Common words, uncommon meanings: Evidence for widespread gender differences in word meaning.

Anonymous CogSci submission

Abstract

Communication relies on a shared understanding of word meaning; however, recent evidence suggests that individual variation in meaning exists even for common nouns. Understanding where and how this variation arises is therefore integral to circumnavigating misunderstandings and facilitating more efficient communication. This study investigated the degree to which men and women ascribe different meanings to the same words. Experiment 1 used a constrained word association task where participants generated three adjectives for each of 42 words. These data were used in Experiment 2, where a separate sample judged the association strength between word pairs. Both experiments investigated the role of gender in word meaning variation and found evidence for gender-specific meaning for a substantial fraction of the 42 words (Experiment 1: 12 or 29%; Experiment 2: 13 or 31%). Experiment 2 also investigated whether conceptual diversity can be explained by gender. Using Gaussian mixture modelling, we found evidence for 62 clusters (indicating concepts), with over 30% of words mapping onto multiple concepts. Evidence for gender-specific concepts was found for nearly half (46%) of the words with multiple clusters. Moreover, gender differences in meaning were not restricted to gender-stereotypical words but included apparently neutral words as well. Altogether, the results demonstrate how male and female speakers of the same language may have slightly different conceptual representations, even of common English nouns.

Keywords: word meaning; concepts; semantic diversity; individual differences; gender

Introduction

Although communication relies on a shared understanding of word meaning, recent research has shown that even common words like "penguin" can refer to different concepts for different people (Martí, Wu, Piantadosi, & Kidd, 2023; Wang & Bi, 2021), leading to misunderstandings and unnecessary conflict. For instance, over the past few years, it has become clear that people have different concepts of *freedom*, as showcased during anti-lockdown and anti-vaccination protests. Conceptual differences may be tied to an individual's values, reflecting one way that individual differences can create different concepts. They also might arise from differences in environment: no two individuals live the same lives, and some researchers have proposed these experiences lead to differences in word meaning (Vivas et al., 2022). However, the causes and degree to which common word meanings differ across individuals are not well understood.

Finding limited differences is expected if word meaning is mostly universal. If speakers are exposed to similar linguistic environments, one could assume that exposure to a sufficiently large amount of language for different speakers leads to strongly convergent concepts for words like *chair* or *freedom*. This might explain why word reading times are largely the same between males and females (Majeres, 1999). It is also consistent with current large-scale models of distributional semantics such as word2vec (Mikolov, Chen, Corrado, & Dean, 2013), where corpora are typically aggregated across sociodemographic factors, this ignores people's unique environments as well as their agency to shape their own language use and experiences.

An additional complexity is that subtle differences in meaning of the sort we are discussing are often hard to detect. Consider gender: men and women are each nearly 50% of the world's population and differ from each other (at least on average) in many ways. One might expect that if individual demographics affects how people conceptualise words, it would be most obvious in the realm of gender. Nevertheless, evidence on gender effects in some studies of word processing and semantic cognition has been mixed. Some research suggests that the emotional valence for most words is similar for men and women; a conservative estimate is that less than 100 words (0.8%) are different between men and women in an extensive English dataset (Warriner, Kuperman, & Brysbaert, 2013). These words are a minority reflecting specific topics such as sexuality, family, taboo, and violence (see also, Heise, 2010). This suggests that gender differences could be relatively marginal and restricted to a small set of domains.

The present study explores the role of sociodemographic factors in meaning by investigating the systematic differences between lexicalized concepts in men and women. To do so, we consider both neutral words and words where gender differences should be expected if the content of our concepts reflects differences in lived experience. Thus, gender can affect word meaning in at least two ways: through the gender of the individual language user or through the perceived gender connotations of the word.

What evidence is there that such gender differences exist? One piece of evidence comes from studies that show structural differences in semantic representations by gender. For example in picture naming tasks with Alzheimer patients, women were more impaired at naming man-made objects, such as vehicles and furniture, and men were more impaired in natural categories like animals and vegetables (Laiacona, Barbarotto, & Capitani, 1998). Similar gender

differences were found in healthy participants, with women having greater semantic knowledge of fruit and men of tools (Capitani, Laiacona, & Barbarotto, 1999; Barbarotto, Laiacona, Macchi, & Capitani, 2002).

