinates when the effect of English was partialed out. For the 81 concrete concepts, only D2 was considered as valence was found strongest correlated for this dimension (cf. Figure 2). A significant effect of valence norm differences in concrete words was found for both Mandarin Chinese, $r = -.24$, $CI_{95}[-.42, -.02]$, and Spanish $r = -.29$, $CI_{95}[-.47, -.07]$. Altogether, these results indicate that differences between the coordinates on the valence dimension for the tested language pairs can be predicted from direct human judgments of a word's valence.

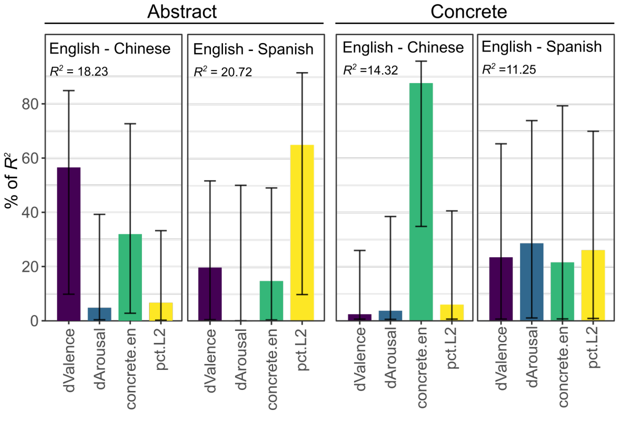

Figure 3: Relative importance and 95% confidence intervals for semantic correspondence scores predicted by differences in valence and arousal, (English) concreteness and English → L2 translation probability (pct.L2).

Does language encode cross-cultural connotation?

To investigate to what extent language, as opposed to psycho-experimental measurements, encodes culture-specific connotations, word embedding models trained to predict word co-occurrence were taken from the multilingual aligned Facebook fastText vectors (Conneau, Lample, Ranzato, Denoyer, & Jégou, 2017). These embeddings are similar to the popular word2vec embeddings (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and further improve them by incorporating sublexical information. This is especially useful for languages like Mandarin Chinese in which characters can be combined into a large number of words (Bojanowski, Grave, Joulin, & Mikolov, 2017).[1] For reasons of space, only the main results are described below. A first analysis aimed at predicting the monolingual relatedness judgments for concrete and abstract words. In each language, human and language-based similarities were correlated for both abstract and concrete words. The data were based on the upper-triangle of the similarity matrix derived from human relatedness ratings and the corresponding cosine-similarities calculated using the 300-dimensional word embeddings. For $n = 3,240$ abstract and concrete word pairs: English: $r = .55$, $CI_{95}[.53, .57]$ and $r = .58$, $CI_{95}[.56, .60]$; Spanish: $r = .48$, $CI_{95}[.45, .50]$ and $r = .51$, $CI_{95}[.49, .54]$; Mandarin: $r = .42$, $CI_{95}[.39, .45]$ and $r = .47$, $CI_{95}[.44, .59]$. Regardless of the language and concreteness, language embeddings only correlated moderately with human judgments.

Next, the role of affective connotation was investigated using interval MDS. Like the previous section, an 8-

[1] The embeddings are trained on a Wikipedia corpus. This might raise concerns about the extent Wikipedia is suited to capture affective connotation. This possibility was investigated using embeddings trained on subtitles (van Paridon & Thompson, 2019) and Chinese embeddings trained on blogs. In all cases, the Wikipedia-based fastText gave the best results.

dimensional solution was derived, leading to solutions with Stress-1 values between 0.09 and 0.10. The first two dimensions, which lead to the largest reduction in stress, did not correlate with valence or arousal for abstract words across all three languages. Only concreteness was significantly correlated with English abstract words on D2, $r = .42$, $CI_{95}[.22, .58]$. For English concrete words, there was a moderate correlation with concreteness on D1, $r = .37$, $CI_{95}[.17, .55]$ and valence on D2, $r = .36$, $CI_{95}[.15, .54]$. For Mandarin Chinese concrete words, there was a moderate correlation with concreteness on D2, $r = .50$, $CI_{95}[.31, .64]$. Altogether, the effect of valence was consistently absent in abstract words and only moderately present in English concrete words in the first two dimensions. Instead, language-based representations organized words mostly in terms of their concreteness. In other words, language-based similarity representations are more in line with an organization in terms of relatedness (ignoring valence). In contrast, a psycho-experimental task that stressed judgments of relatedness clearly distinguishes this dimension, even though no similarity instructions were given.

Discussion

Word meaning measured through human relatedness ratings strongly encoded affective connotations in both concrete and abstract words. This result confirms earlier work using semantic differentials and extends it to a relatedness ranking task, which avoids the limitations of bipolar adjective scales (Osgood et al., 1957). Consistent with cross-cultural work by Osgood et al. (1975), valence, and to lesser extent arousal, were universally found to be the strongest predictors of meaning. Affective connotation also contributed to the meaning of concrete concepts, which suggests that affective connotations contribute to the meaning of a large class of concepts (cf. Kousta et al., 2011).

The second series of analyses sought to determine whether culture-specific aspects of meaning could be directly related to differences in human ratings of affective connotation. Consistent with previous work suggesting systematic connotative differences between Western and East-Asian cultures, a significant difference was found between the meaning of abstract words in Indo-European languages (English and Spanish) and Mandarin Chinese, but not between both Indo-European languages. Moreover, differences in connotations for concrete concepts were not significant, even though such differences were predicted based on English and Mandarin Chinese affective ratings for these words. This could indicate a trade-off, where concrete concepts encode additional sensorial information compared to abstract concepts, which reduces the overall contribution of affective connotations. Although speculative, this interpretation is supported by our finding that in contrast to abstract words, where valence mapped on the first dimension, valence correlated primarily with the second dimension, and the correlation was overall somewhat smaller.

