Power, gender, and individual differences in spatial metaphor

Winter, Bodo; Duffy, Sarah; Littlemore, Jeannette

DOI:
10.1080/10926488.2020.1794319

Power, gender, and individual differences in spatial metaphor:

The role of perceptual stereotypes and language statistics

Bodo Winter[1*], Sarah E. Duffy[2], Jeannette Littlemore[1]

[1] University of Birmingham, Dept. of English Language & Linguistics

[2] Northumbria University, Dept. of Humanities

* corresponding author

10,410 words (including references)

4 figures

Key words: metaphor comprehension; corpus analysis; primary metaphor;

vocational preferences; social role theory

Abstract

English speakers use vertical language to talk about power, such as when speaking of people being "at the bottom of the social hierarchy" or "rising to the top". Experimental research has shown that people automatically associate higher spatial positions with more powerful social groups, such as doctors and army generals, compared to lower spatial positions, which are associated with relatively less powerful groups, such as nurses and soldiers. However, power as a social dimension is also associated with gender. Here, by means of a reaction-time study and a corpus study, we show that professions that display greater gender asymmetries, such as doctor/nurse, exhibit stronger vertical associations. Moreover, we show that people's perception of vertical metaphors for power depends on their own gender, with male participants having stronger vertical biases than female participants, which we propose is due to the fact that men are more prone to thinking about power in bodily terms, as being associated with physical dominance. Our results provide clear evidence for individual differences in metaphor comprehension, thus demonstrating empirically that the same metaphor is understood differently by different people.

Acknowledgements

We would like to thank Alex Dixon for his assistance with the corpus-based study, Frazer Heritage for his help with an earlier version of the study, and the anonymous reviewers for their insightful comments on the manuscript.

1. Introduction

Metaphor pervades thinking about our social lives (Gibbs, 2011; Landau, 2017; Landau et al., 2010; Winter et al., 2018), such as our relations to other people and our positions within social hierarchies. One of the many metaphors that is used for thinking about social relations is the metaphor POWER IS UP or SUBMISSION IS DOWN (Dancygier & Sweetser, 2014; Lakoff & Johnson, 1980; Littlemore, 2019), as evidenced by the following expressions from the British National Corpus (BNC):

The people at the *bottom* end of society see people at the *top* end indulging

He is at the *height* of his powers

I'm right at the *peak* of my career

This study investigates POWER IS UP metaphors in English via an experiment and a corpus-based study. Together, these two studies show that gender is an important component in the understanding of POWER IS UP metaphors. First, behavioral responses to vertically presented power asymmetries (e.g., "doctor" shown to be on top of "nurse") differ as a function of the participant's gender. Second, behavioral responses differ with respect to the gendered nature of profession terms, with stronger power effects observed for professions that exhibit strong gender asymmetries in the corpus. Thus, our results show that POWER IS UP metaphors are understood differently by different people and in different linguistic contexts, pointing to a view of metaphor that is less monolithic and more context-dependent.

2. Background

POWER IS UP can be classified as a primary metaphor (Grady, 1997), that is, a conceptual metaphor that derives from repeatedly experiencing a correlation between the source and target domain, such as, in this case, the correlation between physical height and physical strength. For humans and other animals, physical stature is critical for negotiating power dynamics, particularly during childhood and adolescence. Children are faced with taller parents and older siblings, among others, who are able to assert power over them, thereby further reinforcing the mental association between power and size. Schwartz et al. (1982) demonstrate that elevation is one of the strongest nonverbal cues of social dominance; for example, standing on a platform increases a person's apparent status. Research has also shown that taller people tend be more esteemed by others, they tend to have higher average salaries, and they are more likely to ascend into leadership positions (for a review, see Judge & Cable, 2004). In fact, US presidents that are generally considered "great" are actually taller than presidents considered "failures" (Young & French, 1996).

The correlation between power and vertical position is also evident in other domains. For instance, athletes being awarded the gold medal stand on the highest platform of the victory rostrum, managing directors often occupy offices on the top floors of their companies, and priests leading sermons generally stand on elevated podiums. The association between verticality and power is also used for expressive purposes in various types of media, such as when films depict a powerful villain or threatening monster from a low camera angle to make them appear taller (Winter, 2014), or when high-angle shots (viewing actors from above, thus making them appear smaller) are used to indicate that characters are losing control (Ortiz, 2011).

Experimental research shows that people automatically associate verticality with power. Schubert (2005) asked participants to make rapid judgments about pairs of relatively powerful and powerless groups, such as employer—employee, coach—athlete, and teacher—pupil. Importantly, one of the group labels was presented on top of the other label. People were more quickly to judge which one of the two groups was more powerful when the vertical orientation was congruent with the POWER IS UP metaphor, that is, with the powerful group being on top. Giessner and Schubert (2007) showed that when participants viewed an organization chart that was more vertically extended, participants judged leaders to be more powerful. In another experiment, they showed that information about a leader's power influenced the vertical position where participants placed the leader in the organization chart. Lakens et al. (2011) furthermore showed that it is relative spatial contrast and relative power differences that matter in this association, rather than absolute differences.

However, much of this research can be characterized as falling prey to what Dąbrowska (2016) has identified as one of the "deadly sins" of cognitive linguistics, which is to ignore individual differences (see also Littlemore, 2019). In their original work on conceptual metaphors, Lakoff and Johnson (1980) often refer to the '*prototypical* person' (p. 132), which gives rise to the question of what exactly a prototypical person is. Indeed, in the context of Western society, 'normal' is conventionally typified as adult, white, male, middle-class, heterosexual, right-handed, physically-, and mentally-abled individuals (Littlemore, 2019); yet, many people do not fit this profile. If we aim to develop a fuller understanding of the ways in which metaphors are used and understood, we need to give closer consideration to individual differences in the ways in which people use embodied experiences to reason metaphorically.

There is already existing evidence that metaphors are understood differently by different people. Individual differences in personality traits influence people's understanding of time metaphors (Duffy & Feist, 2014), and being a musician influences one's understanding of music metaphors (Julich, 2019). Other research has found that people's propensity to use metaphor in everyday interactions is associated with their susceptibility to psycholinguistic effects in lab-based studies of metaphor (Fetterman et al., 2016). Specifically with respect to POWER IS UP, it has been found that people who think of themselves as more dominant were faster to respond to higher spatial probes, compared to people who think of themselves as more submissive (Robinson et al., 2008). Together, these results suggest that different people can understand the same metaphor in different ways.

When it comes to power, one particularly relevant individual difference is gender.[1] Psychologists and social scientists have long since noted that power and

[1] We explicitly acknowledge that gender is a non-binary category that is characterized by an underlying continuum. However, for the purposes of this paper only, the binary difference between men and women is the most relevant. This is for four reasons. First, a lot of the research that is directly relevant to our study focuses on the binary distinction between men and women (e.g., Misersky et al., 2014). Second, while gender is actually an underlying continuum, it is often conceptualized in a binary way

gender are not independent social dimensions (e.g., Carli, 1999; Chen et al., 2009; Eagly et al., 2000; Rudman & Glick, 2012). Compared to women, men describe themselves as more assertive and they report to have higher self-esteem than women (Feingold, 1994). Men are more likely to view aggression and bodily force as instrumental to gaining influence and power than women—for whom bodily force may often be associated with a breakdown of self-control and a loss of power (Campbell et al., 1992). Consistent with a different stance towards aggression and physical dominance, Schubert (2004) found that implicitly encouraging men to make a fist heightened their sense of self-control; much more so than for women. This suggests that merely performing a gesture can trigger a person's gender-specific perceptions of power. These mental associations between power and gender correspond to real-world power disparities, such as the fact that in a large number of countries, women tend to occupy lower-status positions and earn less than men (e.g., Blau & Kahn, 1992, 1996).

Schubert's (2005) study of the automatic activation of vertical power metaphors discussed above included stimuli such as doctor, nurse, and general, all of which are highly gendered concepts. It is thus possible that the vertical effects for power observed in this study are confounded with, or moderated by, gender. This is

---

(for metaphorical conceptualizations of the gender binary, see Lederer, 2015). Third, to assess the association with gender experimentally, we need to constrain ourselves to a dimension that can be categorized, so as to establish experimental conditions (a male condition and a female condition). Fourth, all of our participants identified as either male or female, even though they had other response options. Future research can explore to what extent non-binary ways of thinking about gender could interact with vertical power metaphors.

especially likely given the fact that a large number of Schubert's stimuli include labels for professions, and it is known that women and men often have different vocational preferences (Berings & Adriaenssens, 2012; Ferriman et al., 2009; Garnham et al., 2015; Lippa, 1998; Teig & Susskind, 2008). In addition to actual differences in vocational preferences, both men and women also have stereotypes about which activities or professions are considered more male or female (Garnham et al., 2015; Misersky et al., 2014; Ramaci et al., 2017). Some of these gender stereotypes relating to professions arise early in childhood (e.g., Garrett et al., 1977).

