Dialogue & Discourse 17(2) (2026) 94–135 doi: 10.5087/dad.2026.203

Gendered language and liking: Interactions of likeability, gender, and linguistic content in computer-mediated dyadic conversations

Teona Banu BANUT@TCD.IE
Trinity College Dublin, the University of Dublin, Ireland

Carl Vogel VOGEL@TCD.IE
Trinity College Dublin, the University of Dublin, Ireland
and *Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, Hungary*

Editor: Rebecca Clift

Submitted 10/2025; Accepted 08/2026; Published online 08/2026

Abstract

We investigate how gender affects social interaction outcomes, and to what extent gendered language mediates the association between gender and interpersonal liking. From a multimodal corpus of dialogue interactions in which participants rated and estimated likeability (CANDOR), we use inferential statistical analyses to identify expressions associated with gender, likeability, and the liking gap, the difference between how much an interlocutor likes another and how liked interlocutors imagine they are. Results reveal a Dunning Kruger effect of likeability, as well as evidence that an interaction of gender and liking is partly explained by gendered language independent of speaker gender. These findings bridge and expand previous work on liking, gender attitudes, gendered language, and likeable language. Our results suggest that aspects of stereotypically feminine language can enhance likeability regardless of the speaker's gender identity.

1. Introduction

The social function of talk has been identified as at least as important as the informational content of ordinary conversation (in contrast to transactional interactions), according to theoreticians (Brown and Yule, 1983; Dunbar, 1998) and people in general (Clark et al., 2019). While transactional interactions are typically task-based, and evaluated in terms of task success or collaboration (Reitter and Moore, 2007, 2014; Noseworthy et al., 2017; Murat et al., 2022), social interactions are focused on the establishment and maintenance of social relationships and are evaluated in terms of their ability to fulfil social needs (Brown and Yule, 1983; Clark et al., 2019). We consider interpersonal *liking* as an outcome measure of social interactions.

When studying liking, Boothby et al. (2018) argue that one must distinguish between two different perceptual levels: actual liking (how much A likes B), and perceived liking (how much B thinks

A likes them); the "liking gap" is a known and studied phenomenon whereby a significant difference arises between measurements at these two perceptual levels (Boothby et al., 2018). Both actual and perceived liking are different from self liking (how much B likes themselves). Literature has tended to exclusively address actual liking. Gender and gender roles are two traits that have been shown to interact with biases, expectations, and mental associations in ways which suggests a potential link to liking (Eagly and Mladinic, 1994; Leaper, 1987; Rudman and Goodwin, 2004); a tendency for more favourable implicit attitudes towards women has been called the "women-are-wonderful" effect (Eagly and Mladinic, 1994). A series of gender distinctions with respect to language use have been described in literature (see §2.3.1). Separately, it is known that the language that people use influences how they are perceived (see §2.3.2 for more on this). In this paper, we report empirically identified patterns of language use that interact with gender and with liking.

In line with the European Institute for Gender Equality, the American Psychological Association and other organizations, we understand gender identity to mean one's psychological sense of their own gender, which is internal and individual, but influenced by their perception of others. Throughout this paper we use the terms *woman*, *man*, *(one's) gender*, and *gender identity* to refer to one's self-reported gender identity. This is a component of social gender (Jäggi et al., 2025) which is expected to interact with linguistic and social behaviour (Acton, 2011; Eckert, 1989) and which is commonly recorded in data sets, typically as the answer to a question such as 'What is your gender?'. In the independently constructed data that we analyse, the answer to this question is encoded as one of three options: 'male', 'female', 'other or prefer not to answer'. We take this answer to be their self-reported gender identity,[1] making this data suitable for the purposes of our analysis. Throughout this work, we use a binary classification of gender as a simplification of the multi-dimensional gender spectrum on which one might be able to place their identity. This simplification, however, is justified by the proclivity of participants to indicate a female or male gender identity. In the data we describe here, the distribution is as follows: 98.4% (n=3203), compared to 1.6% (n=53) indicating 'other or prefer not to answer'. While future data sets may pose questions and record responses in a more nuanced way, and while this study could be replicated with data set collected mindful of the research reported here, we think it a responsibility of researchers to extract what can be gleaned from existing data before creating new multimodal corpora, given the vast resources that inhere in the construction of any such corpus. Thus, we report here on results that emerge from a particular unscripted corpus (English language, constructed during the Coronavirus pandemic, and with other comparable inherent contextual parameters), but with reason to believe that results mostly generalize. It will be a task of future work to more completely specify and determine the roles of parameters that interact with generalizability.

1. We do not have information, nor do we speculate on their biological sex; we have no information about any participant's sex chromosomes. Neither do we have information nor do we speculate on how the participants perceive their gender in relation to their sex or to their gender assigned at birth, a relationship that we refer to as *gender modality* (e.g. cisgender, transgender, fluid etc).

We refer to *femininity* and *masculinity* as the characteristics specific of women and men respectively. In our analysis, whether a word or a participant's vocabulary is characteristic of women or men is determined by probability patterns, and does not necessarily refer to cultural norms, nor does it stem from biological traits. The participants may or may not be conscious of these characteristics, and the characteristics may or may not be culturally perceived as gender-specific.

In assembling lists of distinctive words, we use the terms *gender-preponderant*, *feminine*, and *masculine* words to refer to the words that are used preponderantly by people who reported themselves to be women or men, respectively. In constructing our vocabulary classification system, we use the terms *gender-preponderant*, *feminine*, and *masculine* vocabularies to refer to word choice tendencies characteristic of women or men. In both of these cases, we compare women to men, as opposed to comparing women to non-women. A 'non-woman' category in this case would represent men and people who indicated a nonbinary identity or declined to answer. Participants declining to answer might still identify as women; therefore, counting them as 'non-women' is likely to introduce noise in the analysis. Nonbinary identities might include, among others, identities that combine aspects of womanhood and manhood (American Psychological Association, n.d.b), and, therefore, counting them as 'non-women' is also likely to introduce noise in the analysis. We conclude that, based on the available data, the comparison of women to men is the most appropriate here. This is a consideration that is likely to apply in a comparable form to many extant corpora.

This paper studies the points of convergence and bridges these previously unconnected four bodies of research: liking, gender attitudes, gendered language, and likeable language. We show that the "liking gap" emerges not just from a general tendency to underestimate oneself, as it has been previously described, but from larger interacting patterns of perception and metaperception: while most participants scored very high in likeability (actual liking), metaperceptions (perceived liking) tended to be mildly positive across the board, regardless of how positive (or negative) external perceptions (actual liking) were. Our findings point towards a manifestation of the Dunning-Kruger effect in interpersonal liking whereby individuals proficient in a "skill" (likeability) underestimate their ability, while deficient individuals overestimate it. We further suggest that gender interacts with social outcomes in dyadic conversations, and that the interaction between gender and liking can be captured by differences in language, measured as statistically significant differences in word usage patterns. We find, *inter alia*, a subset of the language used in the data set which is distinct from what is thought of as the usual language of politeness, except inasmuch as part of the language of politeness includes expressions associated with 'relational work' (Locher, 2005), that is shown to interact with likeability. In the politeness literature, relational work is the activity, partly the use of language, through which interlocutors negotiate their relationships, and the focus on relational work is meant to de-emphasize the prospect that any linguistic expressions convey politeness in any reliable sense. Nonetheless, the language associated with naming and categorizing relationships is reasonable to think of as the language of relational work.

Our findings uncover gendered language effects which have evidently not been previously described in the literature; we further contribute to the development of a trait-based vocabulary classification method which in this paper we employ to classify vocabularies as feminine or masculine. This work also refines the analysis surrounding the liking gap.

We review relevant literature in the aforementioned research areas (§2). We describe our methods for analysing the data (§4) and discuss the results that emerge (§5) and (§6). We conclude with observations and interpretations (§7) of the interactions among liking, gender, and vocabulary reported; address limitations; and suggest directions for further research.

2. Background

We describe current practices of defining and measuring different levels of liking, as well as a branch of research formed around comparing these levels (§2.1). We note a related but separate branch of research on the relation between gender and liking, with particular respect to gender role conformity and nonconformity (§2.2). We consider language – in particular, the vocabulary people use – as a behaviour which interacts with both liking and gender (§2.3).

2.1 Liking levels

Research draws a distinction between different levels of liking. The perceptual level (how much people like each other) is contrasted to the self-perceptual level (how much people like themselves), and to the meta-perceptual level (how much people think others like them). This distinction is important because previous findings by Harnish et al. (1990) show a gap between the perceptual and self-perceptual level, with an apparent self-serving bias. On the other hand, research by Boothby et al. (2018) shows a gap between the perceptual and meta-perceptual level, with an apparent modesty bias. The latter has received extensive media attention and has been dubbed "the liking gap". The liking gap was observed and defined by Boothby et al. (2018), who drew the distinction between the perceptual level, which they call actual liking, versus the meta-perceptual level, which they call perceived liking. The difference between actual and perceived liking has been observed across communication media (Oviedo et al., 2025), age groups (Wolf et al., 2021), and group compositions (Mastroianni et al., 2021). We conclude that while these different perceptual levels might be related, they exhibit persistent differences which are of scientific interest.

However, in the data analysed by Boothby et al. (2018), most participants rated their partners as 6 or 7 on the 7-point liking scale. We argue that this leaves open the question whether the tendency to underestimate one's liking is stable across levels of actual liking, or driven by there being no room for overestimation for the majority of participants. We first aim to replicate their results using an alternative corpus. We further expand on these results by investigating how perceived liking and

the liking gap fluctuate across different levels of actual liking. We largely adopt their terminology – we expand upon this and also upon the corpus we are using in section §3.

2.2 Gender and liking

While Trofimovich et al. (2023) finds no difference in liking measures between genders, other studies investigating the perceptual level (actual liking) exclusively, find either a tendency to rate women more favourably (Eagly and Mladinic, 1994; Rudman and Goodwin, 2004), or to rate anyone exhibiting communal behaviour more favourably (Rojahn and Willemsen, 1994; Wildman and Clementz, 1986; Leaper, 1987), which corresponds to feminine gender roles (Leaper, 1987). Eagly and Mladinic (1994) coined the term "women-are-wonderful" to denote this effect. Rudman and Goodwin (2004) identifies not only a pro-woman bias in general, but also a stronger in-group bias for women. Eagly and Mladinic (1994) and Rudman and Goodwin (2004) both suggest that these effects stem from attitudes and expectations surrounding gender roles. These effects appear to be sex and gender-agnostic – Leaper (1987) studies the difference between gender (woman or man) and gender-roles (femininity/communion and masculinity/agency) and finds that femininity, rather than being a woman, correlates with likeability.

However, in some contexts, gender-role (non)conformity, rather than femininity or masculinity, has been shown to relate to social perception (Bailey et al., 2020; Chelune, 1976; Bocskor, 2024). Bailey et al. (2020) finds that images of women elicit positive reactions but only when posed in sex-congruent, contracted stances. Chelune (1976) finds that self-descriptive statements, which they note as being typically feminine, are positively correlated with liking when the speaker is a woman, but negatively correlated with liking when the speaker is a man. Kleiser Polk and Mayeux (2023) and Bocskor (2024) survey pre-adolescent and young adolescent pupils and show opposing effects of gender-role conformity for girls versus boys, depending on their countries or even on their schools.

We conclude that two distinct gender-related traits and their interaction appear to relate to liking and social perception. These are the gender identity (the gender they identify with, for example woman or man); their characteristic gender roles (their traits and behaviours that are preponderantly associated with gender, for example feminine or masculine traits); and conformity to gender roles (whether they tend to possess characteristics congruent with their gender identity, such as feminine women, or incongruent with their gender identity, such as masculine women). Gender role effects are important to consider in this research because they are foundational to identifying the traits and behaviours which underlie gender difference effects.

We first aim to replicate the "women-are-wonderful" effect of Eagly and Mladinic (1994) as well as the in-group bias effects of Rudman and Goodwin (2004) in a new corpus based on post-interaction surveying thereby expanding the existing works of Eagly and Mladinic (1994) and Rudman and Goodwin (2004) which were based on general attitudes and mental association surveying.