Another way gender can influence word meaning is through gender-stereotyped knowledge of words (Kennison & Trofe, 2003). Few English words are gender-specific in the way that words like *brother* and *sister* are, yet norming studies have found that when people are asked directly they reliably attach gender stereotypes to apparently neutral words(Kennison & Trofe, 2003; Scott, Keitel, Becirspahic, Yao, & Sereno, 2019). For example, *bra* is rated as feminine whilst *beard* is masculine (Scott et al., 2019).

Multiple studies have found that gendered word stereotypes are accessed and used during language comprehension (Carreiras, Garnham, Oakhill, & Cain, 1996; Kennison & Trofe, 2003). In these studies, participants were presented with a sentence with a strongly stereotyped word, such as builder for men and nurse for women, followed by a sentence with a gendered pronoun (he or she). When the first word aligned with the pronoun, comprehension was fluid. However, when they were mismatched, comprehension took significantly longer (Carreiras et al., 1996; Kennison & Trofe, 2003). This demonstrates that gender-stereotyped information is relied upon during language processing.

Whilst it is established that gender stereotypes for words exist, few studies have directly investigated their contribution to meaning diversity between women and men. Altogether, little is known about the extent to which women and men have shared or divergent conceptual representations for stereotyped and neutral words. While separate studies provide useful clues, how and to what degree these differences occur among women and men remains unclear.

The Present Study

This study investigates how the representation of lexicalized concepts in English varies between men and women. Experiment 1 uses a constrained word association production task in which participants provide three adjective associates to a list of 42 nouns. A gender-balanced sample of those responses is then used as the stimuli of an association verification task in Experiment 2. This study aims to determine the prevalence of conceptual diversity in neutral and gender-stereotyped words using two different methods: association generation and association ratings. We also aim to investigate the number of concepts represented by the participants by clustering their responses in Experiment 2 (cf. Martí et al., 2023). Is gender an explanatory factor for this conceptual diversity?

Constrained Word Association Task

Method

Participants. A total of 105 first-year undergraduate students (36 male, 68 female, 1 other, mean age = 19.9) at the University of ANONYMIZED received course credits for their participation. Three non-native English speakers (all women) and the non-binary participant were removed from the sample.

Materials and Measures. Cue words consisted of 42 words listed in Appendix A of the supplemental materials. [1] The words were chosen from the Glasgow Norms based on normed gender, concreteness, and valence ratings (Scott et al., 2019). Normed gender was measured on a 7-point rating scale anchored by 1 = Feminine and 7 = Masculine. Concreteness was also measured by a 7-point rating scale, with anchors 1 = Abstract, and 7 = Concrete. Valence was measured on a 9-point scale anchored with 1 = Negative value and 9 = Positive value. The 42 nouns consisted of three groups of 14 words, split based on being feminine, masculine, or neutrally normed words (rated below 3.4 for feminine, between 3.2-4.7 for neutral, and above 4.7 for masculine). Within each group, the 14 words were then divided based on concreteness, with seven words being normed concrete (rated 5.2 and above) and seven normed abstract (rated 4.5 and below). Lastly, the seven concrete and seven abstract words were chosen based on valence, with two words being positive (5.7 and above), three words being neutral (rated between 3.5 and 5.3) and two words being negative (rated 3.5 and below). Words of varying concreteness and valence were included to provide a stimulus set that represents a broad range of concepts. As expected, t-tests revealed that there were significant differences in rated gender association between the feminine, masculine, and neutral words. There were no significant differences between gendered groups in overall concreteness and valence, making the groups balanced, see Table 1.

Table 1: Average gender stereotype, concreteness, and valence rating of the stimuli.