Within social psychology, a seminal theory that attempts to account for gender stereotypes relating to professions is social role theory. This theory stipulates that the beliefs people hold about gender are underpinned by the observed social role performances of men and women in relation to their personal qualities (Eagly, 1987; Eagly et al., 2000). In this regard, expectations about women and men reflect the sexual division of labor and the gender hierarchy of society: in general, men are more likely to be employed, particularly in positions of authority, while women are more likely to occupy caretaking roles at home and in employment (Eagly et al., 2000). A key premise in social role theory is that beliefs about gender derive from communal and agentic attributes (Eagly, 1987). Communal attributes, such as affection, sensitivity, and sympathy are more strongly ascribed to women, while agentic attributes, such as dominance, ambition, and aggression are more strongly ascribed to men. The fact that communal and agentic attributes are differentially ascribed to men and women explains two types of prejudice that are shown towards females in positions of power

(Eagly & Karau, 2002). The first type of prejudice is that women are less likely than men to be viewed as good potential leaders because of a perceived inconsistency between the primarily 'communal' qualities that are associated with women and the agentic characteristics that are believed to be necessary in order to become a successful leader. The second type of prejudice is that, when women are in leadership positions, evaluations of their performance in the role tend to be less favorable than those of men; successful female leaders tend to be less liked than equivalently successful men (Heilman et al., 2004).

Consistent with social role theory, research has shown that gender constitutes one of the strongest basis for categorizing people (Fiske et al., 1991; Stangor et al., 1992; van Knippenberg et al., 1994). Stereotypes about women and men are also automatically activated; they have the potential to inhibit perceptual processing by influencing the amount of perceptual information that is encoded. Because stereotypes provide perceivers with a rich background of information about social groups, people do not pay as much attention to social group members when they can rely on social schemas, or stereotypes, compared to when they cannot (Macrae et al., 1994). For instance, Banaji and Hardin (1996) conducted a reaction time experiment showing that stereotypes about gender influence processing even if gender is a task-irrelevant dimension, suggesting that unconscious processes in beliefs about gender can operate automatically in judgment. Taken together with the fact that there are gender stereotypes about professions, the automaticity of stereotype activation suggests that the vertical metaphorical power associations in Schubert's (2005) task were potentially driven by gender, or at least influenced by gender connotations of the profession labels used in this study.

In addition to gender stereotypes, language users may also draw on another source of information to infer a person's gender when it is not explicitly mentioned. Specifically, *masculine generics* are masculine forms that are used to refer to people of unknown or unspecified gender or to groups of people of mixed gender (e.g., Braun et al., 2005; Gabriel & Mellenberger, 2004; Hamilton, 1988). A consequence of the tendency to employ masculine forms to refer to people in general (e.g., "to each *his* own"), is that when the generic masculine is used, people are overwhelmingly more likely to think of men than women, to the extent that even professions typically perceived as female, such as beautician, may be interpreted as male (Gygax et al., 2008; for review, see Stahlberg et al., 2007).[2] On the other hand, avoiding masculine generics and explicitly marking roles for gender by means of feminine-masculine word pairs (e.g., *Geschäftsführerin/Geschäftsführer*, 'CEO, fem./CEO, masc.') can serve to reduce the male reading of the role and can help to avoid male bias in employment contexts (Gabriel et al., 2008; Horvath et al., 2016; Horvath & Sczesny, 2016).

A limitation of previous work on metaphor is that, with the exception of a small number of papers, few attempts have been made to investigate the extent to which the reaction time behavior reflects distributions within language corpora. One exception

[2] This pattern is particularly pronounced among speakers of languages with grammatical gender, such as German, which explicitly represents gender via determiners (e.g., *die*; *der*) and morphological variants of nouns (e.g., *Anwalt*; *Anwältin* 'lawyer (m/f)') (cf. Gygax, Gabriel, Sarrasin, Oakhill and Garnham 2008).

to this is Hutchinson and Louwerse's (2013) study in which they compared the ways in which male and female participants responded to primary metaphors in a reaction time study. They showed that vertical associations in line with the HAPPY IS UP metaphor were more driven by corpus-based frequency differences (e.g., how often 'happy-sad' was mentioned as opposed to 'sad-happy') for women than for men. The authors explain this result in terms of the presumed greater general and spoken language ability than males. Their study thus not only shows a difference in metaphor processing (based on gender), but also that corpus frequencies can drive behavioral results, thus underscoring the importance of looking at naturally occurring language in the context of reaction time studies designed to test people's representations of primary metaphors.

The present study is composed of two parts. Firstly, we conducted a replication-extension of Schubert's (2005) study, adding a gender dimension to the task. To do this, we explicitly encoded gender in the stimuli by employing feminine-masculine word pairs in a male or a female linguistic context, such as "male doctor — female nurse" versus "female doctor — male nurse". Our aim was to replicate the finding that participants responded faster when the more powerful group was presented on top of the less powerful group. In addition, we set out to assess whether there was an association between gender and vertical position as well (with either male or female labels being processed more quickly when presented on top), and if the factor of gender interacted with the factor of power. In the second part of the study, we aimed to control for the gendered nature of particular professions, and the usage-based knowledge that participants may have acquired about gender roles, by creating a new corpus-based measure of gender stereotypes for professions, which we correlated with the results from our behavioral study. Our study addresses the topic of individual differences in a two-fold way. The behavioral experiment demonstrates individual differences across participants (based on gender). The corpus study demonstrates differences with respect to linguistic items, and that gender asymmetries based on language statistics carry over to behavioral tasks. From a methodological perspective, our results furthermore demonstrate that corpus analysis can fruitfully be combined with experiments in metaphor research.

3. Methods

3.1. Participants

There were 63 participants (51 women, 11 men), all of whom were undergraduate students in English Language at the University of Birmingham, UK. The average age of participants was 18.9 (range: 18-21).

3.2. Stimuli

We created a list of 64 profession terms, starting with a list of professions drawn from Schubert's (2005) group label terms, and extending this with a list of professions given by Misersky et al. (2014). Our final set of stimuli includes 32 pairs of professions that differ in power, such as "doctor — nurse", "referee — footballer", and "officer — cadet" (see Appendix A).

3.3. Procedure

Following Schubert (2005), participants were asked either one of two questions: which one of the two groups is the more powerful group, or, which one the group is the *less* powerful group (between-participants factor). Each item was displayed on a white computer screen either in a metaphor-congruent position (powerful on top) or metaphor-incongruent position (less powerful on top) (within-participants factor). The distance between the top and the bottom group label was 5cm on the screen. Participants had to press the UP arrow or the DOWN arrow on the keyboard to indicate which of the two groups was more/less powerful.

Each item was presented either with the more powerful group in a “female” linguistic context (“female doctor — male nurse”), or with the less powerful group in a “female” linguistic context (“male doctor — female nurse”) (within-participants factor). Thus, each of the 32 item pairs was shown a total of 4 times (2 power positions X 2 genders), resulting in a total of 128 stimuli.

Throughout the experiment, there were 24 comprehension questions inserted at random points asking whether either the male or the female was at the top. These comprehension questions were included because the information about gender was task-irrelevant, which means that participants could easily solve the task without paying attention to the male/female labels. The list of 128 stimuli was separated into two blocks of 64 stimuli each. Participants were allowed to take a short break in between the two blocks.

Each trial was presented with a fixation cross which was displayed for 750ms and followed by a white screen (inter-trial-interval) of 500ms. After consenting to the study, participants performed 6 practice trials that included 2 gender comprehension questions. Participants were instructed to respond as accurately as possible while also responding as quickly as possible. Participants were instructed to only use one finger, their index finger. The experiment lasted approximately 15 minutes.