We additionally seek to extend our analysis to the meta-perceptual level (perceived liking) and determine whether the gender of the participants or the gender make-up of a communicating pair affect their liking gap and/or the actual and perceived liking scores given and received by them.

We advance current research by developing a measure of characteristic gender role. We investigate the relation between this behavioural trait and liking, with special attention to interaction with actual gender identity, to ensure that we capture the effect of gender role conformity and how this may differ between women and men.

2.3 Gender, vocabulary, and liking

To construct an appropriate measure of characteristic gender role which might help account for the gender differences in liking, we need to identify a behavioural trait which interacts with both gender and liking. To be able to draw strong conclusions about the ability of this measure to explain real life phenomena, we need to be able to reliably extract or construct it for large scale corpora using transparent statistical methods. Natural language processing and text analysis methods can be used to analyse the language in conversation corpora in a reproducible and replicable manner.

2.3.1 GENDER AND VOCABULARY

Gender differences in language usage, pertaining to topic choices as well as phrasing choices, have been identified by numerous studies, but we acknowledge that effects witnessed and reported by others (like those we report) may not generalize across contexts. Women's speech and writing has been found to be marked by usage of pronouns (Newman et al., 2008; Argamon et al., 2003), references to family (Newman et al., 2008; Aries and Johnson, 1983; Tannen, 1990), references to emotions and feelings (Newman et al., 2008; Leaper and Ayres, 2007; Demzik et al., 2021; Riissanen and Watson, 2014), and affiliative language indicating active understanding and support (Leaper and Ayres, 2007). Other potentially feminine language traits are the usage of present tense verbs and asking questions (Newman et al., 2008). Men's speech and writing has been found to be marked by increased usage of articles, prepositions, and numbers (Newman et al., 2008; Argamon et al., 2003), references to sport (Newman et al., 2008; Aries and Johnson, 1983), and assertive language indicating suggestions and task-focused contributions (Leaper and Ayres, 2007). Differences in talkativeness between men and women have also made the subject of a number of papers, in part motivated by an interest in the topic for pop-science (Mehl et al., 2007; Tidwell et al., 2025), but with conflicting results. With respect to what appears to be total token count, Mehl et al. (2007) and Newman et al. (2008) found no significant gender gap in word count, but a replication of Mehl et al. (2007) by Tidwell et al. (2025) which controls for age in samples from multiple countries found that women use significantly more words, an effect driven by a particularly large gap in the middle-aged adult cohorts. A meta-analysis by Leaper and Ayres (2007) found no difference in word count, also likely referring to total token count, but noted that women take more turns and

men have larger speaking durations. Some traits typically associated with women's language, such as self-references and emotional disclosures, have even been found to be inversely correlated with third party perceptions of competence, dominance, and warmth (Berry et al., 1997). However, usage of present tense is found by Berry et al. (1997) to be positively correlated with warmth perceptions, a measure which included liking.

2.3.2 VOCABULARY AND LIKING

Without particular reference to gender, previous findings show that language is indeed related to liking and social perception. One may consider relevant work on the use of language that is deemed politic or positively polite (Watts, 2003; Brown and Levinson, 1987). Within the politeness literature, one may find claims that women are more polite than men (Brown, 1980); but the focus at the moment is on preferences expressed for polite language. It has been found in analysis of online community support fora, contributions that temper claims of absolute certainty about answers are liked more than those that do not (Vogel and Mamani Sanchez, 2012) (e.g. "I think the answer is this" is preferred over "the answer is this"). Greater success in achieving financial transactions mediated online has been found to correlate with use of more polite language (Wang et al., 2024). The use of polite language does not necessarily decrease psychological distance among interlocutors; the contrary has been reported (Stephan and Liberman, 2010), and this suggests that in relation to perceptions of liking, to the extent that individuals use more polite language they may be perceived as maintaining greater distance and lower intimacy, which in turn, might be perceived as liking less. Our goal here is to discover features of language use that correspond to liking more, sensitive to interaction with gender of the language producer and consumer.

In addition to intimacy being nurtured by specific topic, lexical, or syntactic choices, liking seems to rely on perceived responsiveness. In conversations, people who ask follow-up questions are particularly well regarded (Huang et al., 2017). However, the positive effect of questions in conversations is weaker on third party observers, on whom the intimacy of the conversation and responsiveness of the speaker has no effect (Brooks and Yeomans, 2025).

2.3.3 ANALYSIS METHODS

Some research specifically discussed the relation between gender, topic choices, and interpersonal relations. However, these studies tend to be small-scale or use qualitative methods. Studies of conversations between same-gender close friends in childhood and teenage years (Tannen, 1990) and adulthood (Aries and Johnson, 1983) reveal a tendency of girls and women to discuss family matters more often or in more depth than men, while men discuss sports more often and in more depth. In contrast to younger boys, who struggle with free conversations, older boys in the study conducted by Tannen (1990) discuss remarkably vulnerable topics such as substance abuse and social alienation; however, these topics are approached in a parallel and unilateral way, with each

boy dismissing the other's concerns. This is interpreted as a cultural difference, not an empathy deficit by Tannen (1990). In spite of that interpretation, in adulthood, men discuss these topics less often or in less depth than women do (Aries and Johnson, 1983), which might contribute to a difference in perceived intimacy and, in turn, relate to liking differences.

A computational pipeline to identify the interaction of gender, topic/word choice and social interaction outcomes at a large scale can be constructed based on existing methods. To identify gender-preponderant words, chi-squared tests can be used. This test is a common method of feature selection when dealing with categorical data such as gender or word usage (Kotu and Deshpande, 2015).

Naive Bayes (NB) classifiers have been used for text-based gender classification tasks (Miller et al., 2012; Burger et al., 2011; Vashisth and Meehan, 2020). We implement a NB classifier in order to split the participants into subgroups based on whether their vocabulary is characteristically feminine or masculine. The classification accuracy of this method, as it pertains to gender prediction, seems to vary depending on task design, but is generally lower than that of more complex machine learning models (Vashisth and Meehan, 2020; Burger et al., 2011). However, human annotators also tend to be much less accurate than machine learning models at text-based gender classification tasks (Burger et al., 2011). Therefore, we argue that, based on historical performance, a NB classifier is suitable for this analysis where it is used to simulate a facet of human perception. Additionally, we elect to use methods that are fully replicable; thus, we avoid classification powered by large language models which continue to develop (cf. King (2023)).

We conclude from past work that language differences are visible between women and men and also between people higher and lower in likeability. In this paper, we study the extent to which one might identify features of the language (word choices) preponderantly associated with women and men as contributing to gender differences in any of the liking levels. Gender-preponderant words are identified as such through statistically significant contrasts within usage patterns. These form the basis of a vocabulary classification system which represents a measure of characteristic gender role expression. We explore whether differences in gender-preponderant language, as expressed through gender-preponderant vocabulary usage, explain the difference in liking between genders.

3. Data

We describe the corpus used for this study (§3.1) and detail how we processed the survey (§3.2) and conversation transcripts (§3.3).

3.1 Corpus

The data comes from the CANDOR corpus, assembled by Reece et al. (2023) through hosting a series of two-person dialogues and administering surveys prior to and after each conversation.

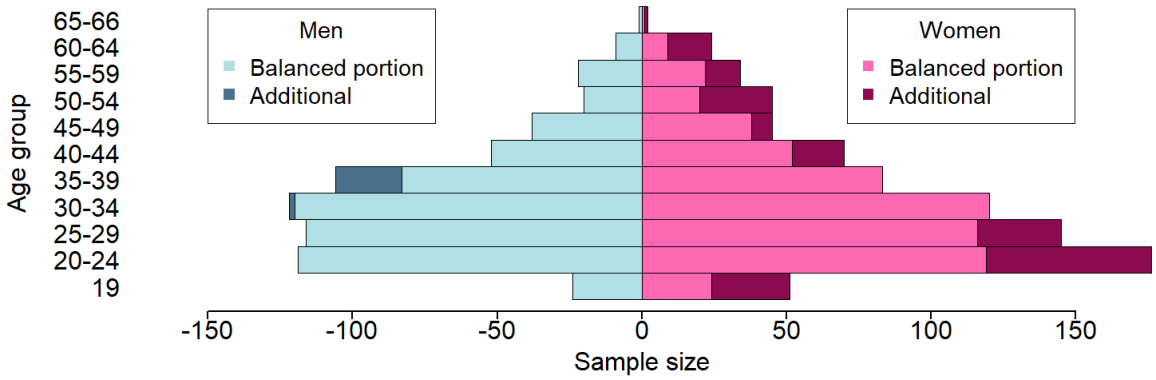

Figure 1: Demographics of CANDOR corpus participants

The dialogues were unscripted; the only direction given to participants was to have a conversation for approximately 30 minutes. The dialogues were carried and recorded over video-chat during November-January 2020, and the participants were adults living in the United States (Reece et al., 2023). Figure 1 shows the age distribution by gender of unique participants.

The corpus includes surveys that were completed by participants prior and after their dialogues, as well as automated transcripts of the dialogues. The corpus was chosen partly for its unscripted nature, which allows for naturalistic language analysis, and partly for its extensive post-conversation survey. The unscripted nature is noteworthy, as Oviedo et al. (2025) finds that the correlation between the liking gap and other conversation outcome measures, such as memory, depends on the purpose of the interaction, with differences between task-based interactions and unscripted conversation. We expect our results to generalise past the CANDOR corpus for the casual conversations of English speaking populations for which the US American adult sample is representative. We expand on the representativeness limitations later (§7.2).

3.2 Survey processing

From the survey, we used the participants' self-reported gender identity and their liking scores. A total of 1456 unique participants took part in one or more of the 1656 dialogues, resulting in 3312 dialogue-specific surveys being completed. The surveys were completed by women 1740 times (52.5% of the responses), by men 1463 times (44%), and by people indicating a nonbinary gender identity or who chose not to answer 53 times (1.6%). The remainder of 56 survey responses (1.7%) were unavailable, with no indication of gender identity. Due to most participants indicating a binary gender identity, our experiments only consider women and men when differentiating by gender.

In the survey, participants rated (1 to 7) how much they liked their conversation partner (actual liking) and how much they thought their conversation partner liked them (perceived liking). The gap between these measures is the liking gap, defined by Boothby et al. (2018) as $gap = mean(\text{perceived liking}) - mean(\text{actual liking})$. We analyse the liking gap at an individual level

Table 1: Processing example showing, for participants P1 and P2, their initial answers and resulting measurements.

				Received			Given		
	Q_1	Q_2		Actual	Perceived	Gap	Actual	Perceived	Gap
P_1	1	2	$\rightarrow$	5	2	-3	1	7	6
P_2	5	7		1	7	6	5	2	-3

Q_1: "How much did you like your partner?"

Q_2: "How much do you think your partner liked you?"

and use it to refer, for every participant, to the difference between how much their conversation partner actually liked them and how much they think their partner liked them. To compare a participant's perceived liking to their own actual liking indicated by their partner, and therefore to compute a participant's own liking gap, we differentiate between the ratings *given* by a participant (how much they like someone) and ratings *received* by a participant (how much they are liked). We therefore consider the two survey questions as referring to their actual given and perceived received liking scores, respectively. We then construct, for each participant, their actual received (AR) liking, perceived received (PR) liking and the gap in their received (GR) liking on the one hand, and their actual given (AG) liking, perceived given (PG) liking, and the gap in their given (GG) liking on the other hand. Table 1 exemplifies how one gets, for a pair of participants, from two sets of two values each to the two sets of six values each that we just described. Note that, for each pair of participants P1 and P2, P1's received values are the same as P2's given values and vice-versa. A negative gap value suggests an underestimation of one's likeability.

3.3 Text processing

We annotated the conversation transcripts using TreeTagger, developed by Schmid (1997, 1999) and the TreeTagger Python Wrapper developed by Pointal (2015). We used an English parameter file trained on the British National Corpus (BNC) with the BNC Basic Tagset. TreeTagger annotations contain the initial word token, the part of speech (POS), and the lemma. Punctuation, unclassified items (a possible POS outcome in the BNC Basic Tagset), and alphabetical symbols were excluded because they were frequently associated with auto-transcriber errors.