	N	Stereotypy	Concreteness	Valence
Feminine	14	2.52	4.57	4.76
Masculine	14	5.29	4.62	4.81
Neutral	14	3.85	4.60	4.75

Procedure. Participants completed a 30-minute online constrained word association task which asked them to provide three adjective associations per cue for a list of 42 cue words using a procedure similar to De Deyne, Navarro, Perfors, Brysbaert, and Storms (2019). After the word association task, participants completed a shortened version of the LEAP-Q (Marian, Blumenfeld, & Kaushanskaya, 2007) to ascertain their language and cultural background.

Results

Responses were normalized in several ways. Punctuation was removed, and responses were spell-checked and changed to Australian English where applicable. If more than three responses were given, only the first three were retained. Meaning differences between male and female response frequency

[1] https://figshare.com/s/5c3e5e95326e39cd27e2

distributions were determined by calculating the cosine distances between the log-transformed response distributions of both groups. The significance of the distances was determined with a permutation test by randomly assigning gender to cue-response observations. This procedure was repeated 1,000 times. Twelve out of 42 words (29%) had significantly different cosine distances. Three feminine words were significant: *bonnet*, $p < .001$; *bra*, $p = .008$; *insecurity*, $p = .026$; six masculine words: *arrow*, $p = .014$; *glory*, $p = .033$; *oblivion*, $p = .025$; *regime*, $p < .001$, *satire*, $p = .003$ and *tornado*, $p = .005$. In addition, the responses for three neutral words also differed between male and female participants: *desire*, $p = .032$, *scissors*, $p = .041$, *slime*, $p < .001$. Altogether this suggests that gender differences can be observed in constrained word association data, even for words that are not gender-stereotypical.

Judgements of associative strength

While Experiment 1 provides evidence for gender differences among gender-stereotypical words in association generation data, the strongly skewed response frequencies might underestimate how prevalent these differences are if they occur in the mid or tail section of the response distribution. To overcome this limitation typical for language production tasks and to investigate the robustness of the findings, Experiment 2 uses a receptive task rather than a productive one. In this experiment, participants judge the associative strength between the noun cues and a gender-balanced sample of adjective responses taken from Experiment 1.

Method

Participants. A total of 141 participants were recruited at the University of ANONYMIZED ($n = 104$), and through social media ($n = 37$). These participants were second-year undergraduate students who received 2 course credits for their participation; all other participants received an e-gift card valued at $20AUD upon completing the study. Twenty-eight non-native English speakers and 21 participants with low response reliability (see below) were removed. The final sample comprised 88 participants (43 male, 45 female) aged 18 to 54 ($M = 23$).

Materials Responses from Experiment 1 were annotated using the part-of-speech information in SUBTLEX-US database (Brysbaert, New, & Keuleers, 2012) after which the correctness of the part-of-speech was manually double-checked. Words that were not adjectives and responses which were too similar to the cue word (e.g., *glory* and *glorious*) were removed. We extracted a gender-balanced sample of 36 women and 36 men and, for each gender, chose the top 12 associate responses per word as the stimuli for Experiment 2. Aggregating across both genders resulted in 15 to 22 distinct associates per cue.

Procedure. Participants in Experiment 2 completed a 60-minute online session. This consisted of a familiarity rating task in which they rated their familiarity with each of the 42 cue words. Familiarity judgments were based on a 9-point scale, where 1 = very unfamiliar and 9 = very familiar. If a word was unknown, participants were asked to enter 0. The order of words was randomized. After this, participants completed the association verification task, where they were required to rate how much they associated a list of adjectives with each cue word. Judgments of associative strength were made on a scale ranging from *0 = no association* to *100 = very strong association* (see Appendix B, Supplemental Information). Participants were instructed to respond intuitively and using their personal judgment in an absolute sense. Each of the 42 cue words had a specific set of associates to rate. The scale design presented all the stimuli simultaneously, which allowed participants to position items in accordance with others. They were instructed to start with the word which, in their judgment, was the most strongly associated. Further, the dynamic scale allowed participants to zoom and pan with the mouse to accurately position words for nuanced responding (see Appendix B for an example). This way, the benefits of making relative judgments in ranking procedures are combined with the resolution of a continuous scale. The trials and adjectives were randomized. Similar to Experiment 1, participants completed a shortened version of the LEAP-Q (Marian et al., 2007) to check their language background.[2]