3.4. Statistical analysis

All statistical analyses were conducted with R version 3.6.0 (R Core Team, 2019) and the packages tidyverse version 1.2.1 and ggrepel version 0.8.1 for data processing and data visualization (Slowikowski, 2019; Wickham, 2017). The package brms version 2.9.0. (Bürkner, 2017) was used to perform Bayesian regression analysis (see details below). All data and code is available under the following publicly accessible Open Science Framework repository: https://osf.io/mbv24/

Data from one participant was excluded as this participant did not appear to have solved the task at all (response accuracy was exactly at chance level, 50%, and response times were around 500ms, much below the response times of all other participants, which were >1000ms). Thus, the final analysis is based on a total of 62 participants (51 women, 11 men). In addition, we excluded individual trials that were 2 standard deviations away from each participant's response time mean. We furthermore excluded response times above the 95th percentile of the overall response time distribution (3,949ms). Finally, response times were log-transformed using the natural logarithm.

The log response times were analyzed with a Bayesian linear mixed effects regression. A linear mixed effects model is needed to adequately analyze the data in this case because the experimental design has multiple data points per participants (repeated measures), as well as multiple data points for each item. A mixed model allows modeling such multiple sources of variation via so-called random effects (one random effect for participant, another random effect for item). For an applied introduction to linear mixed effects modeling with R for linguistic data, see Winter (2013).

Mixed models can be embedded in a 'traditional' null hypothesis significance testing framework, or within a Bayesian analysis framework.[3] The latter facilitates interpreting the evidence for a hypothesis in a continuous manner (rather than amassing indirect evidence against null hypotheses and relying on hard cut-offs, such as $p < 0.05$). Bayesian models furthermore facilitate the interpretation of null results (Dienes, 2014), and they have several advantages for model fitting (e.g., the

[3] Technical details on the model fitting process: Models were estimated with Hamiltonian MCMC and 4 chains, each of which was run for 4,000 iterations (2,000 warmup samples), yielding a total of 8,000 posterior samples. There were no indications of any major convergence issues (no divergent transitions and all Rhat values < 1.02). We used weakly informative / regularizing priors for all regression slopes, specifically, normal distributions centered at zero with standard deviations of 0.2. Posterior predictive checks indicated some mismatch between the data and the posterior predictions. Running a comparable ex-Gaussian or Gamma model did not lead to better posterior predictive checks, and there also was no difference whether the cleaned or raw response times (no exclusion) were used. All fixed effects predictors (which are all binary) were deviation-coded, with the female participant and female stimulus conditions as the reference level (-0.5), compared to the male participant and the male stimulus condition (+0.5). This coding scheme was used to aid the interpretation of main effects in the presence of interactions.

ability to estimate effects more conservatively via the use of weakly informative priors). For a conceptual introduction to Bayesian analysis, see Dienes (2008). For a helpful tutorial on analyzing linguistic data with Bayesian mixed models, see Franke and Roettger (2019). For readers unfamiliar with Bayesian techniques, we recommend focusing on the 95% credible intervals: although there are important conceptual differences, a 95% CI that does not include zero can be interpreted as being roughly analogous to a 'significant' result. See Morey and colleagues (2016) for a detailed conceptual discussion of credible intervals in comparison to confidence intervals.

Our main response, the logarithm of response times was modeled as a function of the following fixed effects predictors:

- **Task:** "more powerful?" versus "less powerful?" (between)
- **Participant Gender:** male versus female (between)
- **Hierarchy Condition:** powerful = up versus powerful = down (within)
- **Gender Condition:** male = up versus male = down (within)

In addition, we included the following fixed effects two-way and three-way interaction terms:

- **Task * Hierarchy Condition**
- **Task * Gender Condition**

- **Hierarchy Condition * Gender Condition**
- **Participant Gender * Hierarchy Condition**
- **Participant Gender * Gender Condition**
- **Participant Gender * Hierarchy Condition * Gender Condition**

These interaction terms describe the degree to which one condition variable moderates the effect of another condition variable in determining response times.

We included random effects for participant and item. Specifically, we included by-participant and by-item varying intercepts. These random intercepts describe the degree to which particular participants or particular items lead to overall slower/faster responses. In addition, we added by-participant and by-item varying slopes for the within-participants and within-items condition variables: Hierarchy Condition, Gender Condition, and the Hierarchy Condition * Gender Condition interaction. These random slopes estimate the extent to which specific participants or specific items differ in how they are affected by the gender or hierarchy manipulations. Thus, our model assumes that both people and items differ with respect to the condition manipulations, and it estimates the degree to which they do so.

In section 4, we present the findings from the reaction time study. In section 5, we present the results of our corpus study which was conducted in order to help explain by-item differences observed in the experiment.

4. Results: Reaction time study

4.1. Hierarchy Condition effects

Across all conditions, participants took about 1,790ms to respond (*SD* = 707ms). Participants were on average 332ms faster when they had to pick the more powerful response (*M* = 1,628ms, *SD* = 634ms) than when they had to pick the less powerful response (M = 1,960ms, SD = 739ms, see Figure 1). This was reflected in strong evidence for a main effect of Task, with responses to the "powerful?" question being overall much faster than responses to the "powerless?" question (log coefficient: -0.17, 95% Bayesian credible interval: [-0.27, -0.07], posterior probability of the effect being above zero: 0.01). The fact that the 95% credible interval is below zero indicates strong support for the main effect of Task. This is also reflected in the low estimated probability of the effect being above zero.

Participants were only 13ms faster when the more powerful profession was presented on the top of the screen (*M* = 1,784ms, *SD* = 707ms) compared to the bottom of the screen (*M* = 1,797ms, *SD* = 708ms). Corresponding to this small difference, the Bayesian model indicated that there was overall no strong evidence for a main effect of Hierarchy Condition. This is suggested by a small coefficient for this main effect (-0.01, estimate of the log difference between powerful = bottom and powerful = up), whose 95% credible interval spans zero [-0.04, 0.01]. The posterior probability of powerful = top being slower is relatively low (= 0.16). While this suggests a trend that is consistent with vertical metaphorical associations, it is not low enough to constitute strong evidence for a main effect of Hierarchy Condition.

In his Experiment 2, Schubert (2005) found a reliable interaction between Task and Hierarchy Condition. Our results indicated that when participants were asked to pick the 'powerful' one of each pair, response times were almost equivalent for whether the bottom (*M* = 1,627ms, *SD* = 629ms) or the top (*M* = 1,629ms, *SD* = 640ms) was powerful. When participants were asked to pick the *less* powerful of each pair, participants were slightly *faster* when the powerful was presented on top (*M* = 1,945ms, *SD* = 737ms) compared to the bottom (*M* = 1,976ms, *SD* = 742ms). However, the Bayesian model indicated no strong evidence for such an interaction (log coefficient of the interaction +0.02, 95% CI [-0.01, 0.05], posterior probability of the effect being below zero = 0.14), as shown in Figure 1. That is, whether the powerful position was presented at the high or low position on the screen did not yield reliable differences with respect to whether the task asks for the more or less powerful role.

These results suggest that *at least on average*, there was no metaphor-congruency effect in our experiment. A number of factors that could explain the discrepancy between this result and those obtained by Schubert (2005) will be explored in the discussion section. Either way, at least given the exact specifications of our experiment (with the specific task we used, the specific participant population, and the specific set of items), we did not attain strong evidence for the hypothesis that across the board, people associate a physically higher position with more power. However, this null result may in part be due to the gendered dimension of our task, which will be explored in the next section.