Using the tagged transcripts, we created lists containing each word uttered. We recorded, for each participant, whether they uttered each word and if so, how many times. These two records form the basis of two counting methods we contrast in this paper: people counts and utterance counts. People counts represent how many people said a word, while utterance counts represent

how many times a word was said. For a single pair of a participant and a word, the people count takes binary values of 1 (the participant said the word) or 0 (the participant did not say the word), while the utterance count takes integer values (including zero) corresponding to how many times the participant said the word. For each participant, the people counts result in the set of all words said, while the utterance counts result in the multiset of all words said. The individual participant-word counts can be then aggregated across participant groups of interest, as we describe in section §4.3.

4. Methods

We focused on developing an analysis pipeline based on deterministic mathematical models and fundamental statistics. This ensures that our methods are transparent and our results are explainable and reproducible, and do not rely on black-box methods or random initialisation. This also enabled us to ensure that the underlying assumptions of the statistical models and tests used were fulfilled. We describe the methods used to test for the presence and properties of a liking gap (§4.1) as well as the interaction of gender with the liking levels and the liking gap (§4.2). We further describe the methods used to compile a list of gender-preponderant words (§4.3) and to analyse the interaction of gender and gender-congruency of individual vocabulary, with the liking levels (§4.4).

4.1 Liking levels

Using R we created data visualisations to show the distribution of actual and perceived liking scores across participants and how participants tend to rate themselves given their partners' actual opinions of them. Also in R, we computed explanatory statistics for the data. For both the explanatory statistics and the visualisations, we split the perceived and gap data across levels of actual liking. Since the actual liking data does not follow a normal distribution, we used the non-parametric Wilcoxon signed rank test to compare actual and perceived liking, testing the null hypothesis that the median liking gap is zero.

4.2 Gender and liking

Using python, we kept track, for each participant, of their partner's gender, compiling a dataset that contained, for each participant, their own reported gender, their partner's gender, and the participants' actual liking (reported by their partner) and perceived liking (reported by themselves). Using R, we created visualisations to depict the differences between the way genders report and perceive liking scores. Also in R, we computed explanatory statistics for the data. For actual and perceived liking scores, as well as their gap, we used Wilcoxon rank-sum test to compare the score distribution between women and men in general and between woman-woman, woman-man, man-woman, and man-man pairs specifically. The null hypothesis for each test is that there is no difference between subsamples in how they report, perceive, or misestimate liking.

4.3 Gender and vocabulary

We compiled a list of words which can be shown to be gender-preponderant based on deviations from statistically expected usage patterns that were determined to be statistically significant using Chi-squared tests (with the threshold accepted for α or p adjusted in light of multiple testing).

Based on gender, we aggregated the individual participant word counts obtained using each of the two counting methods described in section §3.3. In this context, people counts represent, for each group (e.g. "Women") and each word (e.g. "book") how many participants said the word at least once during their recorded conversation. Utterance counts represent, for each combination of group and word (e.g. "Women" - "book") how many times the word was said by participants in that group. Using these values we created two series of contingency tables (one for each counting method) comparing gender sample splits (e.g. "Women v Men" - "book"). Using the people counting method, we compared, for each word, the number of people of each gender who said the word vs the number of people of each gender who never said the word. Using the utterance counting method, we compared, for each word, the number of times the word was said by people of each gender vs the number of times any other word was said by people of each gender.

We then performed a series of Chi-squared tests for each table to identify gender-preponderant words. As a threshold for inclusion, we required a minimum value of 5 in each cell of the contingency table and expected values table. Lemmas for which we rejected the null hypothesis and observed different usage between women and men were added to the word list corresponding to the gender which used the word more often. Because we performed a test for each lemma, we used a Bonferroni-adjusted p-value of $\frac{5\%}{\#lemmas}$ as the significance level threshold for rejecting the null hypothesis.

To gain insight into the gender-significant words, we classified them into categories. The categories that emerged represent semantic fields (e.g. Family), parts of speech (e.g. Interjections), or functional categories (e.g. Modality). Since these categories are used for discussion and display purposes, and not for modelling, this process was done simply by manual annotation. Each word was included in a single, best-fitting category. The categories were determined by manually passing over the list of strings and sequentially expanding the category inventory and updating the elements as needed. When similar categories overlapped with each other they were sometimes merged; "Family" and "Partners", overlapping on "wife" and "husband" were merged into "Family". Categories were not merged if merging would combine multiple disparate categories; "Sports" overlapping with "Games" ("player", "game") and "Games" overlapping with "Tech" ("vr", "console") were kept separate to avoid a "Sports-Tech" megacategory. In these cases, words were sorted into what we deemed was the best-fitting category. Where the best-fitting category was ambiguous due to context-dependence, we used a personal implementation of a Key Word In Context (KWIC) search, which takes in a string and a context window size and outputs, for every case-insensitive instance of that string in the transcripts, the turn in which it appeared and neighbouring turns within

the specified window. In this case, context-dependent words were classified based on the context which appeared to be dominant. We reported how many words fell into each category, separated by corresponding gender, and we provided the full list of words for each category, separated by corresponding gender.

4.4 Vocabulary and liking

Based on the gender-preponderant words identified in the previous step, we used a simple model which predicted participants' gender, with the goal of analysing the association between gender, gender-congruency, and liking.

We used a naive Bayes (NB) classifier, a method which we argue to be suitable based on past results (see §2.3) and our own research goals. The classifier computed, for every participant, the probability that they are a woman based on whether or not they used each of the gender-significant words. The probability was computed using formula 1, where:

- $p \in P$ is a participant p in the set of all participants P
- $♀_p$ and $♂_p$ are the events where participant p is a woman or a man, respectively. These events are disjoint. $G_p \in \{♀_p, ♂_p\}$ is the event where participant p has gender G.
- $P(♀_p)$ and $P(♂_p)$ are the probabilities that participant p is a woman or a man, respectively. $P(♀_p) + P(♂_p) = 1$. $P(G_p)$ is the probability that participant p has gender G.
- V_p is the vocabulary set of participant p
- $w \in F \cup M$ is any word w belonging to either of the gender-preponderant sets F or M;
- $\mathbb{1}_{V_p}(w)$ is the indicator function of V_P and takes value 1 if $w \in V_p$ and 0 otherwise
- $P(♀_p|V_p)$ is the probability that participant p is a woman, given the participant p's vocabulary. More specifically, given whether the participant did or did not utter each of the gender-preponderant words, it is the probability that the participant is a woman. $P(♂|V_p) = 1 - P(♀_p|V_p)$
- $P(\mathbb{1}_{V_p}(w)|♀_p)$ is the probability of the outcome of $\mathbb{1}_{V_p}(w)$ assuming that the participant is a woman. When $\mathbb{1}_{V_p}(w) = 1$, this is equivalent to $P(\mathbb{1}_{V_p}(w) = 1|♀_p)$, or the probability that the word w is in participant p's vocabulary V_P given that p is a woman. When $\mathbb{1}_{V_p}(w) = 0$, this is equivalent to $P(\mathbb{1}_{V_p}(w) = 0|♀_p)$, or the probability that the word w is *not* in participant p's vocabulary V_P given that p is a woman. $P(\mathbb{1}_{V_p}(w)|♂)$ is analogous.

The NB probability measures were computed in the same way using the people and utterance counting methods, with the two differing only in the composition of the $F \cup M$ gender-preponderant word sets.

$$P(♀_p|V_p) = \frac{P(♀_p)\prod_{w\in F\cup M} P(\mathbb{1}_{V_p}(w)|♀_p)}{\sum_{G\in\{♀_p,♂_p\}} P(G_p)\prod_{w\in F\cup M} P(\mathbb{1}_{V_p}(w)|G_p)} \quad (1)$$

Based on the predictions, participants were then split depending on whether their vocabulary was rather feminine or rather masculine. Together with their reported genders, this resulted in four groups:

1. $♀_F$: women with feminine vocabularies (gender-congruent, "accurately" classified)
2. $♀_M$: women with masculine vocabularies (gender-incongruent, "inaccurately" classified)
3. $♂_F$: men with feminine vocabularies (gender-incongruent, "inaccurately" classified)
4. $♂_M$: men with masculine vocabularies (gender-congruent, "accurately" classified)

The goal of the classifier is not to maximise accuracy; rather, our aim is to study the accurately and inaccurately classified groups. We are interested in the (relative) size of these groups. The accuracy of the NB model using each of the counting methods, and the size of each subgroup were reported and compared. We further report the Spearman correlation between gender and vocabulary on the one hand, and liking on the other. We also report the average liking ratings for each of the gender-vocabulary groups. We use un-paired Wilcoxon signed rank tests to test the hypothesis that there are differences in terms of given or received liking ratings for each of the six pairwise comparisons between the four groups.

Other classification methods were considered, including a Logistic Regression based on the number of preponderantly feminine and masculine words, respectively, used by each participant. We intend to report regression-anchored analyses separately.

5. Results

We describe and visualise the results that replicate past work by Boothby et al. (2018) and confirm the presence of a liking gap, as well as our novel results that uncover new and interesting properties of this phenomenon in §5.1. We further expand on these by showing how gender and pair gender make-up interact with levels of liking in §5.2. We describe the list of gender-preponderant words in §5.3. Finally, in §5.4 we show how gender and vocabulary interact with the liking levels. We discuss the implications of these results in section 6.

5.1 Liking levels

Figure 2 shows the distribution of the actual and perceived liking scores. Actual liking mean and median values are very close, 5.99 and 6, respectively, but the mode value is equal to the maximum possible value of 7. This means that most participants gave actual liking values of 6 or 7, with

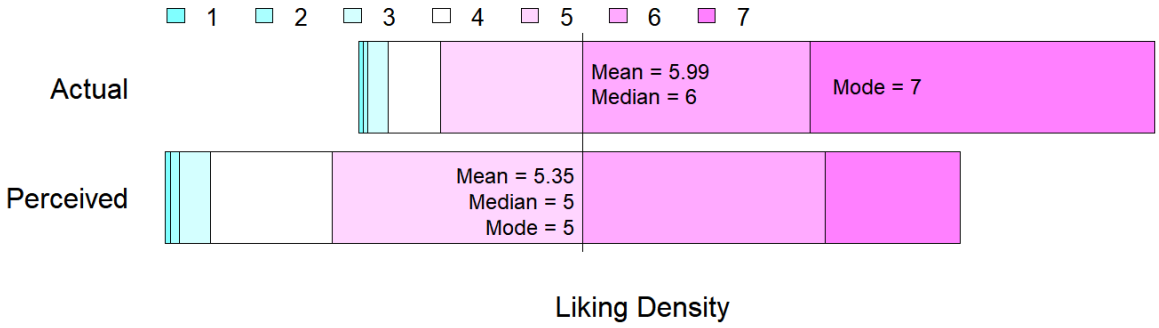

Figure 2: Distribution of actual and perceived liking liking

7 being the most common, and lower values being decreasingly common. Perceived liking, on the other hand, has its mode and median values at 5, and a mean of 5.35. Explanations of the psychological mechanisms behind the liking gap need to account for these differences in median and mode values between actual liking and perceived liking.

Actual and perceived correlation is positive ($r = 0.21$, $n = 3206$), showing that, although participants are rarely perfectly accurate in their perception of their own likeability, their perception still mirrors reality somewhat. This is consistent with the previous finding that the liking gap tends to be off by only a few points. A Spearman rank correlation test leads to rejecting the hypothesis of $\rho = 0$ ($p < 0.001$) in favour of the alternative $\rho \neq 0$.

However, the paired Wilcoxon Rank Sum Test did reveal a statistically significant difference ($p < 0.001$) between actual and perceived liking. This confirms the findings of Boothby et al. (2018) of a statistically significant liking gap. The gap has its median and mode values at -1, and a mean of -0.65 with a standard deviation of 1.49. Participants underestimated their likeability a majority of times: 55%, compared to 27% accurate perception and 18% overestimating. However, when participants do under or over-estimate their likeability they usually do so by a small margin (1-2 points away on the ordinal scale)

Figure 3 shows the frequency of each pair of actual-perceived values. The size of the bubble is proportional to the number of times a participant received the actual liking value corresponding to the x-axis bubble position and reported the perceived liking value corresponding to the y-axis bubble position. The count is also listed inside each corresponding bubble. While the fact that low values are uncommon for both actual and perceived liking is consistent with previous results, this bubble graph also shows that perceived liking values are most often in the 4-6 range with modes of 5 for most actual liking values. The median perceived liking value is also equal to 5 for all values of actual liking, except 7/7 where median perceived liking is 6. The weakly positive correlation we found earlier is visible, but so is its low magnitude.

