The Polarity Effect of Evaluative Language

Lucien Baumgartner (lucien.baumgartner@philos.uzh.ch)
University of Zurich, Department of Philosophy, Zürichbergstrasse 43, 8044 Zurich, Switzerland

Pascale Willemsen (pascale.willemsen@uzh.ch)
University of Zurich, Department of Philosophy, Zollikerstrasse 117, 8008 Zurich, Switzerland

Kevin Reuter (kevin.reuter@uzh.ch)
University of Zurich, Department of Philosophy, Zürichbergstrasse 43, 8044 Zurich, Switzerland

Abstract

Recent research on thick terms like 'rude' and 'friendly' has revealed a polarity effect, according to which the evaluative content of positive thick terms like 'friendly' and 'courageous' can be more easily cancelled than the evaluative content of negative terms like 'rude' and 'selfish'. In this paper, we study the polarity effect in greater detail. We first demonstrate that the polarity effect is insensitive to manipulations of embeddings (Study 1). Second, we show that the effect occurs not only for thick terms but also for thin terms such as 'good' or 'bad' (Study 2). We conclude that the polarity effect is indicative of a pervasive asymmetry that holds between positive and negative evaluative terms.

Keywords: polarity effect; thick terms; thin terms; evaluative language; moral judgments; praise; blame

1 Introduction

The terms we use to make evaluative judgments fall into at least two main classes (e.g., Eklund, 2011; Väyrynen, 2013). First, thin terms like 'great' and 'awful' evaluate, i.e., praise or blame, a person or state of affairs without providing any descriptive information as to what it is that the person or state of affairs is considered praise- or blameworthy for. Second, thick terms like 'generous' and 'honest' also evaluate, but additionally communicate the descriptive features in virtue of which someone or something is evaluated. For instance, by saying that Sally is generous and by calling her honest, we evaluate her behaviour positively. However, being generous is clearly different from being honest—generosity is concerned with sharing things with others, honesty is about telling the truth. While 'generous' and 'honest' share the same evaluative component, they differ in the descriptive features that are the basis for the positive evaluation.[1]

[1] More recently, researchers have identified another class of evaluative concepts, so-called dual character concepts (Knobe et al., 2013; Del Pinal & Reuter, 2017; Reuter, 2019; Reuter, Löschke, and Betzler 2020). Given that dual character concepts have two independent dimensions for categorization, we will not empirically investigate this class of concepts in this paper. Also, some philosophers suggest that pejoratives and slurs constitute independent classes of evaluative concepts (for a discussion, see Cepollaro, 2020). However, both pejoratives and slurs only communicate negative evaluations and do not have positive counterparts. The aim of this paper is to investigate whether positive and negative evaluations of terms within the same class behave differently. Therefore, we omit pejoratives and slurs.

Thick concepts have received a lot of attention in the literature because they hold together descriptive and evaluative content, with one of the main questions being whether we can detach the evaluation and use thick concepts non-evaluatively (e.g., Kirchin, 2010; Putnam, 2002; Roberts, 2011; Willemsen et al., 2022, ms; Williams, 1985). Some philosophers deny this and argue that statements like 'What Tom did was cruel, but by that I am not saying something negative about him' are contradictory and, thus, infelicitous (Elstein & Hurka, 2009; Hare, 1952; Kyle, 2019). Others believe that a non-evaluative use of thick concepts is possible and that statements like this are felicitous (Blackburn, 1992; Cepollaro, 2020; Cepollaro & Stojanovic, 2016; Hare, 1963; Väyrynen, 2021).

Willemsen and Reuter (2020, 2021) tested these two opposing intuitions by using the cancellability test for conversational implicatures (see Grice, 1989; Sullivan, 2017; Zakkou, 2018). Here are some examples of the experimental stimuli that were used, distinguishing between attributions of thick terms to people (Character) and attributions of thick terms to behavior (Behavior):

(1) Negative Character: Amy is rude, but by that I am not saying something negative about Amy.

(2) Negative Behavior: Amy's behavior last week was rude, but by that I am not saying something negative about Amy's behavior that day.