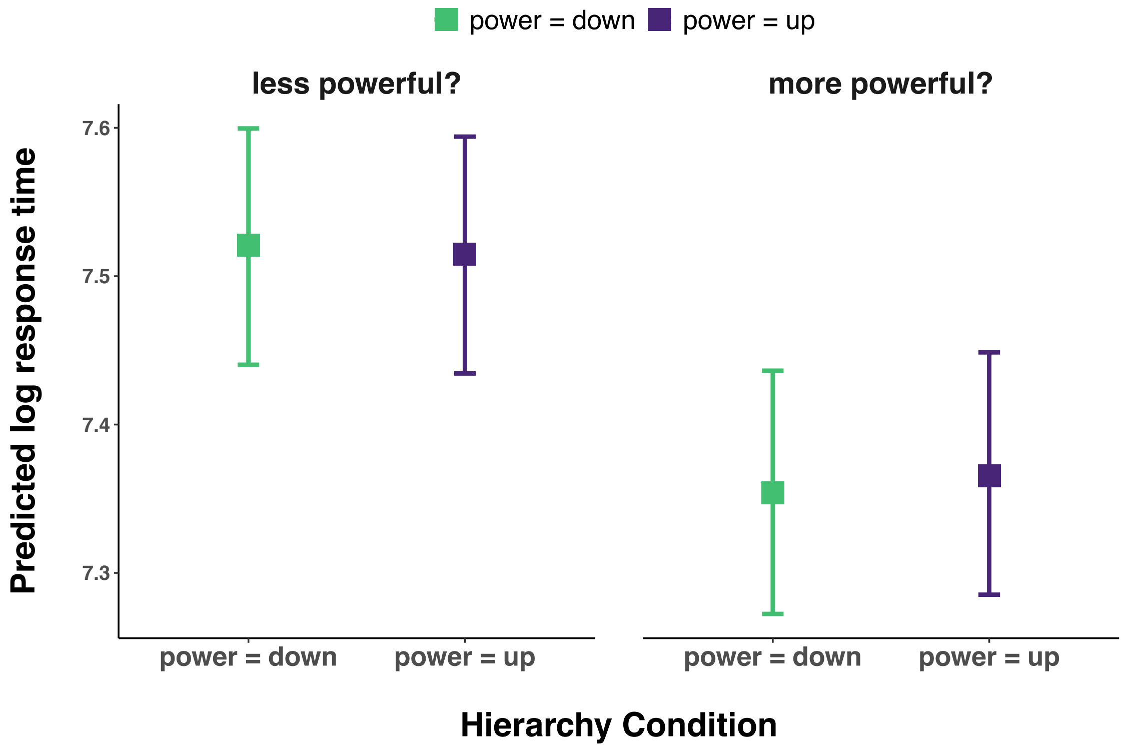

Figure 1. Marginal effects (averaged over other conditions) of the Bayesian regression model for the Task ("more powerful?" versus "less powerful?") by Hierarchy Condition (power = up versus power = bottom) interaction; as can be seen, there is no indication of an interaction or a Hierarchy Condition effect; instead, the "less powerful?" responses were much slower; error bars are Bayesian 95% credible intervals

3.2. Gender Condition effects

Participants were overall slightly faster (13ms) when the female profession was presented at the top position (*M* = 1,784ms, *SD* = 698ms) compared to the bottom position (*M* = 1,797ms, *SD* = 717ms), however, the corresponding main effect of

Gender Condition was not strongly supported by the Bayesian mixed-effects regression (estimate of the difference between female = up and male = down: 0.01, 95% CI: [-0.01, 0.03], posterior probability of the effect being below zero: 0.24).

Did this Gender Condition effect differ as a function of task? When participants were asked to pick the more powerful profession, responses to female = up (*M* = 1,628ms, *SD* = 623ms) were on average identical to responses to female = down (*M* = 1,628ms, *SD* = 646ms). When the task asked for the *less* powerful profession, participants were on average slightly faster for female = up (*M* = 1,945ms, *SD* = 734ms) than male = up (*M* = 1,976ms, *SD* = 744ms). However, altogether, there was no strong evidence for an interaction between Gender Condition and Task (log estimate of interaction -0.02, 95% CI: [-0.05, 0.01], posterior probability < 0 = 0.91).

There was, however, a notable two-way Hierarchy Condition * Gender Condition interaction (log coefficient: -0.10, 95% CI: [-0.14, -0.06], posterior probability of this being below zero: 1.0). Specifically, when the female profession was shown on top of the male one, responses were 26ms *slower* for powerful = up (*M* = 1,797ms, *SD* = 701ms) than powerful = down (*M* = 1,771ms, *SD* = 694ms). When the male profession was on top, responses were 52ms *faster* for powerful = up (*M* = 1,771ms, *SD* = 712ms) than powerful = down (*M* = 1,823ms, *SD* = 721ms). Another way of putting this result is that responses were slightly faster when the powerful role was at the top position *and* the powerful role was also the male one; the reverse was the case when women were shown to be the powerful profession at the top position. This result shows that indeed, it is both gender and power that matter with respect to vertical power metaphors in our task. In our experiment at least, vertical spatial effects only occurred when the stimulus array matches stereotypes about gender (male = powerful, female = less powerful).

3.3. Participant Gender effects

On average female participants (*M* = 1,809ms, *SD* = 722ms) responded 104ms slower than male participants (*M* = 1,705ms, *SD* = 628ms). However, there was no strong evidence for this difference being reliable, i.e., no Participant Gender main effect (log coefficient: -0.03, 95% CI: [-0.15, 0.09], posterior probability < 0 = 0.71). The factor Participant Gender furthermore did not appear to interact strongly with Hierarchy Condition (-0.01, 95% CI: [-0.05, 0.04], posterior probability: 0.63), which suggests that male participants were not overall more susceptible to metaphorical representations of power. In addition, there was no strong evidence for an interaction between Participant Gender and Stimulus Gender (-0.0, 95% CI: [-0.04, 0.04], posterior probability: 0.52).

There was, however, evidence for a three-way interaction between Participant Gender, Hierarchy Condition and Stimulus Gender (log coefficient of three-way interaction: -0.15, 95% CI: [-0.23, -0.08], posterior probability of this being below zero: 0.99). A look at Figure 2 helps to understand the nature of this complex relationship: For female participants, there was only a weak Gender * Hierarchy two-way interaction. When the male was also the powerful profession displayed on top of a female powerless profession, they responded 12ms faster (*M* = 1,801ms, *SD* = 724ms) than when power and vertical position were incongruent with gender stereotypes (M = 1,812ms, SD = 721ms). For male participants, this two-way interaction was much more pronounced. They responded 92ms faster when male = powerful = top (M = 1,637ms, SD = 639ms) than when power and vertical position were incongruent with gender stereotypes (M = 1,729ms, SD = 623ms). Another way to paraphrase this complex result is that male participants responded fastest when power stereotypes matched a view of their own gender as being powerful, that is, when the more powerful profession was described as male *and* shown on top.

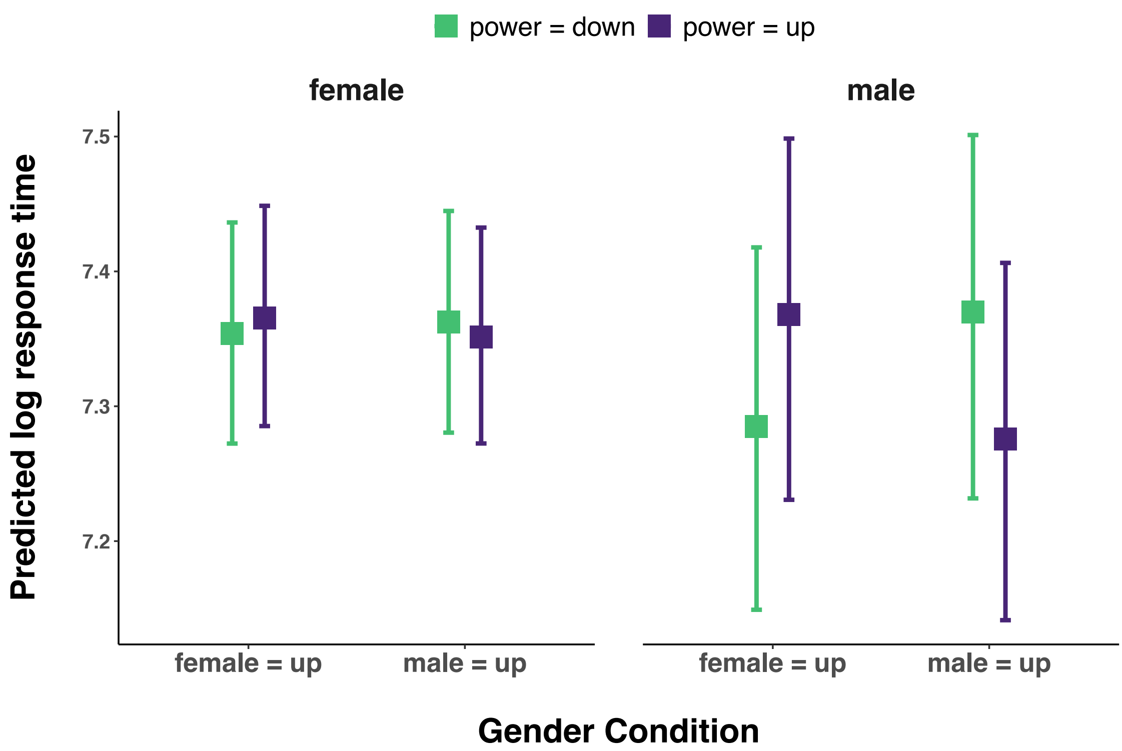

Figure 2. Interaction between Hierarchy Condition (power = up versus power = down) and Gender Condition (male = up versus male = down) for male and female

participants; as can be seen, male participants show a much stronger Hierarchy Condition * Gender interaction; error bars are Bayesian 95% credible intervals

5. Results: Corpus study

5.1. Creating a corpus-based measure of gender asymmetry

So far, we have presented all results averaged across items. However, clearly, not all professions are alike. As mentioned in the introduction section, some professions are more associated with men, whereas others are more associated with women. In this section, we explore whether participants' responses differ for different items in our study, specifically, the degree to which the powerful and powerless pair also exhibit a gender asymmetry. For example, "doctor — nurse" and "pilot — flight attendant" are two pairs where power goes along with stereotypical views of gender, with people generally being more likely to assume that doctors and pilots are male, and that nurses and flight attendants are female (see data in Misersky et al., 2014). If the association between gender and power changes the degree to which people are prone to vertical metaphorical conceptualization, we specifically expect that the interaction between Hierarchy Condition (powerful = top, powerless = top) and Stimulus Gender (male = top, female = top) should be affected by the perceived gender asymmetry for a profession pair.