Given that perceived liking values tends to be around 5, and a majority of actual liking values given are 6 and 7, it follows that the liking gap will be most often negative. In the 43% of times when

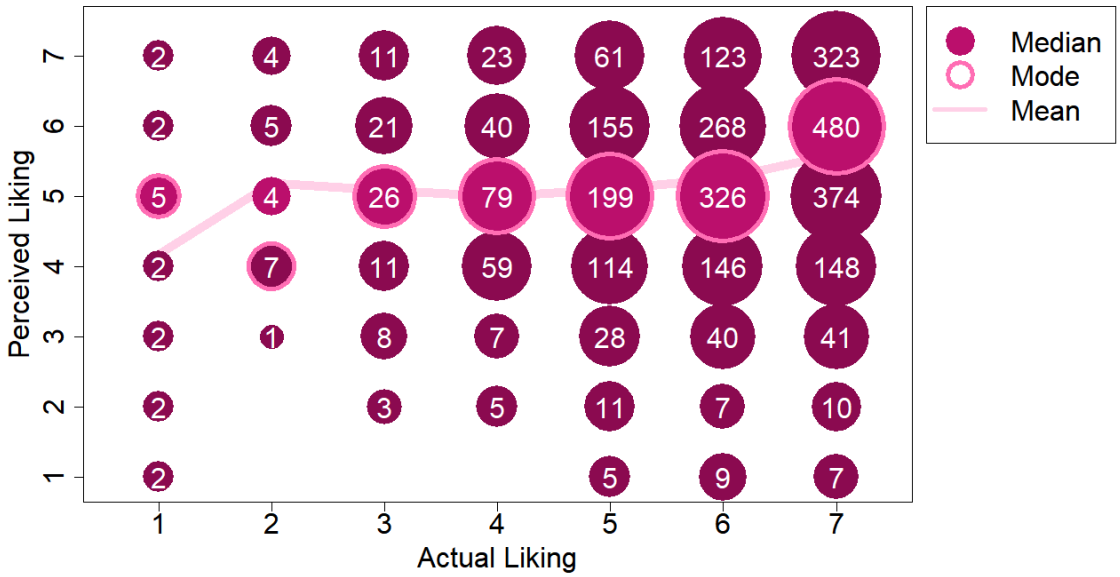

Figure 3: Frequency of perceived liking levels across actual liking levels

participants received maximum actual liking scores, it would be impossible for them to overestimate their likeability. Moreover, participants very rarely (17% of times) even report perceived liking ratings of 7. However, they just as rarely (21% of times) report perceived liking ratings of 4 or less. This leads to the low-actual/low-perceived quadrant in figure 3 being the most sparsely populated. In fact, participants who received actual liking ratings of 4 or less more commonly overestimate their likeability rather than under or correctly estimate it.

Figure 4 further decomposes the gap distribution. Participants who received actual liking scores of 5 tend to correctly estimate, while participants who received scores below or above 5 tend to overestimate and underestimate, respectively, with the size of the miss-estimation increasing two- and three-fold at the extremes. The liking gap is thus not evenly sized and directed, but instead varies across levels of actual liking. The negative median values are the result of the high frequency of high actual liking reports.

If there were a general tendency of people to slightly underestimate their likeability with a mean and variance similar to the one observed in the CANDOR corpus data but evenly spread across the actual liking range, we would expect the bottom-left quadrant of figure 3 to be more populated, while the top-left quadrant would be nearly empty. Furthermore, we would expect the mean liking gap values in figure 4 to be negative across most of the actual liking range except for the lowest values, and we would expect the magnitude of the gaps, especially the positive gap at the lower end of the actual liking spectrum, to be smaller.

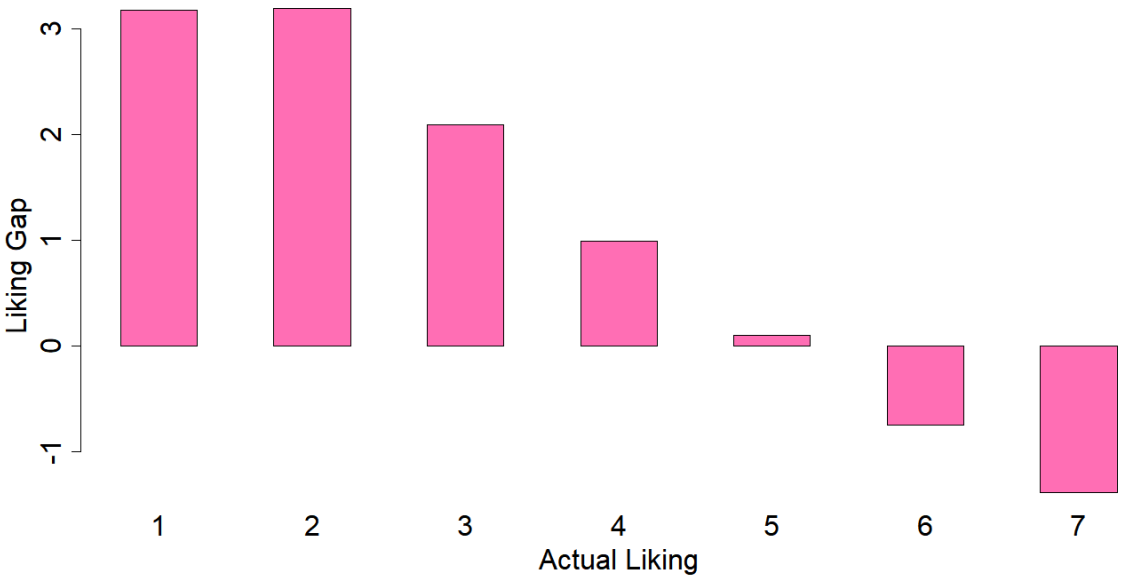

Figure 4: Average liking gap at each level of actual liking

5.2 Gender and liking

Table 2 shows the average liking ratings received and given by women and men and the results of the two-sample Wilcoxon Rank Sum tests for each type of liking rating. Given that the data is ordinal, the decimal value of the means only serves to determine whether the tendency of a group is to give or receive smaller or larger ratings compared to the other group. Based on the means, the gender which is associated with higher values is indicated (W for women and M for men) alongside the significance level for the cases where we rejected the null hypothesis. Women give and receive higher ratings, perceive higher ratings, and are perceived as giving higher ratings. A two-sample Wilcoxon Rank Sum test leads to rejecting the null hypothesis that AR, PR, AG, or PG liking is equal between men and women in the represented population (see section §3.2, ff. 9, for a detailed explanation of actual received (AR), perceived received (PR), actual given (AG), and perceived given (PG) liking). We conclude that women have higher AR, AG, PG ($p < 0.001$) and PR ($p < 0.01$) liking ratings. This can mean that women are seen as more likeable than men, believe themselves to be more likeable than men do, like people more than men do, and make people feel more liked than men do. Similarly, we reject the null hypothesis that both genders have the same received liking gap. Table 2 shows that men's liking gap is still negative but smaller in magnitude than women's. This means that while men also underestimate their liking more often than they overestimate it, they underestimate less often than women. We cannot reject the null hypothesis regarding the difference between genders in terms of given liking gap. While women's conversation partners have an average given liking gap that's smaller in magnitude than men's conversation partners' ($|-0.63| < |-0.67|$), this difference is not statistically significant.

Table 2: Average liking rating depending on participant gender and the test results for differences between them.

	Received			Given		
	Actual	Perceived	Gap	Actual	Perceived	Gap
Women	6.11	5.4	-0.71	6.05	5.43	-0.63
Men	5.86	5.3	-0.58	5.93	5.26	-0.67
Women vs Men	♀***	♀**	♂**	♀***	♀***	

* ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

Table 3: Average value of the liking ratings given by participants

Giving (P1)	Receiving (P2)	Actual (P1 rating P2)	Perceived (P2's perception)	Gap (P2's error)
Women	Women	6.19	5.49	-0.7
	Men	5.9	5.35	-0.54
Men	Women	6.03	5.29	-0.74
	Men	5.81	5.23	-0.58

This means that the gender of one's dialogue partner cannot be shown to interact with one's accuracy in estimating one's likeability.

Table 3 shows the mean liking rating values given by participants broken down by the participants' gender and their dialogue partner's gender. The exact mean values establish the ordering. Table 4 shows the results of the two-sample Wilcoxon Rank Sum tests performed between each gender pair subset of participants. Figure 5 shows the distribution of each liking type for each gender pair subset. Women gave women the highest actual and perceived scores, higher than any other pairing ($p < 0.001$). This implies that, compared to all other pairs, women like each other more and feel more liked. This is visible in the figure 5 distributions, with women having the highest tendency to give each other the maximum rating and being the only group to guess each other's rating as 6 more often than 5. Men give women higher ratings than men give other men ($p < 0.01$) - the distributions show this comes from more 6 and 7 ratings and fewer 4 and 5 given by men to women than to men. This means that men like women more than men like men. Finally, the relative given liking gaps resulting from women talking to men are higher (lower in magnitude if negative) compared to men or women talking to women ($p < 0.05$). This means that men talking to women underestimate their likeability less than women talking to women or women talking to men.

5.3 Gender and vocabulary

There were 400 words significantly associated with the gender of the speaker. Using the people counting method, we identified 45 preponderantly feminine words and 45 preponderantly masculine

Table 4: Results of two-sample Wilcoxon Rank Sum tests for the difference between given liking ratings given depending on participant - pair gender combination

	Test	Question	Answer
Actual	WW vs WM	Whom do women prefer?	♀***
	MW vs MM	Whom do men prefer?	♀**
	WW vs MW	Who likes women?	♀***
	WW vs MM	Who likes their own gender?	♀***
Perceived	WW vs WM	Who feels liked by women?	♀***
	WW vs MW	Who makes women feel liked?	♀***
	WW vs MM	Who feels liked by their own gender?	♀***
Gap	WW vs WM	Who underrates when talking to women?	♀*
	WM vs MW	(...) to the opposite gender?	♀*

* ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

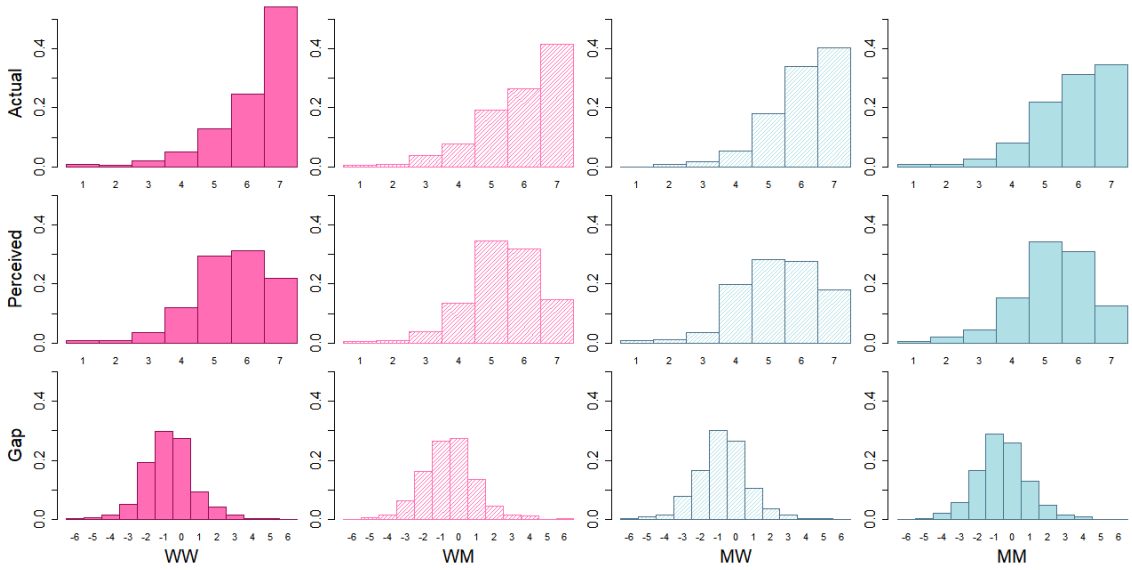

Figure 5: Distribution of given liking ratings for each pair gender combination

Table 5: The number of words in each category that saw significantly higher usage by women or men. The number of words identified using the people-counting method as well is shown in (brackets) if it is larger than zero.