Results

Twenty-one participants who correlated $r < .20$ with the mean familiarity ratings or mean associative strength rating were removed from all further analyses. The reliability of the data was calculated on the remaining sample using the Spearman-Brown split-half correlation. Familiarity was highly reliable, $r_{splithalf} = .977$. The average number (percentage) of unknown adjectives was low, 1.64 (0.21%) and 1.75 (0.23%) for respectively females and males. The mean familiarity for the 42 words across both genders, males: 7.29, females: 7.28, was not significantly different, $t(81.9) = -0.036$, $p = .97$. The associative strength judgments were also highly reliable, $r_{splithalf} = .981$. The average reliability for each cue word was $r_{splithalf} = .961$ and .947 for, respectively, female and male participants.

Gender meaning differences Similar to Experiment 1, we compared the response distributions to investigate meaning differences. First, all ratings were standardized using z-scores by participant and cue. Because the distributions consist of continuous ratings on a relatively small number of adjectives, Euclidean distances were used [3]. Significance was again established using a permutation test repeated 1,000 times. A total of 13 out of 42 cues (31%) were significantly different between male and female participants. This included 4 feminine words: *bonnet**, $p < .001$; *bra** $< .001$, *mistress*, $p = .002$ and *pill*, $p = .022$, 5 masculine words: *oblivion**, $p =$

[2] The detailed data will be used in a follow-study to investigate language differences.

[3] This choice is consistent with (Martí et al., 2023) and qualitatively similar results were obtained using a cosine measure.

.04; *regime**, p = .001; *satire**, p = .006, and *tornado**, p = .001 and 4 neutral words: *desire** p = .01; *koala*, p = .009, *slime**, p <.001; and *temptation*, p = .01. Seven of these words, marked with asterisks, were also identified in Experiment 1, despite the use of a different procedure, the sampling of a subset of responses, and the use of a different distance measure.

Similar to Experiment 1, the proportion of cues was primarily gender-stereotyped (9 out of 13). However, given that 4 out of 14 neutral words (29%) were significant compared to 8 out of 28 (32%) gender-stereotyped words, this suggests that meaning differences are not restricted to feminine or masculine words. To illustrate these gender differences, Figure 1 highlights potential differences in meaning for the word *oblivion*. Females associate it more strongly with, *ignorant*, *unaware*, and *unconscious*. In comparison, males associate it more strongly with *cruel*, *destructive*, and *dead*. This could indicate the term is interpreted by men as a state after destruction, whereas women think of being a state of being forgotten or unaware, which is consistent with its dominant dictionary senses (Cambridge University Press, 2023).

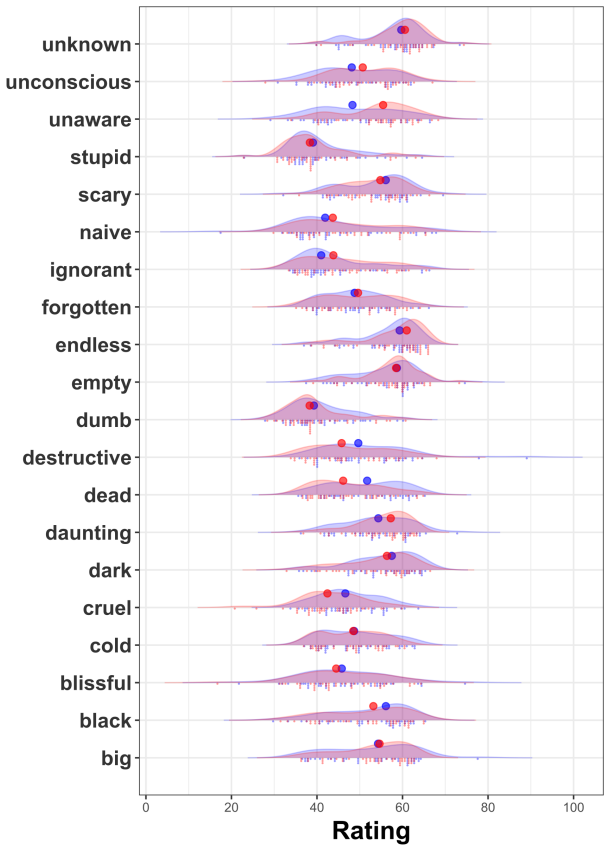

Figure 1: Gender Differences in Adjective Associate Ratings for *oblivion*. Bigger dots represent the mean rating of an associate for each gender. The smaller dots represent the individual data points, and the curves represent their distribution.