To assess this, we constructed a corpus-based gender asymmetry ratio which quantifies the degree to which a particular profession is more explicitly talked about as male or female in text. We then validated this corpus-based measure by comparing it to perceptual stereotypes coming from a rating study (Misersky et al., 2014). Finally, we correlate the corpus-based measure with the behavioral responses from our experiment. This analysis adds a very important component to our study: it shows that the results not only differ by participants, but also by items, and these item differences are predictable from corpus-derived language statistics, specifically, how often particular professions are *talked about* as being male or female.

To construct the corpus-based gender asymmetry measure, we used the spoken section of the British National Corpus. We used sketch engine to search for each keyword (each of the profession terms listed in Appendix A), asking for 1,000 results per query. We used sketch engine's part of speech function to restrict queries to singular nouns. The results for each search term were shuffled once within sketch engine and then exported to a spreadsheet.

The resultant spreadsheets were hand-annotated for whether the linguistic context clarified that a mentioned profession was male or female. We also categorized contexts as "unspecified" if there was not enough information to determine the gender of a mentioned profession. We began analyzing a narrow context of 10 words to each side of the keyword, but if this information was insufficient to determine the contextually implied gender, we broadened the context to 100 words in each direction. We marked concordance lines where the keyword was used not as a profession term as cases to be "disregarded" (for example if the word "conductor" referred to the electrical component rather than the transport profession). This process of hand annotation was continued until we had either 200 results for each key word that were not marked as "disregarded", or until we had exhausted all the corpus examples (not all searches returned enough results to produce 200 lines). The file containing our annotations can be found on the OSF repository for this paper (https://osf.io/mbv24/).

For each profession, we computed a ratio quantifying the degree to which male text mentions were more frequent than female text mentions. For example, 66 occurrences of the word *surgeon* referred to men, whereas only 2 occurrences referred to women, which yields a ratio of 66 / 2 = 33. However, since there were many 0 mentions and ratios with 0 in the denominator are undefined, we added +1 to all male and all female values before calculating the ratios.

5.2. Validation of the corpus-based measure with gender stereotype ratings

To validate that our corpus-based asymmetry measure has anything to do with gender stereotypes, we performed a comparison with existing data for perceptual gender stereotypes. Misersky and colleagues (2014) performed a questionnaire-based study where participants had to "estimate the extent to which the presented social and occupational groups actually consisted of women and men" on an 11-point rating scale ranging from 0% women to 100% men. The rating study was conducted in seven languages (Czech, English, French, German, Italian, Norwegian, and Slovak), but here we focus on the English data, which was based on 281 undergraduates from the University of Sussex (UK). We sought to find the closest match in the Misersky et al. (2014) ratings, such as "sports coach" for "coach", which we were able to do in all but 14 cases (for all matching decisions, see https://osf.io/mbv24/).

Initial exploration showed that the perceptual gender stereotype ratings taken from Misersky et al. (2014) track the logarithm of the corpus-based ratio, rather than the actual ratio. This is consistent with the observation that human behavior and human perception tracks the logarithm of word frequencies, rather than raw frequencies (e.g., Smith & Levy, 2013; Zipf, 1949). Thus, all following analyses are reported for the log10 transformed ratios.

The correlation between perceptual stereotypes and the text-based male/female ratio is shown in Figure 3. As can be seen, professions that more frequently occur as male in the corpus (such as “soldier”, “footballer”, and “architect”) are also judged to be more stereotypically male professions in the Misersky et al. (2014) norming study. Similarly, professions that more frequently occur as female in the corpus (such as “flight attendant” and “nurse”) are also judged to be more stereotypically female.

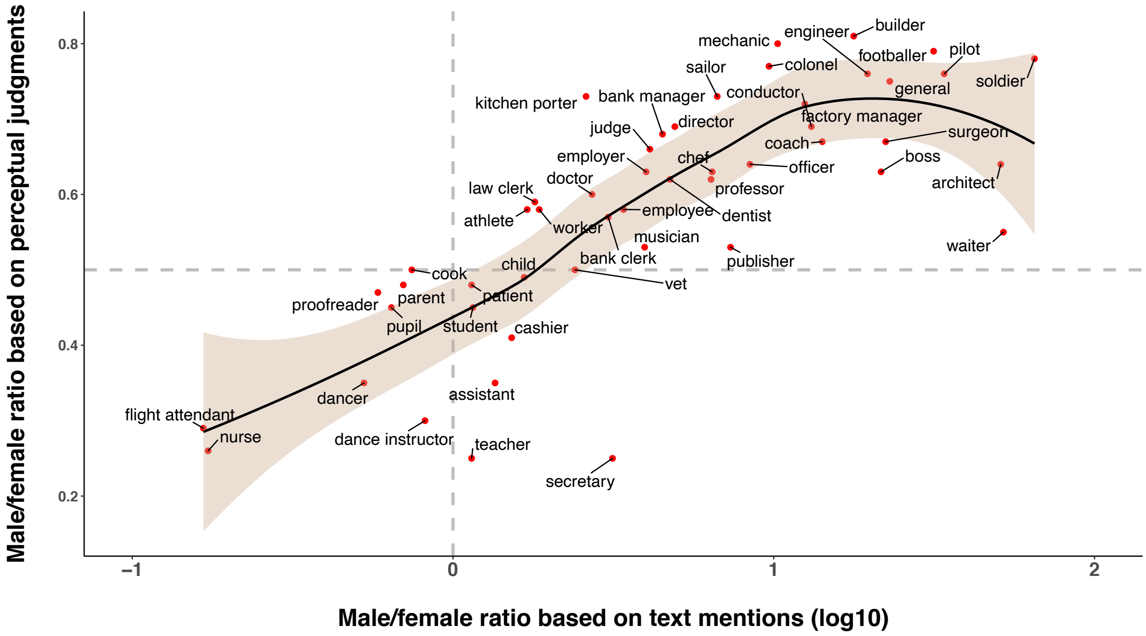

Figure 3. Correlation between corpus-based and perceptual stereotypes (with rating data from Misersky et al., 2014); the x-axis represents our logarithm text mention ratio (male/female), whereas the y-axis represents the perceptual ratings; the dashed lines indicate the point where male/female ratios are equal on the respective measures; the superimposed model is a loess smoother which shows that the relationship plateaus out for very high values of text mentions

Our corpus-based gender asymmetry ratio correlated strongly with the Misersky et al's. perceptual ratios (Pearson's $r = 0.73$). The Bayesian 95% credible interval of this correlation coefficient ranged from 0.53 to 0.79. The posterior probability of the correlation being below zero is $p = 0.0$, which indicates strong evidence for the two measures being positively correlated. Interestingly, a look at Figure 3 reveals some nonlinearity in the relationship between perceptual stereotypes (ratings) and the corpus-based measure of stereotypes. Specifically, for professions that occur almost exclusively as male in corpora (and rarely ever as female), perceptual judgments are slightly lower than what a linear trend would suggest.[4]

This nonlinear trend has two possible explanations: It may be that participants in the Misersky et al. (2014) rating study disprefer the edges of the rating scale, i.e., they may attempt to be conservative and avoid judging any profession as being either 100% male or female. Alternatively, it may be that the way people talk about particular professions as male or female is skewed. That language use differs from reality is evidenced by a comparison to the real world gender ratios aggregated by Garnham et al. (2015) from the UK Office for National Statistics: For example, 8% of all soldiers in the UK are female, but in the corpus, 97% of all mentions were about male soldiers. This may reflect a bias of English speakers to talk more about male soldiers, than female soldiers.