	Women	Men		Women	Men
Health	14 (2)	0	Tech	0	30 (5)
School	10 (2)	0	Finance	0	25 (2)
Pets	6 (3)	0	Sports	0	15 (6)
Holidays	3	0	Games	0	9 (4)
Appearance	3	0	Politics	0	10 (1)
Feelings	26 (12)	4 (1)	Entertainment	0	8
Family	22 (14)	3 (1)	Quantities	1	43 (12)
Household	6 (1)	3	Location	3	26 (1)
People	3 (2)	6 (3)	Quality	4 (1)	16 (2)
Social	1	2	Time	1	6 (1)
Interjections	14 (6)	7 (2)	Articles & Prepositions	0	5
Verbs	7	4	Pronouns	4 (1)	7
Links	5	5 (1)	Fillers & Hedges	1	20 (2)
Intensifiers	5 (1)	3 (1)	Questions	2	2
Total				**141 (45)**	**259 (45)**

words. Using the utterance counting method, we identified 141 preponderantly feminine words and 259 preponderantly masculine words, including all the words identified by the people-counting method.

The words were manually classified into semantic fields, part of speech categories, or functional categories. Table 5 the number of preponderantly feminine and masculine words, respectively, from each category. Table 6 shows some examples from each category identified. Table A in appendix §A shows the full list of gender-preponderant words, grouped by category.

Compared to men, women tended to use words related to health, school, pets, holidays, and appearance.

A few categories were dominated by women, although men also had some preponderantly masculine words belonging to these categories. Women used a wider range of words denoting or implying feelings, while men only used some words expressing enjoyment more often. Women tended to use a lot of different words denoting family members, partners, and family relationships, compared to men. Women use a wider range of words related to the household, however men also have some food and home specific words. Feminine speech was also marked by more preponderantly feminine verbs and interjections, compared to masculine speech.

Compared to women, men tended to use words related to technology, finance, sports, videogames, politics, and entertainment. Their speech was also marked by specific articles or prepositions, and filler words or hedges.

Table 6: Preponderantly feminine/masculine word examples from each category identified.

	Feminine	Masculine
Health	sick, therapy, mask	
School	undergrad, school, teacher	
Pets	cat, bark, pet	
Holidays	Christmas, holiday, birthday	
Appearance	makeup, hair, wear	
Tech		programmer, Java, hardware
Finance		buck, stocks, tax
Sports		baseball, Kobe, NBA
Games		gaming, Nintendo, Discord
Politics		senator, war, Trump
Entertainment		sequel, casino, stream
Feelings	hope, cry, lovely	fan, enjoy, appreciate
Family	parent, sister, pregnant	folk, girlfriend, wife
Household	home, grocery, pie	burger, build, beer
Quantities	many	miles, million, mathematics
Location	Pennsylvania, Philly, to	Orlando, area, west
Quality	allowed, nice, old	aspect, decent, correct
People	girl, woman, friend	dude, guy, buddy
Time	then	quick, month, ago
Social	tell	call, mention
Articles & Prepositions		a, by, for
Interjections	hi, gosh, please, um	hey, damn, pleasure, uh
Verbs	be, do, have	make, get
Pronouns	I, she, he	this, that, it
Fillers & Hedges	like	whatever, kinda, whatnot
Links	also, because, but	example, if, or
Intensifiers	really, super, totally	gotta, actually, certainly
Questions	who, why	question, search

Table 7: Subgroup size and prediction accuracy under each counting method

	People		Utterance	
	Feminine	Masculine	Feminine	Masculine
Women	1428 (82% ♀)	312 (18% ♀)	891 (51% ♀)	849 (49% ♀)
Men	299 (20% ♂)	1164 (80% ♂)	47 (3% ♂)	1416 (97% ♂)

Some categories were dominated by men, but women also had some preponderantly feminine words representing those categories. Men used a lot of numbers and quantitative topics, while women only used one. Men used a wide range of preponderantly masculine words describing and locations, while women only had three preponderantly feminine location words. Men had a larger number of preponderantly masculine words to express qualitative opinions and facts that do not necessarily evoke feelings compared to women. Men referenced more people in contexts other than family, sports, or entertainment than women. Men used words related to time, while women only had one preponderantly feminine time word. Men had a larger number of preponderantly masculine pronouns than women.

Finally, a few categories of words were more evenly split between preponderantly feminine and masculine words, such as words related to socialising, linking words, intensifiers, and interrogative words or words marking unknown answers.

5.4 Vocabulary and liking

Figure 6 shows the distribution of the NB model predictions using each of the two counting methods. The predicted value is the probability, according to each model, that a participant is a woman. The lower extreme represents participants who were considered by the model to be likely men, and the higher extreme represents likely women. Participants tend to be classified as either very likely men or very likely women. The ratio of women to men in the corpus is approximately 6:5 in favour of women. Using the people-counting method, 1730 participants are classified as likely women and 1473 as likely men, a ratio of approximately 6:5. Using the utterance method, only 997 participants are classified as likely women, and 2206 as likely men, a ratio of approximately 2:5 in favour of men.

Table 7 shows the prediction accuracy and the number of participants in each group. The model predicts women's gender with 82% accuracy and man's with 80% accuracy when using the people counting method. However, when using the utterance counting method, the model predicts participants' gender with only 51% accuracy if they are women, and 97% if they are men.

Table 8 shows the correlation of liking scores with reported gender identity and with the vocabulary-based gender predictions computed using each counting method. The correlations of AR, PR, AG,

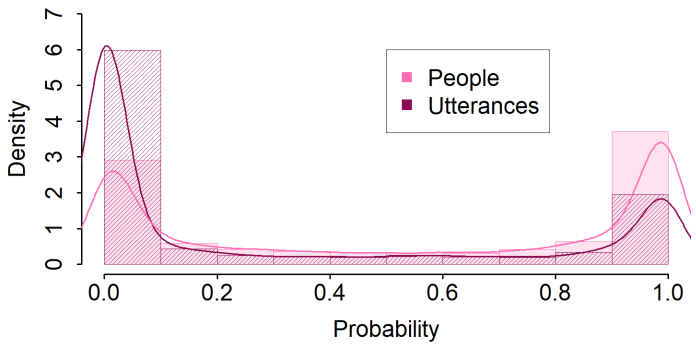

Figure 6: Distribution of the probabilities that a participant is a woman

and PG liking with vocabulary type are similar in magnitude between the vocabulary types determined using the two counting methods, and also similar in magnitude and equal in sign to the correlation with reported gender identity. The null hypothesis that the correlation between these liking scores and vocabulary type in the represented population is rejected for both vocabulary classifications; all the confidence levels are above 90%, and the specific p-values are indicated in table 8. This means that, regardless of the vocabulary model used, people with a feminine vocabulary tend to be better liked, to feel more liked, to like others better, and to make others feel more liked. We also note that the correlation of these four liking scores with vocabulary is higher in magnitude than the correlation with the reported gender identity when using the people-count-based vocabulary model.

The correlations of GR with vocabulary under each of the models is equal in sign to the correlation of GR with reported gender identity but smaller or equal in magnitude (see section §3.2, ff. 9, for a detailed explanation of received liking gap (GR) and given liking gap (GG)). Moreover, only when using utterance counting can the null hypothesis of no correlation in the larger population be rejected at a 95% or higher confidence level. This can be taken to mean that women and, by some measures, people with feminine vocabularies, overestimate their likeability to a greater extent. Neither reported gender identity nor vocabulary type correlate with GG, meaning that neither one's gender nor the femininity or masculinity of their vocabulary can be shown to correlate with others' accuracy in estimating their own likeability.

Table 9 shows the average liking scores for the four subgroups formed based on self-reported gender and predicted gender based on vocabulary using each of the counting methods. The table contains six sectors formed by combinations of the two counting methods and three liking score types (Actual, Perceived, Gap). Within each of these sectors, the cell colours follow a range from blue corresponding to the lowest magnitude, to pink at the highest magnitude.

Looking at actual and perceived liking, there seems to be a divide between vocabulary type, rather than gender, when using the people counting method. Women and men with feminine vocabularies tend to have high AG, PR, and PG scores, while women and men with masculine vocabularies

Table 8: Correlation of liking levels with gender and vocabulary using people (P) and utterance (U) counting

	Received			Given		
	Actual	Perceived	Gap	Actual	Perceived	Gap
Gender	0.11***	0.04**	-0.05**	0.08***	0.06***	0.01
Vocab (P)	0.12***	0.08***	-0.03	0.12***	0.1***	0
Vocab (U)	0.09***	0.03.	-0.04*	0.07***	0.06***	0

. ($p < 0.1$), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

tend to have lower scores. The AR scores show women and men with gender-congruent vocabularies having, on average, the highest and lowest scores, respectively; people with gender-incongruent vocabularies have average scores in-between the gender-congruent extremes. This means that, at the metaperceptual level, people with feminine vocabularies experience and participate in more positive social experiences (feel more liked and make others feel more liked). They are also the ones to give higher scores. However, the best and least liked are feminine women and masculine men, respectively.

When using the utterance counting method, men with masculine vocabularies tend to have the lowest actual and perceived liking scores, smaller GR, and larger GG. Women tend to have higher actual and perceived liking scores, especially those with feminine vocabularies. The group of men with feminine vocabularies is very small (only 47 data points), but they appear to have high AR and PG scores and low AG and PR scores. This means that, compared to men, women tend to be and feel better liked, report liking their partners more, and make their partners feel better liked, all effects stronger in women with feminine vocabularies using utterance-counting. Men with feminine vocabularies might be better liked and make their partners feel better liked, but they themselves report feeling less liked and liking other less.

Looking at the liking gap, the divide seems to be along the gender lines rather than vocabulary lines when using the people-counting method. Women tend to have larger GR and lower GG than men. This means that the NB model with people-counting does not capture traits or behaviours that interact with the size of the liking gap. A similar effect is visible when using the utterance-counting method, with the exception of the small group of men with feminine vocabularies, who have the largest GR and elicit the lowest GG.

To test which of the previous effects are generalisable for the entire represented population, table 10 shows, for each counting method and combination of gender-vocabulary subgroups, which subgroup is associated with a higher score. For the liking gap scores, the subgroup with the more positive gap (lower in magnitude if negative) is indicated.

When comparing participants with gender-congruent vocabularies, we rejected the null hypothesis and found significant differences in terms of actual and perceived liking between the subgroups;

Table 9: Average liking for each identity & vocabulary subgroup

		Actual		Perceived		Gap	
		Received	Given	Received	Given	Received	Given
People	♀$_F$	6.14	6.10	5.43	5.47	-0.71	-0.63
	♀$_M$	5.95	5.84	5.23	5.21	-0.71	-0.64
	♂$_F$	5.97	6.10	5.43	5.44	-0.53	-0.66
	♂$_M$	5.83	5.89	5.27	5.22	-0.57	-0.67
Utterance	♀$_F$	6.15	6.10	5.40	5.48	-0.75	-0.62
	♀$_M$	6.06	6.01	5.39	5.37	-0.67	-0.63
	♂$_F$	6.00	5.96	5.23	5.43	-0.76	-0.50
	♂$_M$	5.85	5.93	5.30	5.26	-0.56	-0.67

women give and receive higher scores regardless of counting method used. We found significant differences between women and men with gender-incongruent vocabularies only when using people-counting. Specifically, we found that when defining vocabulary type using the people-counting method, men with feminine vocabularies have higher AG, PR, and PG ($p < 0.05$) scores than women with masculine vocabularies. Comparing women, we found that those with feminine vocabularies have higher AR, AG, PG ($p < 0.001$ using people-counting; $p < 0.05$ using utterance counting), and PR ($p < 0.01$ only using people-counting). Similarly, when comparing men, those with feminine vocabularies were associated with higher scores AR, PR ($p < 0.05$), AG, and PG ($p < 0.01$) but only when using people-counting. Among participants with feminine vocabularies, few significant differences arise; women have higher AR scores ($p < 0.5$) only when using people counting and higher AG ($p < 0.1$) scores only when using utterance counting. Comparing participants with masculine vocabularies, women were found to have higher AR ($p < 0.001$), AG, PG ($p < 0.05$), and PR ($p < 0.1$) scores, but only when using utterance-counting to determine vocabulary. Men with masculine vocabularies have less negative GR compared to women regardless of vocabulary type or counting method. No significant differences were found in terms of GG.