(3) Positive Character: Tom is friendly, but by that I am not saying something positive about Tom.

(4) Positive Behavior: Tom's behavior last week was friendly, but by that I am not saying something positive about Tom's behavior that day.

Participants were then asked whether the speaker, Sally, contradicts herself. At first sight, the results seem inconclusive. However, the most crucial finding goes beyond the initial research question and reveals a systematic difference between positive and negative terms. Negative evaluations were significantly harder to cancel compared

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

to positive ones ($\Delta \approx 1.0$ on a 9-point Likert scale), irrespective of whether the thick terms were assigned to the character or the behavior. More specifically, statements like (1) and (2) were judged to be significantly more contradictory than statements like (3) and (4). This asymmetry, called *Polarity Effect*, was previously unknown and provides a challenge for the idea that positive and negative thick terms can be treated alike (see also Väyrynen (2021) and Zakkou (2021)).

So far, the polarity effect has only been recorded for thick terms. One might wonder, though, whether the effect is in fact a more global effect that also holds for other evaluative terms, specifically thin terms like 'good' and 'bad'. It seems plausible to assume that the effect only occurs for thick concepts but disappears for thin ones. Thin concepts are said to be merely evaluative, with their only function being to express approval or disapproval. What does remain if we cancel this sole content of a thin concept? The term should be empty and no longer express anything. Thus, whatever the reason is that the polarity effect occurs for thick concepts, it should not pertain to thin concepts as well. Another reason to think that the polarity effect occurs for thick concepts only, is the descriptive richness of thick concepts. Philosophers have argued that the descriptive content is often very rich and contains disjunctive features (Wiggins, 1993), which may lead to unexpected effects in experimental settings like the cancellability task. To give an example: One person can be called courageous for trying a dangerous trick on a snowboard, while another person demonstrates courage by standing up to the class bully, or simply by being themselves and not caring about other people's opinion. Courage comes in many forms that often cannot be properly reduced to one shared core feature. If this picture is correct, then the evaluation of a thick concept is less central to the semantic content—it is simply one of many things that make up the concept. For thin concepts, however, the evaluation is extremely central and, in fact, all that the concept expresses.

This line of reasoning can still not explain why positive and negative terms behave differently when the evaluation is cancelled, but it provides a suggestion of where to search for the root of the effect. If the polarity effect were a phenomenon restricted to thick terms only, then a promising explanation of the effect, let's call it **thick concept explanation**, would dig into the intricacies of thick terms. First, the way in which the evaluative content combines with the descriptive content might be different for positive and negative terms. Second, as suggested above, the descriptive content of negative thick terms might be descriptively richer or might contain more disjunctive features compared to the descriptive content of positive thick terms.

In contrast, if the effect were to also hold for thin terms, then an explanation that focuses on the descriptive aspects of thick concepts would not take us very far. Thus, in case the polarity effect is a more pervasive evaluative language effect, then the following claim should hold:

Pervasive Linguistic Asymmetry: A negative evaluation is, ceteris paribus, harder to explicitly cancel compared to a positive evaluation.

Consequently, a more encompassing explanation would be required. Willemsen and Reuter (2021) suggest an explanation of the polarity effect that is grounded in different social norms, let's call it **social norms explanation**, that may guide our behavior. They state:

> "Uttering a positive thick term without the intention to commit to a positive evaluation seems relatively harmless. Being misunderstood in cases of negative thick terms has a potentially greater impact. If mistaken, a speaker communicates a negative evaluation they initially did not want to commit to. Since negative evaluations harm others by diminishing their social status and reputation, people are less willing to accept a cancellation of a negative evaluation." (p. 8)

Such an explanation would be consistent with a growing body of empirical evidence that has shown moral valence to have an effect on non-moral judgements, e.g. of knowledge (Beebe & Buckwalter, 2010) and causation (Sytsma et al., 2019, for an overview see Willemsen & Kirfel, 2019). Additionally, the philosophical and linguistic literature is rife with results in which social norms seem to have an asymmetrical influence on praise and blame (Guglielmo & Malle, 2019). Recently, Anderson, Crockett, and Pizarro (2020) argued that while both praise and blame are essential to sustaining social relationships and facilitating social regulation, blaming one another comes with significant social costs, both on the part of the blaming and the blamed party. Being blamed can have serious consequences, such as loss of reputation and social alliances, social exclusion, or punishment. Consequently, the *wrongful* attribution of blame that is unjustifiably causing a person to suffer these negative consequences, is itself an act of severe social impact.[2]