5.3. Item-specific analysis and correlations with corpus data and norms

In the following analysis, we assess whether the behavioral results discussed so far are moderated by the degree to which specific professions are gender-biased. For

[4] To assess whether there was reliable evidence for a quadratic trend, a Bayesian regression was fitted that included the logarithmic male/female ratio as both a linear predictor and a quadratic one. The quadratic predictor was negative (-0.2, 95% CI: [-0.04, 0.0]), indicating an inverse U-shape as seen in Figure 1. The posterior probability of this quadratic effect being inverse U-shaped (< 0) was 0.97, indicating reasonably strong evidence for a plateauing-out of the relationship between the corpus-based measure and the perceptual stereotypes.

this, we extracted the by-item random slopes for the Hierarchy Condition * Gender interaction. These random effect coefficients describe the strength of the association between gender and vertical space for each stimulus pair. For example, for the item pair "officer — cadet", both the powerful and the powerless position were almost exclusively male in our corpus, and likewise, the interaction effect was weaker for this pair (log response time coefficient for this item: -0.09) than for "doctor — nurse" (-0.11), which is a pair with a bigger gender asymmetry in the corpus-data.[5]

Figure 4 shows the correlation between our corpus-based measure and the interaction coefficients from the Bayesian regression model. As can be seen, the behavioral interaction is stronger (more negative values = faster) when there is a bigger gender asymmetry in the corpus (towards the right of the figure). The between the behavioral measure and the corpus-based measure is reasonably strong, with Pearson's r = -0.45 (95% CI [-0.77, -0.16], high posterior probability of being negative p = 0.99). What does this mean in terms of actual behavior? The correlation between the interaction coefficients and our gender asymmetry measure is easiest explained for the pair "doctor — nurse", a relatively asymmetrical pair for which the powerful profession (doctor) is more strongly associated with men in the corpus, compared to the less powerful profession (nurse), which is relatively more female. For such an asymmetrical pair, participants were even faster if the "male doctor"

[5] We only use the corpus-based measure here rather than the Misersky et al. (2014) norms because the latter do not contain enough professions that are used as our stimuli. Since our corpus analysis was targeted, we have data for all item pairs. In this analysis we used the ratio of ratios (not the ratio of logarithmic ratios, which did not show strong correlations).

was presented on the top of the screen, that is, when both the powerful and the male were presented on top. This suggests that strong gender asymmetries (as in our case estimated via language statistics) can strengthen effects of vertical power metaphors observed in experimental studies.

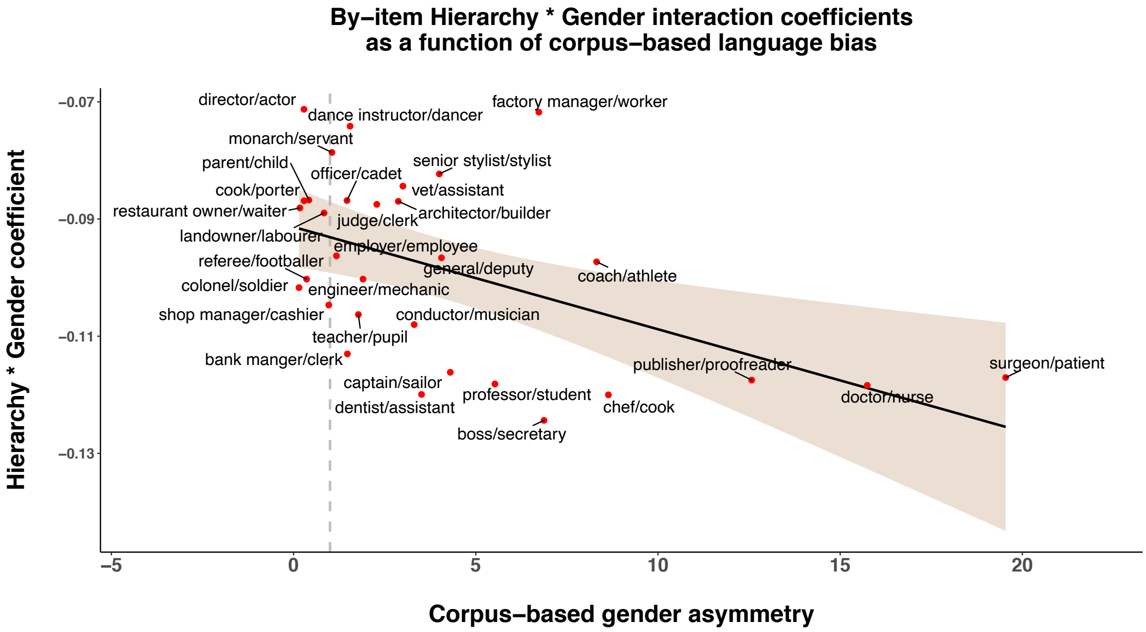

Figure 4. Text-based male/female ratios moderate the Hierarchy * Condition interaction; the *y*-axis represents log response times, with lower values indicating that these were items for which the hierarchy * gender interaction was more pronounced (participants were faster with the male variant if it was presented on top); the *x*-axis represents the degree to which the powerful member of each stimulus pair is associated with male mentions in the text, compared to female mentions; the plot excludes one data point ("pilot — flight attendant") that has an extremely high gender asymmetry (> 200); superimposed simple linear regression fit with 95% confidence interval

6. Discussion

This study investigated the relationship between power and space in line with the metaphor POWER IS UP. We made several theoretical and methodological contributions. First, we sought to replicate Schubert's (2005) finding that people are quicker to make judgments about power when a relatively powerful group is physically displayed on top of a relatively less powerful group, congruent with the way we talk about power in terms of vertical positioning ("at the top of the hierarchy", "to hit rock bottom" etc.). While we did not find evidence for a main effect of vertical metaphorical association, we found a reliable three-way interaction between Participant Gender, Hierarchy Condition, and Stimulus Gender, where responses to "power = up" were slightly faster when the up position was also the male profession. This interaction was much more pronounced for the male participants than for the female participants. When the male role was at the top position, male participants were faster when this position was also the more powerful one. In addition, we also found differences across professions. Using corpus frequencies, we were able to show that those profession pairs that exhibited stronger gender asymmetries lead to stronger vertical metaphorical associations. Thus, we have demonstrated that POWER IS UP metaphors are understood differently as a function of the participant's gender, and they are also understood differently as a function of gender stereotypes associated with particular professions.

How are our results to be interpreted in comparison to Schubert (2005)? Our findings invite the possibility that there may have been gender effects in Schubert's original study that were not tested for. As discussed above, gender stereotypes are intimately tied to particular professions, including the profession labels used in Schubert's study. Power then could be reflected in occupational differences, gender differences, or, both. If only one of these factors is present (e.g., only profession labels, as in Schubert's study), this dimension becomes the dominant factor. If, however, another dimension is made manifest (in our case: gender), this dimension can trump the other (compare Duffy & Evans, 2017). Thus, we do not consider the present study as a "failure to replicate" the idea of metaphorical associations between power and vertical space. Instead, we did find evidence for the idea that people automatically associate power with vertical space, and that this association is moderated by the effect of gender.

There are, however, other differences to Schubert's (2005) study that need to be pointed out. First, given that we have successfully demonstrated individual differences, it is possible, and indeed likely, that differences in the participant sample may explain why we have failed to find Schubert's main effect of faster responses to power = up. Specifically, Schubert's participant sample included a relatively larger proportion of male participants than our participant sample (40% male participants as opposed to only 18% in our case). Given that we have found stronger vertical metaphorical associations for male participants in the present study, it could be that Schubert's (2005) results are disproportionately driven by the male participants. Another important difference is language: Schubert's study was conducted in German, which encodes gender morphologically in profession terms. We explicitly paired each profession with the labels "male" and "female", whereas it was not stated whether Schubert's stimuli used masculine generics (e.g., *Professor*) or gender-marked forms (e.g., *Professor* (masc.), *Professorin* (fem.)). Future studies should seek to investigate the effects of vertical power metaphors in relation to grammatical gender.