6. Discussion

Our findings confirm previous reports of a liking gap (Boothby et al., 2018; Oviedo et al., 2025; Mastroianni et al., 2021) and expand current understanding of this intersubjective perceptual discrepancy. We find that most people in the sample rate themselves as average in terms of likeablity, regardless of how they are actually perceived, while at the same time rating others as high or very high in likeability. We show how these two effects give rise to a liking gap whereby, on average, people underestimate their likeability. Another implication of this finding is that the those low or medium in likeability, albeit few, tend to overestimate how others perceive them. Boothby et al.

Table 10: Test results for differences at each liking level between gender and vocabulary sub-groups, using people (P) and utterance (U) counts respectively.

		Actual		Perceived		Gap	
		Received	Given	Received	Given	Received	Given
Congruent:	P	$♀_F^{***}$	$♀_F^{***}$	$♀_F^{***}$	$♀_F^{***}$	$♂_M^{*}$	x
$♀_F$ vs $♂_M$	U	$♀_F^{***}$	$♀_F^{***}$	$♀_F^{**}$	$♀_F^{***}$	$♂_M^{**}$	x
Incongruent:	P	x	$♂_F^{*}$	$♂_F^{*}$	$♂_F^{*}$	x	x
$♀_M$ vs $♂_F$	U	x	x	x	x	x	x
Women:	P	$♀_F^{***}$	$♀_F^{***}$	$♀_F^{**}$	$♀_F^{***}$	x	x
$♀_F$ vs $♀_M$	U	$♀_F^{*}$	$♀_F^{*}$	x	$♀_F^{*}$	x	x
Men:	P	$♂_F^{*}$	$♂_F^{**}$	$♂_F^{*}$	$♂_F^{**}$	x	x
$♂_F$ vs $♂_M$	U	x	x	x	x	x	x
Feminine:	P	$♀_F^{*}$	x	x	x	x	x
$♀_F$ vs $♂_F$	U	x	$♀_F.$	x	x	x	x
Masculine:	P	x	x	x	x	$♂_M.$	x
$♀_M$ vs $♂_M$	U	$♀_M^{***}$	$♀_M^{*}$	$♀_M$	$♀_M^{*}$	$♂_M^{*}$	x

. ($p < 0.1$), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$)

(2018) propose, as a potential explanation for the liking gap, the incentive to focus on conversational mistakes in order to fine tune conversational skills. Under this paradigm, it is possible that those who do not focus on conversational mistakes will neither exhibit a negative liking gap, nor have good conversational skills, explaining the low/medium likeability group. We have not yet tested, for example, dysfluency rates, or other behaviours that may be deemed conversational mistakes (Tatchell et al., 1983), nor properties such as turn taking balance, which have been shown to relate to external perception of conversational collaboration (Vogel et al., 2023). Our findings could be interpreted as evidence of a Dunning-Krueger effect in social perception. This phenomenon arises when patterns of cognition and metacognition or perception and metaperception interact such that bottom quantiles overestimate their abilities while top quantiles underestimate them (Kruger and Dunning, 1999). Our setting has the property that the majority of participants is concentrated at the high end of the likeability spectrum.

We also found that women reported more enjoyable social experiences in conversations. They were better liked and liked their partners more, compared to men. Our findings are consistent with the "women-are-wonderful" effect (Eagly and Mladinic, 1994) and confirm the stronger in-group bias of women compared to men (Rudman and Goodwin, 2004). However, we also find that men rated women as more likeable than men rated other men. This casts doubt on the interpretation of neutral cultural differences between women's style of conversation and men's style of conversation whereby men would be indifferent or even prefer a masculine conversation style (Tannen, 1990).

The list of gender-preponderant words can help us better understand these effects. We confirm previous findings (Newman et al., 2008; Aries and Johnson, 1983; Tannen, 1990; Leaper and Ayres, 2007; Demzik et al., 2021) that women's speech is marked by references to family and emotions, and additionally reveal a comparatively stronger tendency of women to discuss health, school, pets, and holidays. Men's speech was marked by increased usage of articles and prepositions, references to sport, and increased usage of numbers and references to quantities, confirming previous findings (Newman et al., 2008; Argamon et al., 2003; Aries and Johnson, 1983). We expand on this by showing that, compared to women, men also discussed more often and in much more detail topics such as finance, technology, video-games, and politics.

Another interesting and novel point of contrast between feminine and masculine speech surrounds different ways of addressing the same concept or topic. Women and men had different ways of referring to parents (”parents” vs ”folks”) and friends (”friend” vs ”buddy”), different markers of politeness (”thank” vs ”pleasure”), salutations (”hi” vs ”hey”), and different versions of the same filler word (”anyway” vs ”anyways”). Women and men also had different interjections and phatics: ”goodness/gosh” vs ”damn/hell”, ”yes” vs ”yeah”, ”um” vs ”uh”, the last being in line with the findings of Acton (2011). Women's tendency to directly name specific emotions is also contrasted by men's tendency to use adjectives of quality which imply approval or disapproval but do not directly express their feelings. Finally, gender-preponderant pronouns reveal women's tendency to refer to people or animate objects and contrast it to men's tendency to refer to concepts or inanimate objects.

A number of potential explanations exists for these differences in topic and word choice, likely influenced by socially constrained gender roles. Different priorities or interests could explain some topic differences; for example, the usage of family or appearance related words in feminine speech, and the usage of finance or sports related words in masculine speech. Some of the word choice differences might be explained by different face-saving strategies; for example the presence of “please” in feminine speech, a potential strategy to minimise threat to the partner's negative face, versus “pleasure” in masculine speech, a potential strategy to enhance the positive face of the participant (or of their partner) (Brown and Levinson, 1987; Farenkia, 2012). A difference in these two formulations is in the argument structure, whether arguments are realised or implied: “you” as subject and the request as object for “please”, compared to the request as subject and “me” as object (or possessive determiner) for “pleasure”. In some of word choice differences, we observe a tendency of men to use informal or taboo words. Thus, our findings appear to support literature describing a gendered difference in the use of overt and covert prestige forms (Trudgill, 1972), by showing how lexical differences might contribute to this pattern. We note, however, that there are also some preponderantly feminine informal words identified by our analysis. Differences in media consumption, genre preferences, or role models, might also lead to differences in vocabulary.

As noted earlier, many have studied gender associated vocabulary and other linguistic features and some have tested these in gender profiling (e.g. Verhoeven and Daelemans (2019)) – we note

that this potentially extends to gender-sensitive likeability prediction. We found it is possible to construct a model which identifies gender-preponderant linguistic traits that explain the difference in liking between genders. This finding offers another potential explanation for the "women-are-wonderful" effect. It does so by bridging the gaps between the existing knowledge that women tend to be better liked, women's speech tends to be more affiliative, intimate, and responsive, and traits signalling affiliation, intimacy, and responsiveness correlate with better social outcomes. This also enables us to qualify the implications and extent to which women's social outcomes tend to be more favourable. This effect appears contingent on the contextual desirability of the aforementioned traits. It is also likely limited to the extent to which the likeable traits come across to the receptor, which is limited when the receptor is a third party observer (Huang et al., 2017; Cameron and Shaw, 2016).

Our findings are also consistent with the view of Eckert (2000, 2012), who explains gendered language differences as not merely reflecting some natural categories, but rather as context-dependent variation which is used to construct, differentiate, and signal the speaker's identity. On the basis of gender-related probability patterns, we identify components of interest for linguistic styles which we label as feminine and masculine. However, we then show how these linguistic styles are each adopted by both genders. In line with Eckert (1989), this suggests that the usage of what we have labelled as a feminine vocabulary is neither inherently nor exclusively a feminine behaviour, but rather it is a behaviour common in women which likely reflects a social strategy and partly explains observed social outcome (likeability) differences. Interestingly, by adopting one linguistic style or the other, both genders achieve similar results in terms of social outcomes; our findings do not suggest that liking would be contingent on gender-conformity when it comes to linguistic style. This contrasts observations made by Eckert (2000) regarding gender and other social categories or or social outcomes and signals that our finding should be interpreted carefully: other social outcomes, such as perceptions, self-perceptions, and meta-perceptions of attraction, rapport, or specific traits might all have different relationships with gender and gender-conformity.

7. Conclusions

7.1 Main findings and impact

We set out to gain a better understanding of how gender correlates with liking perceptions and metaperceptions in dyadic interactions, and whether this correlation can be attributed to linguistic trait differences between genders.

We found that people have a tendency to consider themselves average in likeability regardless of their true position, and that this gives rise to a negative liking gap in a population where most individuals are likeable. We also found that women tend to receive and give higher liking ratings, with pairs of women giving and receiving the highest scores, on average. The present work identi-

fies significant gender-related vocabulary differences, both in how women and men tend to discuss different topics, but also how they might refer to the same things differently. Finally, we found some evidence to the hypothesis that gendered speech intersects with likeable speech, resulting in feminine vocabularies being more likeable regardless of speaker gender. This suggests that the associations we found appear to be universal, and do are not affected by gender-congruency.

Our work expands upon existing theories, connecting and showing the interplay of liking, gender, and language. The empirical findings are of particular interest for untangling how women and men socially interact and painting a clearer picture of gendered language and likeable language. They also provide a historical snapshot of the profile of women, men, and the likeable individual during the COVID pandemic, due to the circumstances of the corpus analyzed. The model design can be adapted into natural language processing models for a variety of tasks; the white-box, explainable design makes it particularly suitable for tasks such as policy analysis where transparency is paramount.

7.2 Limitations

We imagine these findings generalize beyond the sample provided by the CANDOR corpus. However, limitations apply. This analysis primarily considered the relationship of gender identity with linguistic behaviours (word/topic choice) and social interaction outcomes (liking). The mapping of gender identity on to a binary represents a simplification. An expanded definition of gender is necessary at both the data collection step and the model design step of research in order to form a more accurate picture of the interaction of gender identity with language and social perceptions. Relatedly, Jäggi et al. (2025) discuss biosocial gender as a multidimensional construct including biological properties, such as chromosomes or hormonal status, as well as socially constructed categories such as gender roles, gender identity and gender expression. When collecting data, these dimensions ought to be considered in relation to the questions one hopes to answer. The present work relied on limited information about gender identity. Other dimensions might be relevant: one's adopted gender roles might relate more closely to linguistic behaviour than gender identity alone would; one's gender expression might relate more closely to external perceptions than one's gender identity alone would; biological properties are less likely to explain our findings, but their interaction with gender identity in determining one's gender modality, such as cisgender or transgender, might offer valuable insight.

Languages other than English, or child and teenage populations were not represented in the corpus, while populations such as older adults were under-represented. These populations might have some topic/word choice similarities to the population sampled, but differences are also expected to exists. No indication of any participant's English status (native or otherwise) is given. Adherence to subcultures based on hobbies, identity, or beliefs might intersect with gender to produce additional specific effects. These would not be fully visible in this corpus unless the subcultures happened to

be large enough for adherents to be paired up by chance, or open enough for outsiders to be willing and able to carry a conversation on that topic (arguably, such was the case for sports or tech/gaming fans). Another point of specificity is the computer mediated conversation setting. While this format is ubiquitous today, many participants might not have been as used to it prior to the boom in online workspaces.