One possibility is that the words with gender differences are simply words with a larger set of adjectives. To investigate this possibility, we calculated the correlation between the number of associates and response reliability was low for both females, $r(40)$ = .09, p = .574, CI_{95} = [-.22,-.38], and males, $r(40)$ = -.10, p = .509. CI_{95} = [-.40,-.21]. This suggests that the number of adjectives per cue word did not affect the overall reliability of a cue's association ratings. We also investigated whether semantic distance was related to gender-laden, concreteness, and valence norms taken from the Glasgow dataset. None of the lexical covariates were significantly correlated with the semantic distance between male and female representations: concreteness, $r(40)$ = -.20, CI_{95} = [-.47, .11], valence, $r(40)$ = .05, CI_{95} = [-.25, .35].

Conceptual diversity While the previous results suggest that many words can be interpreted differently depending on gender, people with the same gender might nonetheless have different concepts for a word. For example, in recent work using an adjective verification task Martí et al. (2023) found 5 to 8 distinct concepts for words referring to animals in a similar two-step procedure of adjective generation and verification. To determine to what degree gender can explain this conceptual diversity, we used a similar analysis strategy as Martí et al. (2023) and measured conceptual diversity as the number of clusters by grouping participants with similar ratings on each cue-adjective pair.

A Gaussian Mixture Model was used to determine the amount of concept variation through clustering the associative strength rating vectors for each participant using the *mclust 5* R package (Scrucca, Fop, Murphy, & Raftery, 2016). In contrast to other methods (e.g., k-means clustering, see Scrucca et al., 2016), this model-based approach provides a way to determine the number of clusters automatically and to detect evidence for a single cluster. For ease of interpretation, we balanced the proportion of males and females to be equal to 43 in each group by randomly removing 2 females. Prior to the analysis, "Unknown" responses were replaced with the mean rating of that associate for each participant. For each of the obtained clusters, we determined the role of gender by applying a Bayesian Binomial test to determine whether the number of males and females was equal within each cluster. This allowed us to determine whether clusters showed evidence of gender-specific meaning.

Figure 2 shows the clusters in a 2D space obtained after Multidimensional Scaling (MDS). The number of clusters varied between 1 and 4 and the average number of clusters was 1.48. Twenty-nine words are mapped onto a single cluster. Seven words are mapped onto 2 clusters, 5 words onto 3 concepts, and 1 word onto 4 clusters (see Figure 2). The Bayes Factors of the binomial tests showed evidence (i.e. BF >1) for gender-specific meaning for 26% of the cue words, with gender-specific clusters for 6 out of 62 clusters (11.36%). Among the 13 words with multiple clusters, gender was significant in 46% of cases. This included one cluster with moderate evidence: *bonnet*, BF = 8.87, and 5 clusters with anecdotal evidence: *bra*, *mistress* 1, *regime*, *satire* and *trophy*. With the exception of *trophy*, these are the same