In addition, we found by-item differences with respect to gender asymmetries, and we were furthermore able to predict these differences using a corpus-based measure of gender asymmetry. Specifically, we found stronger vertical associations of power and gender (Hierarchy * Gender interaction) when the powerful position was more strongly associated with men, and when the powerless position was more strongly associated with women. In other words, the distributions found within the corpus mapped onto the results of the reaction time study. This relates to a further potential difference between our study and Schubert (2005), which is the difference in item composition (our study included more items than Schubert's). Our item-specific analysis suggests that some of the vertical metaphorical associations in other experimental studies may have partially resulted from gender asymmetries that are conflated with power asymmetries. More generally, our results empirically demonstrate that metaphor effects are not the same for all items, which points to the importance of considering item composition in studies of metaphor comprehension.

Our item-specific analysis is particularly interesting as it shows a correlation between language use and psycholinguistic behavior. Specifically, when a profession pair exhibited more of a gender asymmetry in language use (one profession is talked about as being male/female more than the other), associations between gender, power, and vertical space are more pronounced. This demonstrates that corpus analysis and experiments can be used together for a richer understanding of the context-dependent nature of metaphor.

At a more basic level, it is noteworthy that corpus statistics are correlated with perceptual stereotypes of gender at all (Misersky et al., 2014). This opens up new avenues to explore additional research questions: To what extent do corpus-based measures of gender asymmetries for professions reflect occupational statistics? Which professions are reliably more talked about as male or female, more so than the occupational statistics would suggest? For example, we found that in our corpus data, soldiers were almost exclusively talked about as male, even though 8% of UK soldiers are actually female. From a more methodological perspective, our results suggest that when it is impossible or not feasible to conduct a large-scale rating study for gender stereotypes, corpus measures can be used as a stand-in for perceptual stereotypes.

Most importantly, our results have answered the call to investigate individual differences in cognitive linguistics (Dąbrowska, 2016), and in metaphor research more specifically (Littlemore, 2019). We have found that vertical metaphors for power are understood differently by men and women, with male participants having stronger vertical biases than female participants. Our results provide clear evidence for gender-based individual differences in metaphor comprehension, thus demonstrating empirically that the same metaphor is understood differently by different people. This finding is line with previous research showing that men are more likely than women to associate power with physical force. Whereas men associate physical force with exerting influence and gaining power, women are more likely to associate it with loss of control (Campbell et al., 1992) and physical actions, such as making a fist, have been shown to lead men, but not women, to perceive themselves as being more powerful (Schubert & Koole, 2009). These differences are partly a reflection of the different experiences that men and women have with their bodies, and partly a reflection of social expectations regarding behavior.

6. Conclusion

In this study, we have found new evidence for the existence that people think about power in terms of vertical space, but the association between verticality and power was moderated by gender. The influence of gender was three-fold: first, the gender of the label ("male doctor" versus "female doctor") interacted with metaphor congruency; second, the gender of the participant further moderated the metaphor congruency effect; third, the degree to which a profession was associated with gender stereotypes exerted a further influence on metaphorical associations between power and space. Specifically, participants' responses were fastest when the powerful profession was male and shown on top, and this pattern was even more pronounced for the male participants in our study, as well as the more gendered professions. Altogether, our results show that when studying the POWER IS UP metaphor, gender is an important dimension to consider. Moreover, our results show that POWER IS UP metaphors are not comprehended the same way by all people, with men and women showing different responses to the metaphor. This is a clear demonstration of gender-based differences in metaphor perception. Gender should therefore be considered to be a key individual difference variable when discussing embodied metaphor. We need to be careful not to assume that a 'male' world view (Perez, 2019), and the embodied metaphors upon which it is built, is a universal world view.

References

Banaji, M. R., & Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, *7*(3), 136–141. https://doi.org/10.1111/j.1467-9280.1996.tb00346.x

Berings, D., & Adriaenssens, S. (2012). The role of business ethics, personality, work values and gender in vocational interests from adolescents. *Journal of Business Ethics*, *106*(3), 325–335. https://doi.org/10.1007/s10551-011-0999-2

Blau, F. D., & Kahn, L. M. (1992). The gender earnings gap: Learning from international comparisons. *The American Economic Review*, *82*(2), 533–538.

Blau, F. D., & Kahn, L. M. (1996). Wage structure and gender earnings differentials: An international comparison. *Economica*, S29–S62.

Braun, F., Sczesny, S., & Stahlberg, D. (2005). Cognitive effects of masculine generics in German: An overview of empirical findings. *Communications*, *30*(1), 1–21. https://doi.org/10.1515/comm.2005.30.1.1

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Campbell, A., Muncer, S., & Coyle, E. (1992). Social representation of aggression as an explanation of gender differences: A preliminary study. *Aggressive Behavior*, *18*(2), 95–108. https://doi.org/10.1002/1098-2337(1992)18:2<95::AID-AB2480180203>3.0.CO;2-5

Carli, L. L. (1999). Gender, interpersonal power, and social influence. *Journal of Social Issues*, *55*(1), 81–99. https://doi.org/10.1111/0022-4537.00106

Cejka, M. A., & Eagly, A. H. (1999). Gender-stereotypic images of occupations correspond to the sex segregation of employment. *Personality and Social Psychology Bulletin*, *25*(4), 413–423. https://doi.org/10.1177/0146167299025004002

Chen, Z., Fiske, S. T., & Lee, T. L. (2009). Ambivalent sexism and power-related gender-role ideology in marriage. *Sex Roles*, *60*(11–12), 765–778. https://doi.org/10.1007/s11199-009-9585-9

Dąbrowska, E. (2016). Cognitive Linguistics' seven deadly sins. *Cognitive Linguistics*, *27*(4), 479–491. https://doi.org/10.1515/cog-2016-0059

Dancygier, B., & Sweetser, E. (2014). *Figurative language*. Cambridge University Press.

Diekman, A. B., & Eagly, A. H. (2000). Stereotypes as dynamic constructs: Women and men of the past, present, and future. *Personality and Social Psychology Bulletin*, *26*(10), 1171–1188. https://doi.org/10.1177/0146167200262001

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Duffy, S. E., & Evans, V. (2017). The top trumps of time: Factors motivating the resolution of temporal ambiguity. *Language and Cognition*, *9*(2), 293–315. https://doi.org/10.1017/langcog.2016.8

Duffy, S. E., & Feist, M. I. (2014). Individual differences in the interpretation of ambiguous statements about time. *Cognitive Linguistics*, *25*(1), 29–54. https://doi.org/doi.org/10.1515/cog-2013-0030

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Lawrence Erlbaum.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037//0033-295X.109.3.573

Eagly, A. H., Wood, W., & Diekman, A. B. (2000). Social role theory of sex differences and similarities: A current appraisal. In T. Eckes & H. M. Trautner (Eds.), *The developmental social psychology of gender* (Vol. 12, pp. 123–174). Lawrence Erlbaum Associates.

Feingold, A. (1994). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Ferriman, K., Lubinski, D., & Benbow, C. P. (2009). Work preferences, life values, and personal views of top math/science graduate students and the profoundly gifted: Developmental changes and gender differences during emerging adulthood and parenthood. *Journal of Personality and Social Psychology*, *97*(3), 517–532. https://doi.org/10.1037/a0016030

Fetterman, A. K., Bair, J. L., Werth, M., Landkammer, F., & Robinson, M. D. (2016). The scope and consequences of metaphoric thinking: Using individual differences in metaphor usage to understand how metaphor functions. *Journal of Personality and Social Psychology*, *110*(3), 458. https://doi.org/10.1037/pspp0000067

Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology*, *60*(5), 656–674. https://doi.org/10.1037/0022-3514.60.5.656

Franke, M., & Roettger, T. B. (2019). Bayesian regression modeling (for factorial designs): A tutorial. *PsyArXiv*. https://doi.org/10.31234/osf.io/cdxv3

Gabriel, U., Gygax, P., Sarrasin, O., Garnham, A., & Oakhill, J. (2008). Au pairs are rarely male: Norms on the gender perception of role names across English, French, and German. *Behavior Research Methods*, *40*(1), 206–212. https://doi.org/10.3758/BRM.40.1.206

Gabriel, U., & Mellenberger, F. (2004). Exchanging the generic masculine for gender-balanced forms-The impact of context valence. *Swiss Journal of Psychology*, *63*(4), 273–278. https://doi.org/10.1024/1421-0185.63.4.273

Garnham, A., Doehren, S., & Gygax, P. (2015). True gender ratios and stereotype rating norms. *Frontiers in Psychology*, *6*, 1023. https://doi.org/10.3389/fpsyg.2015.01023

Garrett, C. S., Ein, P. L., & Tremaine, L. (1977). The development of gender stereotyping of adult occupations in elementary school children. *Child Development*, 507–512. https://doi.org/10.2307/1128646

Gibbs, R. W. (2011). Evaluating conceptual metaphor theory. *Discourse Processes*, *48*(8), 529–562. https://doi.org/10.1080/0163853X.2011.606103

Giessner, S. R., & Schubert, T. W. (2007). High in the hierarchy: How vertical location and judgments of leaders' power are interrelated. *Organizational Behavior and Human Decision Processes*, *104*(1), 30–44. https://doi.org/10.1016/j.obhdp.2006.10.001

Grady, J. E. (1997). Theories are buildings revisited. *Cognitive Linguistics*, *8*(4), 267–290.