So far we lack evidence of the effect's robustness across embeddings and whether or not it is a thick concept or an evaluative language effect. In this paper, we demonstrate that the polarity effect is not only robust but extends to thin ethical concepts as well, allowing for the conclusion that the polarity effect is indicative of a pervasive linguistic asymmetry. In the empirical part of the paper, we do two things: First, in Study 1, we provide

[2] See also Willemsen, Baumgartner, Cepollaro, & Reuter, ms for a discussion.

a clearer understanding of the polarity effect by investigating how far-reaching it is, viz. in what embeddings it occurs (Section 2.1). In Section 2.2, we provide empirical evidence (Study 2) that the polarity effect holds more globally for both thick as well as thin terms.

2 The Extent and Character of the Polarity Effect

2.1 Study 1: Investigating the Polarity Effect in Different Embeddings

In this study, we investigate the scope of the polarity effect. It might be argued that the previously recorded effect only holds when a thick term is attributed to an individual person ("Amy is rude.")—hereafter, Individual Statement condition—but not in other embeddings, e.g., generic generalizations ("People are rude."). If that were the case, then the polarity effect would have a more narrow application and would be moderated by the subject term.

Two main hypotheses guided the design of our study.[3] First, we predicted to replicate the polarity effect recorded in previous studies:

Polarity Hypothesis (H1): Contradiction ratings in the Individual Statement condition are significantly higher for negative thick terms compared to positive thick terms.

Second, we expected an inverse relationship between the scope of predication and the assertive commitment: the more general an evaluative statement, the smaller the commitment to the evaluation. Generic statements ("people are rude") are notoriously easy to take back, due to their inherent scope ambiguity (e.g., Sterken, 2017; Thakral, 2018). Similarly, limited scope statements (e.g., "some people are rude") do not commit the speaker to the evaluation on a personal level. Individual statements (e.g., "Amy is rude"), in contrast, have higher immediate social costs and thus are most likely to follow social norms. Hence, we hypothesized an embedding effect:

Embedding Hypothesis (H2): The polarity effect is significantly reduced in limited scope statements and for generic generalizations.

Methods 932 participants were recruited via Prolific and completed an online survey implemented in Qualtrics. All participants were required to be at least 18 years old, English native speakers (or bilingual), and to have an approval rate of at least 95%. The remaining 872 participants had an average age of 38.47 years, and the gender distribution in the sample was 55.96% male, 43.81% female, 0.23% non-binary. The 6 positive and 6 negative thick terms we tested were:[4]

- Positive: compassionate, courageous, friendly, generous, honest, virtuous
- Negative: cowardly, cruel, manipulative, rude, selfish, vicious

Here are three exemplary statements we used (including the question that was asked subsequently), illustrating each variant with a different thick term:

Please imagine that [Sally/Tom] said the following sentence:

Individual statement
"[Amy/Steve] is rude, but by that I am not saying something negative about [her/him]."
Limited scope statement
"Some people are friendly, but by that I am not saying something positive about them."
Generic statement
"People are selfish, but by that I am not saying something negative about them."

Does [speaker] contradict [herself/himself]?

Contradiction ratings were recorded on a 9-point Likert scale ranging from 1 = "definitely not" to 9 = "definitely yes". Before participants gave their responses to the test sentences, they were given instructions on how to understand what a contradiction is (see preregistration material). The stimuli included proper names, both for the speaker (Sally/Tom) and the target of the predication in the individual person statement (Amy/Steve), which is a possible source of unexpected gender effects. Hence, the gender of the speaker was randomized evenly in order to control for possible gender effects.[5] Each participant was randomly assigned to one of the 72 stimuli (3 (embeddings) × 6 (concepts) × 2 (polarity) × 2 (gender of the speaker).