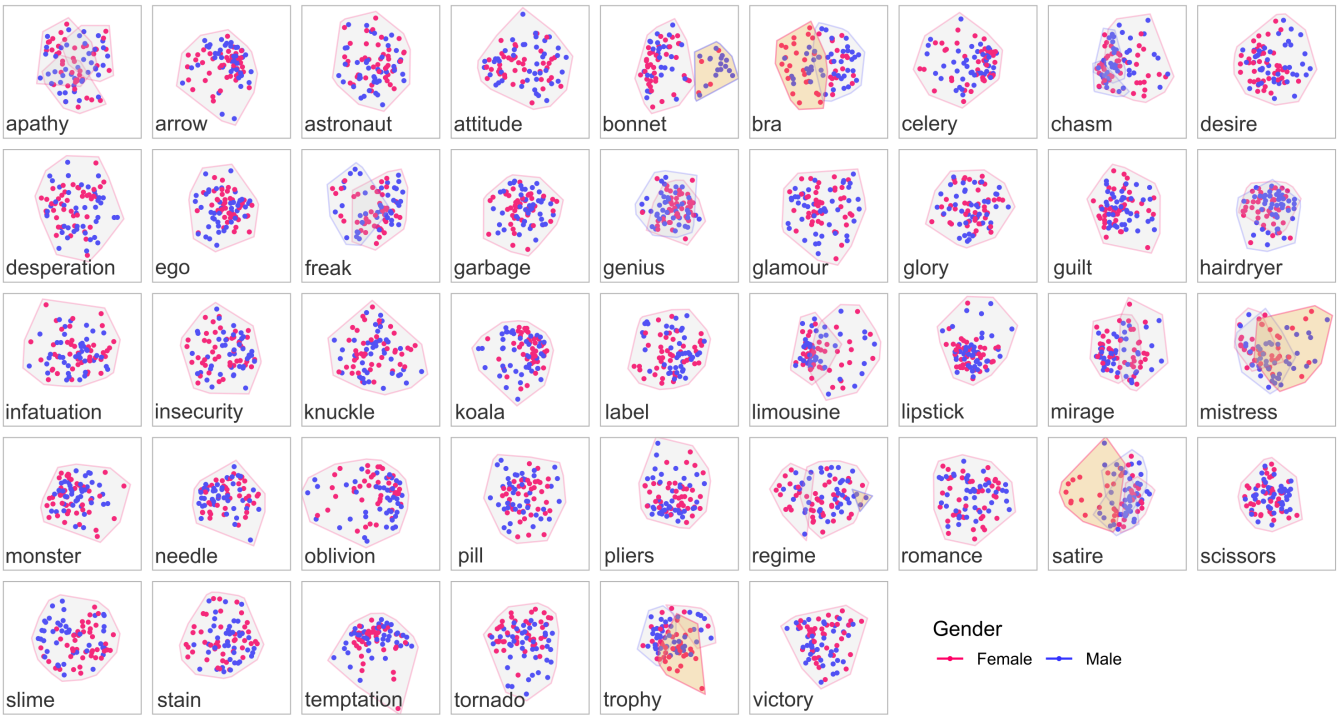

Figure 2: Distribution of men and women in the clusters of associate ratings for 42 cue words. Dots represent individuals within the clusters. Cluster outlines reflect the proportion of gender with red outlines indicating a female-dominant cluster and blue for male-dominant. Clusters with $BF > 1$ are highlighted.

words found in the permutation test of Experiment 2. Out of the 6 clusters, 3 were masculine (*regime*, *satire*, *regime* and 3 were feminine (*bonnet*,*bra*, *mistress*). To illustrate the meaning of the clusters, Figure 3 shows the mean individual ratings of *mistress* by plotting the average ratings for the participants in each cluster, thus providing insight into the concepts of different groups of individuals. For example, participants in the male-dominated cluster 1 gave relatively higher ratings to *strong*, *powerful*, *pretty*, whereas the female-dominated cluster 3 give higher ratings to negative properties such as *callous*, *bad* and *unfaithful*. The second cluster sits somewhere in the middle with less outspoken ratings for these affective adjectives, but higher ratings for more specific terms such as *promiscuous* and *seductive*.

Discussion

The present study investigated how gender affects our understanding of common English words across two experiments. We found gender-specific meaning for 29% and 31% of cue words in respectively Experiment 1 and 2. We expected gender-stereotypical words to exhibit pronounced differences as one gender may have different life experiences and connotations attached to the word. However, this hypothesis was only partially supported as differences were found for neutral words in both Experiments 1 and 2 as well. The same results also showed anecdotal evidence that gender differences were distributed across words that varied in concreteness and emotional valence. All this suggests that subtle gender differences in the conceptual representation of common words may be more widespread than previously assumed (Heise, 2010; Warriner et al., 2013).