Gygax, P., Gabriel, U., Sarrasin, O., Oakhill, J., & Garnham, A. (2008). Generically intended, but specifically interpreted: When beauticians, musicians, and mechanics are all men. *Language and Cognitive Processes*, *23*(3), 464–485. https://doi.org/10.1080/01690960701702035

Hamilton, M. C. (1988). Using masculine generics: Does generic he increase male bias in the user's imagery? *Sex Roles*, *19*(11–12), 785–799. https://doi.org/10.1007/BF00288993

Heilman, M. E., Wallen, A. S., Fuchs, D., & Tamkins, M. M. (2004). Penalties for success: Reactions to women who succeed at male gender-typed tasks. *Journal of Applied Psychology*, *89*(3), 416. https://doi.org/10.1037/0021-9010.89.3.416

Horvath, L. K., Merkel, E. F., Maass, A., & Sczesny, S. (2016). Does gender-fair language pay off? The social perception of professions from a cross-linguistic perspective. *Frontiers in Psychology*, *6*, 2018. https://doi.org/10.3389/fpsyg.2015.02018

Horvath, L. K., & Sczesny, S. (2016). Reducing women's lack of fit with leadership positions? Effects of the wording of job advertisements. *European Journal of Work and Organizational Psychology*, *25*(2), 316–328. https://doi.org/10.1080/1359432X.2015.1067611

Hutchinson, S., & Louwerse, M. (2013). Language statistics and individual differences in processing primary metaphors. *Cognitive Linguistics*, *24*(4), 667–687. https://doi.org/10.1515/cog-2013-0023

Judge, T. A., & Cable, D. M. (2004). The effect of physical height on workplace success and income: Preliminary test of a theoretical model. *Journal of Applied Psychology*, *89*(3), 428. https://doi.org/10.1037/0021-9010.89.3.428

Julich, N. (2019). *Metaphors in verbal descriptions of music* [PhD Thesis]. University of Leipzig.

Lakens, D., Semin, G. R., & Foroni, F. (2011). Why your highness needs the people. *Social Psychology*. https://doi.org/10.1027/1864-9335/a000064

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Landau, M. J. (2017). *Conceptual metaphor in social psychology: The poetics of everyday life*. Routledge.

Landau, M. J., Meier, B. P., & Keefer, L. A. (2010). A metaphor-enriched social cognition. *Psychological Bulletin*, *136*(6), 1045–1067.

Lederer, J. (2015). Exploring the metaphorical models of transgenderism. *Metaphor and Symbol*, *30*(2), 95–117. https://doi.org/10.1080/10926488.2015.1016809

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests: The importance of the people–things dimension. *Journal of Personality and Social Psychology*, *74*(4), 996–1009. https://doi.org/10.1037/0022-3514.74.4.996

Littlemore, J. (2019). *Metaphors in the mind: Sources of variation in embodied metaphor*. Cambridge University Press.

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. *Journal of Personality and Social Psychology*, *66*(1), 37–47. https://doi.org/10.1037/0022-3514.66.1.37

Misersky, J., Gygax, P. M., Canal, P., Gabriel, U., Garnham, A., Braun, F., Chiarini, T., Englund, K., Hanulikova, A., & Öttl, A. (2014). Norms on the gender perception of role nouns in Czech, English, French, German, Italian, Norwegian, and Slovak. *Behavior Research Methods*, *46*(3), 841–871. https://doi.org/10.3758/s13428-013-0409-z

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Ortiz, M. J. (2011). Primary metaphors and monomodal visual metaphors. *Journal of Pragmatics*, *43*(6), 1568–1580. https://doi.org/10.1016/j.pragma.2010.12.003

Perez, C. C. (2019). *Invisible Women: Exposing data bias in a world designed for men*. Chatto & Windus.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ramaci, T., Pellerone, M., Ledda, C., Presti, G., Squatrito, V., & Rapisarda, V. (2017). Gender stereotypes in occupational choice: A cross-sectional study on a group of Italian adolescents. *Psychology Research and Behavior Management*, *10*, 109–117.

Redl, T., Eerland, A., & Sanders, T. J. (2018). The processing of the Dutch masculine generic zijn 'his' across stereotype contexts: An eye-tracking study. *PloS One*, *13*(10). https://doi.org/10.1371/journal.pone.0205903

Robinson, M. D., Zabelina, D. L., Ode, S., & Moeller, S. K. (2008). The vertical nature of dominance-submission: Individual differences in vertical attention. *Journal of Research in Personality*, *42*(4), 933–948. https://doi.org/10.1016/j.jrp.2007.12.002

Rudman, L. A., & Glick, P. (2012). *The social psychology of gender: How power and intimacy shape gender relations*. Guilford Press.

Schubert, T. W. (2004). The power in your hand: Gender differences in bodily feedback from making a fist. *Personality and Social Psychology Bulletin*, *30*(6), 757–769. https://doi.org/10.1177/0146167204263780

Schubert, T. W. (2005). Your highness: Vertical positions as perceptual symbols of power. *Journal of Personality and Social Psychology*, *89*(1), 1–21. https://doi.org/10.1037/0022-3514.89.1.1

Schubert, T. W., & Koole, S. L. (2009). The embodied self: Making a fist enhances men's power-related self-conceptions. *Journal of Experimental Social Psychology*, *45*(4), 828–834. https://doi.org/10.1016/j.jesp.2009.02.003

Schwartz, B., Tesser, A., & Powell, E. (1982). Dominance cues in nonverbal behavior. *Social Psychology Quarterly*, 114–120. https://doi.org/10.2307/3033934

Slowikowski, K. (2019). *ggrepel: Automatically position non-overlapping text labels with "ggplot2."* https://CRAN.R-project.org/package=ggrepel

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Stahlberg, D., Braun, F., Irmen, L., & Sczesny, S. (2007). Representation of the sexes in language. In K. Fiedler (Ed.), *Social communication* (pp. 163–187). Psychology Press.

Stangor, C., Lynch, L., Duan, C., & Glas, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207–218.

Teig, S., & Susskind, J. E. (2008). Truck driver or nurse? The impact of gender roles and occupational status on children's occupational preferences. *Sex Roles*, *58*(11–12), 848–863. https://doi.org/10.1007/s11199-008-9410-x

van Knippenberg, A., van Twuyver, M., & Pepels, J. (1994). Factors affecting social categorization processes in memory. *British Journal of Social Psychology*, *33*(4), 419–431. https://doi.org/10.1111/j.2044-8309.1994.tb01038.x

Wickham, H. (2017). *tidyverse: Easily install and load the "tidyverse."* https://CRAN.R-project.org/package=tidyverse

Wilbourn, M. P., & Kee, D. W. (2010). Henry the nurse is a doctor too: Implicitly examining children's gender stereotypes for male and female occupational roles. *Sex Roles*, *62*(9–10), 670–683. https://doi.org/10.1007/s11199-010-9773-7

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. *ArXiv Preprint ArXiv:1308.5499*.

Winter, B. (2014). Horror movies and the cognitive ecology of primary metaphors. *Metaphor and Symbol*, *29*(3), 151–170.

Winter, B., Daguna, J., & Matlock, T. (2018). Metaphor-enriched social cognition and spatial bias in the courtroom. *Metaphor and the Social World*, *8*(1), 81–99. https://doi.org/doi.org/10.1075/msw.17001.win

Young, T. J., & French, L. A. (1996). Height and perceived competence of US presidents. *Perceptual and Motor Skills*, *82*(3), 1002–1002. https://doi.org/10.1177/003151259608200301

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison Wesley.