Results For the individual statements, the observed mean of positive thick concepts (6.39) was indeed lower compared to negative thick concepts (6.97). As the contradiction ratings significantly deviate from a normal distribution, we used non-parametric alternatives to test our hypotheses. According to a one-sided unpaired two-samples Wilcoxon test ($W = 9348.5$, $p = 0.013$), positive thick concepts have significantly lower average contradiction ratings than negative thick concepts (on 0.05-alpha level). Thus, cancelling negative thick concepts was assessed to be more contradictory than cancelling positive thick concepts. Hence, we cannot reject H1.

[3] The experimental design, predictions, and statistical models were pre-registered with the Open Science Framework. The data file with all the responses can be downloaded here.

[4] We selected the same 12 thick terms that were used in Willemsen & Reuter (2021). Among other reasons for their selection (see https://osf.io/xew6d), these adjectives have the feature of being frequently used in ordinary language.

[5] In the individual statements, Sally only talks about Amy and Tom only about Steve (i.e., gender is held constant across speaker and subject term).

Our second hypothesis was that the difference between negative thick terms and positive thick terms will be largest for individual statements. However, the differences in the estimated marginal means do not support this hypothesis, as shown in Table 1.[6] In fact, the difference for individual statements is the smallest (-0.60). All differences are significant on 0.05-alpha level. Hence, our hypothesis has to be rejected. Lastly, none of the control variables (gender of the speaker, age and gender of the participant) had any significant effect.

Table 1: Pairwise contrasts (positive - negative) of estimated marginal means by embedding. For individual statements, the difference in average contradiction ratings was 0.60.

Embedding	ΔEstimate	SE	t-ratio	p-value
Individual	-0.60	0.30	-2.00	0.047
Limited	-0.65	0.30	-2.14	0.033
Generic	-0.78	0.31	-2.56	0.011

Discussion In Study 1, we replicated the polarity effect for statements in which a thick term is attributed to an individual. Furthermore, the scope of this effect is not limited to statements of the form "[Subject] is [thick term]". Significant differences were found across all three embeddings, providing support for the claim that the polarity effect is rather pervasive. This suggests that the effect does not depend on the linguistic construction used.

2.2 Study 2: Extending the Polarity Effect to Thin Concepts

In previous studies as well as in Study 1 above, it was found that the polarity of a thick term has an effect on contradiction ratings using the cancellability paradigm. In this experiment, we investigated whether the polarity effect shows up for both thin and thick concepts, which would indicate that the effect is more widespread and holds for evaluative concepts more generally, rather than for thick concepts only. We therefore examined whether negative and positive thin terms behave differently from thick terms with respect to cancelling their evaluative content. We thus formulated the following hypotheses:[7]

Main Effect Hypothesis (H3): There is a significant effect of Polarity (Positive vs. Negative) on contradiction ratings, such that the ratings are higher for negative terms compared to positive terms.

Interaction Hypothesis (H4): There is no significant two-way interaction of Concept class (Thin vs. Thick) and Polarity (Positive vs. Negative).

Thin Concept Hypothesis (H5): Contradiction ratings are significantly higher for negative *thin* terms compared to positive *thin* terms.

Thick Concept Hypothesis (H6): Contradiction ratings are significantly higher for negative *thick* terms compared to positive *thick* terms.

Methods 325 participants were recruited via Prolific and completed our online survey implemented in Qualtrics. The same inclusion criteria and instructions were used as in Study 1. The final sample included 303 participants (34.65% male, 63.37% female, 1.98% non-binary) with an average age of 36.69 years.

As stimuli, we used three positive and three negative thin and thick concepts each:

- Thin concepts:
  - Positive: good, great, ideal
  - Negative: bad, awful, terrible[8]
- Thick concepts:
  - Positive: friendly, honest, compassionate
  - Negative: rude, manipulative, cruel

After two test questions, participants were presented with the following vignette[9]:

Please imagine that Sally said the following sentence:

> "What [person] did last week was [thin/thick term], but by that I am not saying something [positive/negative] about [her/his] behavior that day."

Does Sally contradict herself?