As stated previously (§6), (meta)perceptions of other social outcomes might not relate to gender, gendered language differences, and gender-congruency in the same way liking does. Additionally, (meta)perceptions of liking or other social outcomes might all form differently in other interaction contexts and settings, especially in interactions with a significant transactional or task-based component, or in interactions which include third party observers or reviewers. In a different setting, Cameron and Shaw (2016) show how press coverage of British politicians tends to centralise themes of gender (essentialist) differences. *Prima facie*, the attitudes expressed seem consistent with the "women-are-wonderful" bias, by framing feminine attributes as positive. In practice, however, this allows for women to be discredited for both not adhering to these positive, feminine expectations as well as not adhering to assertive, masculine expectations. In this context, we observe the positive association with femininity being incorporated into benevolent sexism – seemingly positive attitudes which ultimately reinforce sexist or misogynistic stereotypes and attitudes. There are, therefore, key differences a setting which is adversarial (political debates), as opposed to casual and in perceptions expressed by external observers (the press) rather than by conversation partners. The relationship of (meta)perceptions and language (gender-preponderant or otherwise) is also obfuscated by the fact that language itself, including its intent, meaning, and effect is shaped by its context (Heritage and Clayman, 2010). Further testing would be necessary to extend our conclusions to settings which differ significantly. Certainly, our method of analysis transfers directly to such settings: we identify words that have distinctive patterns of use between groups if distinctive patterns exist, and then we identify relevant broader categories for those words.[2]

While the relationships identified remain generalisable to social interactions, it is worth considering some of these results in the context of the COVID-19 pandemic, which at the time of the CANDOR conversations was having a significant social and financial impact on the participants. Some gender-preponderant words leave little doubt that the pandemic setting intersected with gender; words that emerged in our analysis as preponderantly feminine, such as 'mask' or 'quarantine', show a gender-preponderant effect of the pandemic on topic/word choice. Less clearly, it is possible that the social and financial strains might have been be reflected in the gender-preponderant topics, with women being more likely to discuss the social strain, the healthcare system, and the

2. This discovery process is the inverse of using a coding manual, such as employed by Banwart and McKinney (2005), but it remains open that while no significant differences emerged with respect to their codes – which it must be noted are described as encoding a much broader range of behaviours and features than word use, but is evidently also informed by word choice – significant differences of word use may be evident in the data they explored. We leave this question to the future, as well.

impact of school closures on education; while the men were more likely to discuss the financial strain, political environment, and effect of emerging technologies on their careers. These topics are largely timeless; however, studying different corpora might shed light onto whether the gender-divide in discussing these topics is similarly timeless, or whether adherence to gender-preponderant behaviours intensifies in times of crisis.

7.3 Future work and final remarks

In order to better understand the extent of the "women-are-wonderful" phenomenon, we have focused on gender as a determinate feature. However, other relevant features in conjunction gender might be identified. Our findings do not suggest a direct causal relationship whereby gender (or some dimension of it) would directly lead to specific language styles, and such an interpretation would also be inconsistent with sociolinguistic literature (Eckert, 2012). Instead, the potential explanations we discussed for the gendered language differences assume that gender interacts with social categories, (sub)cultures, and societal roles. Some of these interactions can be confirmed on infirmed by testing the correlation of gender with other categories in a representative sample. In the CANDOR corpus, data is available documenting the participants' age, race, level of education and employment, as well as psychological traits (e.g. neuroticism) and states at the time of the conversation (e.g. affect). Testing the correlation of gender with these features would be a next step in order to verify whether any of them mediate the relationship we identified between gender and gender-preponderant words and the (meta)perception thereof. Additionally, the interaction of gender with these other features needs to be studied in order to better quantify and qualify the relationship we identified between gender, gendered vocabularies, and liking. For example, further testing is necessary to establish whether these relationships are congruent across age brackets and other demographic and social categories. This applies to both the relationship between gender and potentially gender-preponderant words, and to the relationship between gender, vocabulary, and liking.

In identifying significant differences in language production as a function of gender and in relation to the liking gap, we have focussed on vocabulary. Additional dimensions of language production demand scrutiny. For example, phrasal hedges have been widely studied among gender-discriminating linguistic features with conflicting results reported (Holmes, 1991). Thus, beyond the lexicon, a larger inventory of phrase types, and also paralinguistic features (see Weiss and Burkhardt (2010)), require study. Furthermore, current understanding of the relationship hereby identified might be further improved by considering aspects of interactional sequences and context. The method we used to identify gender-preponderant words does not take into account synonymity, figurativeness, or context, all of which influence the semantic meaning and the interactional effect effect of language. In addition to this, the method we used to label a participant's vocabulary does not account for *when* in the interactional sequence a word was uttered, nor for whether a word that

was not uttered would have been relevant in the first place in the participant's conversation (see Schegloff (1993) for a discussion on how these nuances are often lost in quantitative studies). The computational methods might be refined by using word labels which capture these aspects of meaning, sequencing, and context and allow for improved counting that is more consistent with principles of conversation analysis, such as those described by Schegloff (1997).

Our results raise new questions. The finding that feminine vocabularies are associated with not only higher received liking ratings, but also higher given ratings is surprising. Potential explanations for feminine vocabularies correlating with a more positive overall social experience surround the potential knock-on effect of affiliative or intimate conversation which, when mirrored, makes one's partner more likeable in turn. Topics would have to be identified at the turn level in order to be able to test the hypothesis that mirroring underpins this effect, and also the hypothesis that mirroring or reciprocity is an inherently likeable trait.

Future research should refine the analysis of likeable gendered speech by considering more types of linguistic and paralinguistic behaviours. The question of universal desirability surrounding gender-preponderant linguistic behaviour remains open.

Acknowledgements

This publication records research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223.

Data Availability Statement

Instructions on accessing the data from the CANDOR corpus are provided by Reece et al. (2023). In accordance with the terms of use, we cannot publicly share the data. Our data analysis code is openly available at https://clg.scss.tcd.ie/Gender-Vocabulary-Liking/.

Appendices

A. Gender-preponderant word table

Table 11: Gender-preponderant words. Words in a **bold** font were identified by both counting methods. Regular font words were identified by the utterance counting method only

Women	Men
Health	Tech
therapy, therapist, deaf, counseling, clinic, immune, allergy, med, psychology, nurse, brain, health, sick, mask	**pc, programmer, microsoft, programming, software**, generate, hardware, java, vr, developer, install, python, coding, script, upgrade, server, equipment, electric, digital, update, network, data, code, machine, tech, laptop, website, user, computer, use
School	Finance
undergrad, school, dorm, grad, grade, campus, graduate, teacher, student, class	**buck, dollar**, grind, estate, union, turk, economy, stock, invest, task, ship, price, amazon, industry, cash, tax, company, sell, money, check, pay, business, deal, buy, real
Pets	Sports
pet, dog, cat, bark, puppy, animal	**baseball, sport, basketball, team, beat, win**, brady, boxing, kobe, nba, nfl, golf, league, hit, run
Holidays	Games
christmas, holiday, birthday	**console, player, game, play**, discord, ps, nintendo, gamer, gaming
Appearance	Politics
makeup, hair, wear	**power**, solar, senator, fuel, mike, indiana, war, shoot, trump, right
	Entertainment
	sequel, wayne, casino, release, stream, series, star, movie
Feelings	
cute, lovely, scared, scary, excited, nervous, glad, sad, love, feel, hope, crazy, anxiety, stressful, afraid, cry, exciting, alone, awful, hard, wish, happy, hate, funny, awesome, fun	**fan**, enjoy, sure, into
Family	
husband, boyfriend, pregnant, daughter, son, sibling, mom, baby, child, sister, family, dad, parent, young, sex, grandma, foster, adopt, relationship, married, partner, kid	**wife**, girlfriend, folk

continued on next page

continued from previous page

	Household
home, vegetable, pie, grocery, live, stay	**burger, beer, build**
	Quantities
many	**1000, million, mile, 2, 15, %, 100, 40, $, 30, 20, once**, mathematics, rating, row, 500, cent, average, per, compare, 3, 50, 25, amount, case, base, almost, couple, 10, most, few, four, three, bit, some, big, least, every, much, all, more, one, lot
	Location
philly, pennsylvania, to	**orlando**, serbia, lexington, tampa, singapore, taiwan, toronto, milwaukee, southwest, miami, florida, west, chicago, north, world, around, country, down, area, way, city, there, out, up, from, on
	Quality
old, nice, not	**aspect, decent**, various, quality, available, correct, large, alright, side, low, pretty, prolific, problem, easy, same, new
	People
woman, girl, friend	**buddy, man, dude**, bro, john, guy
	Time
then	**quick**, whenever, minute, ago, month, hour
	Social
tell	**mention, call**
	Articles & Prepositions
	by, the, of, a, for
	Phatics
goodness, gosh, hi, god, bye, thank, yes, please, oh, hmm, wow, sorry, mm, um, mhm	**pleasure, hey**, damn, uh, hell, shit, yeah
	Verbs
quit, need, want, do, have, would, be	**create, seem, make, get**
	Pronouns
she, he, i, we	**these, something, those, you, this, that, it**
	Fillers & Hedges
like	**relatively, basically**, etcetera, whatnot, generally, essentially, kinda, chance, assume, sort, general, mean, probably, point, guess, might, whatever, kind, know, thing

continued on next page

continued from previous page

Links	
because, so, also, and, but	**example**, as, or, than, if
Intensifiers	
such, totally, super, really, very	**gotta**, certainly, actually
Questions	
why, who	**search, question**

References

Eric Acton. On gender differences in the distribution of um and uh. *Penn Working Papers in Linguistics*, 17(2):Article 2, 2011.

American Psychological Association. Gender identity. In *APA dictionary of psychology*. Retrieved March 03, 2025, from https://dictionary.apa.org/gender-identity, n.d.a.

American Psychological Association. Nonbinary. In *APA dictionary of psychology*. Retrieved April 21, 2026, from https://dictionary.apa.org/nonbinary, n.d.b.

S. Argamon, Moshe Koppel, J. Fine, and Anat Rachel Shimoni. Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3):321–346, 2003. doi: https://doi.org/10.1515/text.2003.014.

Elizabeth J. Aries and Fern L. Johnson. Close friendship in adulthood: Conversational content between same-sex friends. *Sex Roles*, 9(12):1183–1196, 1983. doi: https://doi.org/10.1007/BF00303101.

April H. Bailey, Robert Lambert, and Marianne LaFrance. Implicit reactions to women in high power body postures: Less wonderful but still weaker. *Journal of Nonverbal Behavior*, 44(2):329–350, 2020. doi: https://doi.org/10.1007/s10919-019-00327-w.

Mary Christine Banwart and Mitchell S. McKinney. A gendered influence in campaign debates? analysis of mixed-gender uni ted states senate and gubernatorial debates. *Communication Studies*, 56(4):353–373, 2005. doi: 10.1080/10510970500319443.

Diane S. Berry, James W. Pennebaker, Jennifer S. Mueller, and Wendy S. Hiller. Linguistic bases of social perception. *Personality and Social Psychology Bulletin*, 23(5):526–537, 1997.

Akos Bocskor. Gender differences in coolness dynamics among early and middle adolescents. *Journal of Early Adolescence*, 45(7):801–835, 2024. doi: https://doi.org/10.1177/02724316241281981.

Erica J. Boothby, Gus Cooney, Gillian M. Sandstrom, and Margaret S. Clark. The liking gap in conversations: Do people like us more than we think? *Psychological Science*, 29(11):1742–1756, 2018. doi: https://doi.org/10.1177/0956797618783714.

Alison Wood Brooks and Michael Yeomans. Boomerasking: Answering your own questions. *Journal of Experimental Psychology: General*, 154(3):864–893, 2025. doi: https://doi.org/10.1037/xge0001693.

Gillian Brown and George Yule. *Discourse Analysis*. Cambridge University Press, 1983.

Penelope Brown. How and why are women more polite: Some evidence from a Mayan community. In S. McConnell-Ginet, R. Borker, and Furman F., editors, *Women and language in literature and society*. Preager Publishers, 1980.

Penelope Brown and Stephen Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. Discriminating gender on twitter. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011.

Deborah Cameron and Sylvia Shaw. *Gender, Power and Political Speech*. Palgrave Macmillan London, 1 edition, 2016. doi: https://doi.org/10.1057/978-1-137-58752-7.

G. J. Chelune. Reactions to male and female disclosure at two levels. *Journal of Personality and Social Psychology*, 34(5):1000–1003, 1976. doi: https://doi.org/10.1037/0022-3514.34.5.1000.

Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. What makes a good conversation? challenges in designing truly conversational agent. In Stephen Brewster and Geraldine Fitzpatrick, editors, *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page Article 475. Association for Computing Machinery, New York, NY, USA, 2019. doi: https://doi.org/10.1145/3290605.3300705.