We also investigated whether different concepts associated with the same word are gender-specific. This allows us to find out to what degree gender can explain conceptual diversity. Conceptual diversity for words was measured by clustering participants' adjective judgments similarly to Martí et al. (2023). In their study, they found that five to eight different concepts existed for common, concrete nouns. It was hypothesized that there would be multiple concepts for each 42 cue words, and gender-specific concepts would be predominantly found for gender-laden words. The first part of this hypothesis was supported as we found 62 clusters for our 42 cues, with an average of 1.46 clusters per word. This replicates Martí et al. (2023)'s finding as we found multiple concepts per word, although the total number of concepts was lower. The second part of the hypothesis was also supported: we found 6 clusters with gender-specific meanings among the 13 words with multiple concepts. This suggests that distinctions between concepts can be explained by the gender of the participant.

The clustering results show that multiple concepts exist for common nouns, consistent with Martí et al. (2023). However, in contrast to Martí et al. (2023), the average of 1.48 clusters was considerably smaller than their findings using a similar task where the number varied between 5 to 8 for animals and 7 to 12 for politicians. While this difference

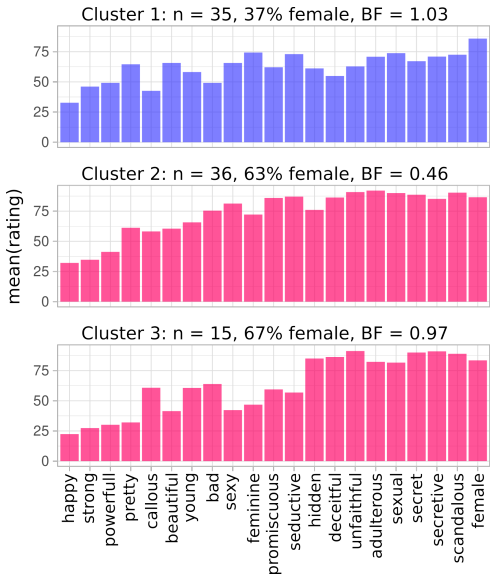

Figure 3: Cluster interpretation for *mistress*. Cluster 1 ($BF > 1$) is mostly a male interpretation, whereas clusters 2 and 3 correspond to female interpretations.

might be related to the specific domain, methodological factors might also contribute to the smaller number of clusters. For example, standardizing the ratings reduces the number of clusters, assuming different participants use the rating scale differently. However, it is also possible that the absolute ratings reflect something about the meaning as well. A second difference is that the current task only presents related adjectives, whereas the study by Martí et al. pools adjectives across a broad domain (animals). Not only does the latter lead to a larger set of in total 105 adjectives, but the majority of these adjectives are also only somewhat related to the target noun (e.g., *whale - melodious*). As a consequence, the presence of a large proportion of fairly unrelated adjectives might inflate the number of distinct concepts. A second difference is the fact that Martí et al. (2023) recruited over 500 participants per domain, whereas the current study had 141 participants. Within the samples of both studies, there might be differences in the homogeneity of the participants, as ours recruited primarily psychology students, and Martí et al. (2023) recruited participants through Prolific (an online platform). This could also lead to an inflation of the number of concepts in their study. While more work is needed to determine what affects the absolute number of clusters or concepts, the current findings do support the idea of a high degree of conceptual diversity, even among common words.

Reassuringly, the adjective judgments in Experiment 2 using a newly developed dynamic scale had high reliability. However, the current study is somewhat limited by the number of words that were tested. The current procedure of having two participant samples generating and judging word-association pairs requires a large number of participants. In terms of stimulus items, a similar limitation related to sample size is manifest in two ways: the associate set gathered from Experiment 1 and the number of ratings in Experiment 2. While the number of associates did not impact the reliability of associate ratings, it may have limited the number of clusters found. This may be because of the lack of distinct associations in Experiment 2. For example, *arrow* only had 15 associates to rate, which does not allow for much differentiation. While a small number of response types might already capture the core meaning, future work should therefore not only extend the set of cue words but also the set of generated responses and vary the total number of association types in strength judgments to determine a saturation point at which additional words do not differentiate meaning substantially.