The participants answered on a 9-point Likert scale anchored at 1 = "definitely not" and 9 = "definitely yes". Since the gender of the speaker did not have any significant effect in Study 1, we did not add it as a control variable in Study 2. Instead, we varied the gender of the person Sally is speaking about, but without duplicating the number of vignettes. Accordingly, participants were evenly assigned to one of the 12 vignettes (3 (terms) × 2 (concept classes) × 2 (polarity)).

[6] The estimation is based on a two-way ANOVA of the interaction of polarity and embedding, with the gender of the speaker (male/female), as well as age (continuous) and gender of the respondent (male/female/non-binary) as controls.

[7] The experimental design, predictions, and statistical models were https://osf.io/r9mb5 pre-registered with the Open Science Framework. The data file with all the responses can be downloaded here.

[8] We selected highly frequent thin terms, including 'good', 'great', and 'bad' (2nd, 4th, and 22nd most frequently used adjectives in American English in the Corpus of Contemporary American English).

[9] Whereas in Study 1 we used thick term attributions to persons, in Study 2 thin and thick terms were attributed to behavior. Previous studies have revealed no differences between both conditions.

Results In Study 2, we found the main Polarity Effect again: according to a one-sided unpaired two-samples Wilcoxon test (W = 15712, p-value < 0.001), negative terms have significantly higher contradiction ratings than positive terms (across concept classes), thus supporting H3. Furthermore, the differences of differences based on Aligned Rank Transform (ART) nonparametric ANOVA (t-ratio (299) = 1.284, p-value = 0.2002) showed that there is no significant two-way interaction of concept class and polarity, which is in line with our predictions in H4. The Polarity Effect was also found for thin concepts (H5) and thick concepts (H6) respectively: a one-sided unpaired two-samples Wilcoxon test (W = 4021.5, p-value < 0.001) showed that negative thin concepts have significantly higher contradiction ratings than positive thin concepts; the same was found for thick concepts (W = 3918.5, p-value < 0.001). In summary, none of our hypotheses can be rejected.

In general, thick terms (5.93) have lower average contradiction ratings than thin concepts (7.15). Figure 1 depicts the means and standard error per term, which reveals two outliers, namely the thick terms *manipulative* (5.42) and *honest* (3.08). We thus ran additional tests to check for concept class differences for positive and negative terms respectively, with and without outliers. A two-sided unpaired two-samples Wilcoxon test (W = 2243.5, p-value = 0.01602) showed that there are significant differences between positive thin and positive thick concepts; the same is true for negative thin and thick concepts (W = 2032.5, p-value < 0.001). These differences are no longer significant for positive thin and thick, if we drop the the outlier *honest* (W = 1770, p-value = 0.5546), nor for negative thin and thick concepts after dropping *manipulative* (W = 1667.5, p-value = 0.1868).

Discussion The results of Study 2 paint a clear picture, according to which the polarity effect does not hold for thick terms only, but is a more widespread effect that applies to evaluative concepts more generally. Our results suggest that the Polarity Effect between positive and negative terms is a unified phenomenon for thin and thick concepts.

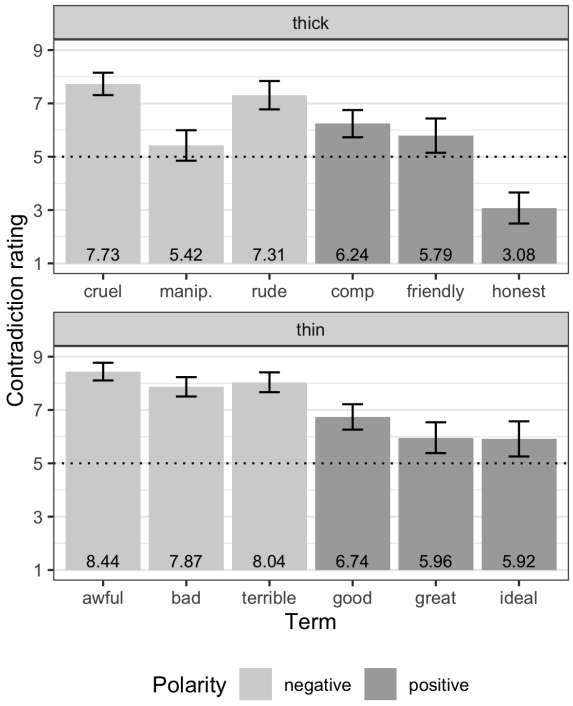

Figure 1: Average contradiction ratings per concept. The error bars display the standard error around the means.