Alysen Demzik, Pauline Filippou, Christopher Chew, Allison Deal, Emily Mercer, Sejal Mahajan, Eric M. Wallen, Hung-Jui Tan, and Angela B Smith. Gender-based differences in urology residency applicant personal statements. *Urology*, 150:2–8, 2021. doi: https://doi.org/10.1016/j.urology.2020.08.066.

Robin Dunbar. *Grooming, Gossip, and the Evolution of Language*. Harvard University Press, 1998.

Alice H. Eagly and Antonio Mladinic. Are people prejudiced against women? some answers from research on attitudes, gender stereotypes, and judgments of competence. *European Review of Social Psychology*, 5(1):1–35, 1994. doi: https://doi.org/10.1080/14792779543000002.

Penelope Eckert. The whole woman: Sex and gender differences in variation. *Language Variation and Change*, 1(3):245–267, 1989. doi: https://doi.org/10.1017/S095439450000017X.

Penelope Eckert. *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Blackwell, 2000.

Penelope Eckert. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41(1):87–100, 2012. doi: https://doi.org/10.1146/annurev-anthro-092611-145828.

European Institute for Gender Equality. Gender identity. In *Glossary and thesaurus*. Retrieved March 03, 2025, from https://eige.europa.eu/publications-resources/thesaurus/terms/1049, n.d.

Bernard Mulo Farenkia. Face-saving strategies in responding to gratitude expressions: Evidence from canadian english. *International Journal of English Linguistics*, 2(4):1–11, 2012. doi: https://doi.org/10.5539/ijel.v2n4p1.

Richard J Harnish, Antonia Abbey, and Kenneth G DeBono. Toward an understanding of 'the sex game': The effects of gender and self-monitoring on perceptions of sexuality and likability in initial interactions. *Journal of Applied Social Psychology*, 20(16, Pt 2):1333–1344, 1990. doi: https://doi.org/10.1111/j.1559-1816.1990.tb01475.x.

John Heritage and Steven Clayman. *Talk in Action: Interactions, Identities, and Institutions.* Wiley, 2010.

Janet Holmes. Language and gender. *Language teaching*, 24(4):207–220, 1991.

Karen Huang, Michael Yeomans, Alisson Wood Brooks, Julia Minson, and Francesca Gino. It doesn't hurt to ask: Question-asking increases liking. *Journal of Personality and Social Psychology*, 113(3):430–452, 2017. doi: https://doi.org/10.1037/pspi0000097.

T. Jäggi, P. M. Gygax, S. Decock, U. Gabriel, S. V. Hoof, H. Verhaegen, and C. Vincent. Beyond she and he: A framework for studying the cognitive, psychological and social effects of gender-neutral pronouns. *Journal of language and social psychology*, 44(6):850–880, 2025. doi: https://doi.org/10.1177/0261927X251346193.

Michael R King. Can Bard, Google's experimental chatbot based on the LaMDA large language model, help to analyze the gender and racial diversity of authors in your cited scien-

tific references? *Cellular and molecular bioengineering*, 16(2):175–179, 2023. doi: https://doi.org/10.1007/s12195-023-00761-3.

Margaret Kleiser Polk and Lara Mayeux. Associations between peer-perceived and self-perceived gender typicality and peer status in early adolescence. *Journal of Early Adolescence*, 43(3): 320–341, 2023. doi: https://doi.org/10.1177/02724316221105604.

Vijay Kotu and Bala Deshpande. *Predictive Analytics and Data Mining*. Elsevier, 2015. doi: https://doi.org/10.1016/C2014-0-00329-2.

Justin Kruger and David. Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, 1999. doi: https://doi.org/10.1037/0022-3514.77.6.1121.

Campbell Leaper. Agency, communion, and gender as predictors of communication style and being liked in adult male-female dyads. *Sex Roles: A Journal of Research*, 16(3-4):137–149, 1987. doi: https://doi.org/10.1007/BF00289645.

Campbell Leaper and Melanie Ayres. A meta-analytic review of gender variations in adults' language use. *Personality and Social Psychology Review*, 11(4):328–363, 2007. doi: https://doi.org/10.1177/1088868307302221.

Miriam A. Locher. Politeness theory and relational work. *Journal of Politeness Research*, 1:9–33, 2005.

Adam M. Mastroianni, Gus Cooney, Erica J. Boothby, and Andrew G. Reece. The liking gap in groups and teams. *Organizational Behavior and Human Decision Processes*, 162:109–122, 2021. doi: https://doi.org/10.1016/j.obhdp.2020.10.013.

Matthias R. Mehl, Simine Vazire, Nairán Ramírez-Esparza, Richard B. Slatcher, and James W. Pennebaker. Are women really more talkative than men? *Science*, 317(5834):82–, 2007. doi: https://doi.org/10.1126/science.1139940.

Zachary Miller, Brian Dickinson, and Wei Hu. Gender prediction on twitter using stream algorithms with n-gram character features. *International Journal of Intelligence Science*, 2(4A):143–148, 2012. doi: https://doi.org/10.4236/ijis.2012.224019.

Anais Murat, Maria Koutsombogera, and Carl Vogel. Mutual gaze and linguistic repetition in a multimodal corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 13th Conference on Language Resources and Evaluation*, pages 2771–2780. European Language Resources Association, Marseille, France, 2022.

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3): 211–236, 2008. doi: https://doi.org/10.1080/01638530802073712.

Michael Noseworthy, Jackie Chi Kit Cheung, and Joelle Pineau. Predicting success in goal-driven human-human dialogues. In Kristiina Jokinen, Manfred Stede, David DeVault, and Annie Louis, editors, *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 253–262, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5531. URL https://aclanthology.org/W17-5531/.

Vanessa Y. Oviedo, Andrew J. Guydish, and Jean E. Fox Tree. The liking gap online: People like us more than we think. *Computers in Human Behavior Reports*, 17:Article 100582, 2025. doi: https://doi.org/10.1016/j.chbr.2024.100582.

Laurent Pointal. TreeTaggerWrapper, 2015. URL pypi.org/project/treetaggerwrapper/2.3/.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):Article eadf3197, 2023. doi: https://doi.org/10.1126/sciadv.adf3197.

David Reitter and Johanna D Moore. Predicting success in dialogue. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815. Association for Computational Linguistics, Prague, Czech Republic, 2007. doi: https://doi.org/10.1017/S0140525X04000056.

David Reitter and Johanna D Moore. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46, 2014.

Anne Riissanen and Greg Watson. Gender difference in the use of thought representation – A corpus-based study. *English Language Teaching*, 7(3):71–84, 2014. doi: https://doi.org/10.5539/elt.v7n3p71.

Krystyna Rojahn and Tineke M. Willemsen. The evaluation of effectiveness and likability of gender-role congruent and gender-role incongruent leaders. *Sex Roles*, 30(1-2):109–119, 1994. doi: https://doi.org/10.1007/BF01420743.

Laurie Rudman and Stephanie Goodwin. Gender differences in automatic in-group bias: Why do women like women more than men like men. *Journal of Personality and Social Psychology*, 87 (4):494–509, 2004. doi: https://doi.org/10.1037/0022-3514.87.4.494.

Emanuel A. Schegloff. Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26(1):99–128, 1993. doi: https://doi.org/10.1207/s15327973rlsi2601_5.

Emanuel A. Schegloff. Whose text? whose context? *Discourse & Society*, 8(2):165–187, 1997. doi: https://doi.org/10.1177/0957926597008002002.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In D. B. Jones and H. Somers, editors, *New Methods In Language Processing*, pages 154–164. Routledge, 1997. doi: https://doi.org/10.4324/9781315072685.

Helmut Schmid. Improvements in part-of-speech tagging with an application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural language processing using very large corpora*, pages 13–25. Springer Netherlands, 1999. doi: https://doi.org/10.1007/978-94-017-2390-9. Revised version of a paper originally presented at the 1995 ACL SIGDAT-Workshop in Dublin.

Elena Stephan and Yaacov Liberman, Nira andTrope. Politeness and psychological distance: a construal level perspective. *Journal of personality and social psychology*, 98(2):268–280, 2010. URL `https://doi.org/10.1037/a0016960`.

Deborah Tannen. Gender differences in topical coherence: Creating involvement in best friends' talk. *Discourse Processes*, 13(1):73–90, 1990. doi: https://doi.org/10.1080/01638539009544747.

R.H. Tatchell, S. van den Berg, and J.W. Lerman. Fluency and eye contact as factors influencing observers' perceptions of stutterers. *Journal of Fluency Disorders*, 8(3):221–231, 1983. ISSN 0094-730X. doi: https://doi.org/10.1016/0094-730X(83)90036-0.

Colin A Tidwell, Alexander F Danvers, Valeria A Pfeifer, Danielle B Abel, Eva Alisic, Andrew Beer, Sabrina J Bierstetel, Kathryn L Bollich-Ziegler, Michelle Bruni, William R Calabrese, Christine Chiarello, Burcu Demiray, Sona Dimidjian, Karen L Fingerman, Maximilian Haas, Deanna M Kaplan, Yijung K Kim, Goran Knezevic, Ljiljana B Lazarevic, Minxia Luo, Alessandra Macbeth, Joseph H Manson, Jennifer S Mascaro, Christina Metcalf, Kyle S Minor, Suzanne Moseley, Angelina J Polsinelli, Charles L Raison, James K Rilling, Megan L Robbins, David Sbarra, Richard B Slatcher, Jessie Sun, Mira Vasileva, Simine Vazire, and Matthias R Mehl. Are women really (not) more talkative than men? a registered report of binary gender similarities/differences in daily word use. *Journal of Personality and Social Psychology*, 128(2):367–391, 2025. doi: https://doi.org/10.1037/pspp0000534.

Pavel Trofimovich, Rachael Lindberg, Anamaria Bodea, Thao-Nguyen Nina Le, Chaoqun Zheng, and Kim McDonough. I don't think you like me: Examining metaperceptions of interpersonal

liking in second language academic interaction. *Languages*, 8(3):200–222, 2023. doi: https://doi.org/10.3390/languages8030200.

Peter Trudgill. Sex, covert prestige and linguistic change in the urban british english of norwich. *Language in Society*, 1(2):179–195, 1972. doi: https://doi.org/10.1017/S0047404500000488.

Pradeep Vashisth and Kevin Meehan. Gender classification using twitter text data. In *31st Irish Signals and Systems Conference (ISSC)*. Institute of Electrical and Electronic Engineers, Letterkenny, Ireland, 2020. doi: https://doi.org/10.1109/ISSC49989.2020.

Ben Verhoeven and Walter Daelemans. Discourse lexicon induction for multiple languages and its use for gender profiling. *Digital Scholarship in the Humanities*, 34(1):208–220, 2019.

Carl Vogel and Liliana Mamani Sanchez. Epistemic signals and emoticons affect kudos. In Péter Baranyi, editor, *3rd IEEE Conference on Cognitive Infocommunications*, pages 517–522, 2012.

Carl Vogel, Maria Koutsombogera, and Justine Reverdy. Aspects of dynamics in dialogue collaboration. *Electronics*, 12(10):Article 2210, 2023. doi: https://doi.org/10.3390/electronics12102210.

Shaoda Wang, Dezhu Ye, and Junyun Liao. Politeness matters: The role of polite languages in online peer-to-peer lending. *Journal of Business Research*, 171:114394, 2024. ISSN 0148-2963. doi: https://doi.org/10.1016/j.jbusres.2023.114394. URL `https://www.sciencedirect.com/science/article/pii/S0148296323007531`.

Richard J. Watts. *Politeness*. Cambridge University Press, 2003.

Benjamin Weiss and Felix Burkhardt. Voice attributes affecting likability perception. In *Interspeech*, pages 2014–2017, 2010.

B. G. Wildman and B. Clementz. Assertive, empathic assertive, and conversational behavior: Perception of likability, effectiveness, and sex role. *Behavior Modification*, 10(3):315–331, 1986. doi: https://doi.org/10.1177/01454455860103004.

Wouter Wolf, Amanda Nafe, and Michael Tomasello. The development of the liking gap: Children older than 5 years think that partners evaluate them less positively than they evaluate their partners. *Psychological Science*, 32(5):789–798, 2021. doi: https://doi.org/10.1177/0956797620980754.