Finally, the current results do not directly address the question of whether participants can take different stances when it comes to gender stereotypes for words and concepts. Following the principle of least cognitive effort and consistent with the current findings, an egocentric viewpoint is likely in this study, but future work might consider an experimental manipulation in which participants are asked to take an allocentric position (from the other gender's position) to see if this information is encoded.

Conclusion

Across two experiments with different methods, we found a consistent pattern of widespread gender-specific meaning. While this includes obvious cases for gender-stereotyped words such as *bra*, gender differences were also found for words that were not gender-stereotyped (e.g. *koala*). Moreover, the widespread nature of these differences has implications for semantic cognition research more generally as it increasingly relies on large language models. Such models are typically trained on texts for which little demographic information is available. Our work questions a shared language experience across genders, although it seems likely that other, non-linguistic factors, such as lived experience, might also drive conceptual diversity. Likewise, while there's increasing awareness of bias differences between languages (e.g., Lewis & Lupyan, 2020), more work is needed to determine biases among different groups of people speaking the same language. This is where improved methods to measure meaning in a demographic-aware fashion could supplement language-based approaches. Second, access to online crowds allows us to scale up studies significantly, but here as well, there's a risk when not all groups are equally represented or when the participant homogeneity is difficult to ascertain. Our study demonstrates that when sample characteristics are carefully checked, methods reliably converge in identifying such differences.

Finally, our extension of Martí et al. (2023)'s clustering analysis provides a novel account to identify factors that contribute to meaning diversity leading to a better understanding of the mechanisms behind concept diversity. The novelty of this account and current results can hopefully motivate new research directions for, and insights into, conceptual diversity.

References

Barbarotto, R., Laiacona, M., Macchi, V., & Capitani, E. (2002). Picture reality decision, semantic categories and gender: A new set of pictures, with norms and an experimental study. *Neuropsychologia*, *40*(10), 1637–1653.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the subtlex-us word frequencies. *Behavior Research Methods*, *44*(4), 991–997.

Cambridge University Press. (2023). Oblivion. In *Cambridge Dictionary.* Retrieved from `https://dictionary.cambridge.org/dictionary/english/oblivion`

Capitani, E., Laiacona, M., & Barbarotto, R. (1999). Gender affects word retrieval of certain categories in semantic fluency tasks. *Cortex*, *35*(2), 273–278.

Carreiras, M., Garnham, A., Oakhill, J., & Cain, K. (1996). The use of stereotypical gender information in constructing a mental model: Evidence from english and spanish. *The Quarterly Journal of Experimental Psychology Section A*, *49*, 639–663.

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.

Heise, D. R. (2010). *Surveying cultures: Discovering shared conceptions and sentiments*. John Wiley & Sons.

Kennison, S. M., & Trofe, J. L. (2003). Comprehending pronouns: A role for word-specific gender stereotype information. *Journal of Psycholinguistic Research*, *32*(3), 355–378.

Laiacona, M., Barbarotto, R., & Capitani, E. (1998). Semantic category dissociations in naming: is there a gender effect in alzheimer's disease? *Neuropsychologia*, *36*(5), 407–419.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, *4*(10), 1021–1028.

Majeres, R. L. (1999). Sex differences in phonological processes: Speeded matching and word reading. *Memory & Cognition*, *27*, 246–253.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, 940–967.

Martí, L., Wu, S., Piantadosi, S. T., & Kidd, C. (2023). Latent diversity in human concepts. *Open Mind*.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, *51*(3), 1258–1270.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, *8*(1), 289.

Vivas, L., Yerro, M., Romanelli, S., García Coni, A., Comesaña, A., Lizarralde, F., … Vivas, J. (2022). New spanish semantic feature production norms for older adults. *Behavior Research Methods*, *54*(2), 970–986.

Wang, X., & Bi, Y. (2021). Idiosyncratic tower of babel: Individual differences in word-meaning representation increase as word abstractness increases. *Psychological Science*, *32*(10), 1617–1635.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207.