A final analysis aimed to predict relatedness from language and determine whether/if cross-cultural affective connotations derived from human relatedness judgments are encoded in language. The monolingual results using word embeddings showed only moderate correlations with the behavioral representations. The correlations were lower for abstract words, which is striking since these are often assumed to rely more on linguistic and less on sensorimotor properties than concrete concepts (e.g., Van Hell & De Groot, 1998). Perhaps more importantly, in contrast to the behavioral data, relatedness derived from language was not strongly determined by valence. One explanation is that word embeddings do not adequately capture affective connotations or sentiment, which is supported by recent findings that show word embedding models require explicit training for sentiment to encode valence in the embeddings (Young, Hazarika, Poria, & Cambria, 2018).

The current approach has a couple of limitations that might need to be addressed in future research. A first issue that might affect the results is the dependence on the specific response alternatives in the partial ranking task. This could potentially explain the fact that no significant cultural differences were found for affect in concrete concepts and correlations with word embeddings were somewhat lower than those reported in previous studies. Since word embeddings consider all words in the lexicon, the moderate correlations might be due to a task artefact. However, a follow-up analysis that correlated human relatedness judgments with similarities derived from word association vectors in Rioplatense Spanish, Mandarin Chinese and English derived from the Small World of Words project (De Deyne et al., 2019) indicated that contextual effects in relatedness task are likely to be minor. For abstract and concrete words, the correlation with the similarity scores for 3,319 ranked word pairs[2] were $r = .72$, $CI_{95}[.71, .74]$ and $r = .77$, $CI_{95} = [.76, .78]$ for English; $r = .72$, $CI_{95} = [.70, .73]$ and $r = .77$, $CI_{95} = [.76, .78]$ for Spanish and $r = .67$, $CI_{95} = [.65, .69]$ and $r = .78$, $CI_{95} = [.76, .79]$ for Mandarin Chinese. These results make it unlikely that the set of response alternatives determines the moderate language-based correlations, or that low-level language-specific factors for Mandarin Chinese could explain the moderate results for word embedding results in that language. A final potential limitation is that culture-specific connotations are manifested on the arousal dimension, yet valence played a more important role in explaining cultural differences. In part, this can be explained by the fact that the measurement of arousal is often not as reliable as that of valence and a single large-scale resource in Mandarin Chinese was not available to select items from but was compiled from a variety of sources. As such, a more systematic and large-scale effort to obtain Mandarin norms might be needed to determine questions about the role of arousal.

Despite these limitations, the current work also has several implications beyond the connotation of monolingual concepts. In bilinguals, for instance, systematic but subtle connotation differences might explain why emotional concepts

[2] Two words, *competitor* and *absence*, were not present in the Mandarin Chinese word associations

are hard to acquire in adults learning a second language (Pavlenko, 2007). As part of a follow-up study, we are currently addressing such questions by replicating the current findings in bilingual. Second, language-specific meanings are likely to be multi-faceted with connotation only one of them. For example, words might have different meanings not only because they have different connotations, but also because they have different senses, prototypicality, centrality, imagery and so on (cf. Šipka, 2015). Moreover, each of these factors is likely to overlap to some degree, which suggests a large-scale systematic follow is needed to quantify how connotation and meaning more generally might converge or conflict across languages.

Acknowledgments

This work was funded by an ARC DECRA project DE-140101749 awarded to SDD. Our gratitude goes to Meghna Kadalbajoo for help with data collection and ethics preparation, and Andrew Wang for help with the translation study.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Bozzeti-Engstrom, M. L. (2002). What's in a word?: Connotation in teaching english speakers of other languages. *Theses Digitization Project*, *2078*.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Cai, Q., & Brysbaert, M. (2010). SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PloS One*, *5*(6).

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *33*, 133–143.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The Small World of Words English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987-1006.

de Leeuw, J., & Mair, P. (2013). *Multidimensional scaling using majorization: SMACOF in R.* (Vol. 31).

Haspelmath, M. (2012). How to compare major word-classes across the world's languages. *Theories of everything: In honor of Edward Keenan*, *17*, 109–130.

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665-695.

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, *140*, 14.

Lim, N. (2016). Cultural differences in emotion: differences in emotional arousal level between the East and the West. *Integrative Medicine Research*, *5*, 105–109.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics* (pp. 174–184).

Osgood, C. E., May, W. H., Miron, M. S., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning* (Vol. 1). University of Illinois Press.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. University of Illinois Press.

Pavlenko, A. (2007). *Emotions and multilingualism*. Cambridge University Press.

Prior, A., MacWhinney, B., & Kroll, J. F. (2007). Translation norms for English and Spanish: The role of lexical variables, word class, and L2 proficiency in negotiating translation ambiguity. *Behavior Research Methods*, *39*, 1029–1038.

Šipka, D. (2015). *Lexical conflict: Theory and practice*. Cambridge University Press.

Stadthagen-Gonzalez, H., Imbault, C., Sánchez, M. A. P., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*, 111–123.

Van Hell, J. G., & De Groot, A. M. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*, 193–211.

van Paridon, J., & Thompson, B. (2019). subs2vec: Word embeddings from subtitles in 55 languages. *PsyArXiv*.

Wen, Y., & van Heuven, W. J. (2017). Chinese translation norms for 1,429 english words. *Behavior Research Methods*, *49*, 1006–1019.

Yao, Z., Wu, J., Zhang, Y., & Wang, Z. (2017). Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, *49*, 1374–1385.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational intelligence magazine*, *13*, 55–75.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., ... Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 540–545).