3 General Discussion

3.1 Summary of the Results

The purpose of the empirical part of the paper was twofold. First, we aimed to replicate the polarity effect, thereby testing the extent to which the effect holds in different embeddings. Second, we aimed to investigate whether the polarity effect is a narrow *thick concept* effect, or whether it holds more widely for a larger set of evaluative terms including thin terms. In regards to the first aim, we successfully replicated the polarity effect for individual subjects. Furthermore, and against our predictions, the effect popped up in all three embeddings we tested, i.e., not only when thick terms are ascribed to persons, but also when being attributed to a group of people, as well as in generic statements. From this, we can conclude that the polarity effect is not (at least not strongly) dependent on the scope of predication in which the thick term appears.

In order to pursue our second aim, we tested not only a batch of thick terms but also six thin terms. The results of Study 2 reveal that statements including positive thin terms are also less contradictory than negative thin terms, mirroring the effect for thick terms. While we cannot rule out that the outcome of Study 2 is the result of two independent effects, the similar results for thick and thin terms in Study 2 do indicate that the same cause is driving the effect in both cases.

3.2 Interpretation and Discussion of the Results

Two accounts were stated in the introduction that may account for the polarity effect of thick terms. First, given that thick terms have both evaluative as well as descriptive content, we hypothesized that the entanglement of descriptive and evaluative content might be stronger for negative thick terms than it is for positive thick terms. The greater entanglement for negative thick terms might be down to the nature of the descriptive content of negative thick terms or it might be explained in the way in which evaluative and descriptive content combine. If

such an explanation were to hold, we would not expect the polarity effect to show up for thin terms. In other words, a positive result for thin terms would indicate the falsity of the *thick concept explanation*.

Second, as suggested by Willemsen & Reuter (2021), certain social roles might be in place that govern the use of positive and negative terms. If a person publicly attributes a negative aspect to a another person, she needs to be able to justify the blameworthy aspect more strongly than when attributing a positive aspect. Consequently, the use of negative terms comes with greater social costs, because they can do serious harm and need to have a more solid grounding. If this social norm hypothesis were true, then the polarity effect might as well show up for positive terms. Thus, a positive result would provide some evidence in favor of the social norm explanation.

The results of Study 2 suggest the **thick concept explanation** to be false. In contrast, the data provide evidence that social norms might be key to understand the polarity effect. The **social norm explanation** is also in line with recent results that show that people are less inclined to permit the use of negative thick terms when these are not intended to be used to blame a person (Willemsen & Reuter, 2020). In any case, the polarity effect is indicative of a pervasive linguistic asymmetry that holds between positive and negative evaluative terms.

One might wonder why the mean value for 'honest' is significantly lower than for all other positive items. Interestingly, this experiment is not the first in which 'honest' is an outlier (see Willemsen Reuter 2020, 2021, and Willemsen, Baumgartner, Cepollaro, & Reuter, ms). We believe that there are two possible factors that drive this effect. First, honesty is one of the virtues that can easily become a vice. Some truths are just tough to bear and often conflict with other norms of politeness, respect, and so on. Thus, calling someone honest does not necessarily involve a positive evaluation. These considerations might have affected participants' interpretations of the stimulus, making the positive evaluation particularly easy to cancel. Second, many uses of 'honest' do not seem to be communicating high praise for an agent, but rather that the agent has merely met a certain minimal standard. We can and should expect others to be honest.

In a recent paper, that second aspect has been investigated more thoroughly. Willemsen, Baumgartner, Cepollaro, and Reuter (ms) provide an alternative explanation for the polarity effect that considers the relevance of social expectations for the interpretation of evaluative language. Let's call this explanation the **evaluative deflation explanation**. They argue that acts that count as morally desirable and are referred to by the use of positive terms, such as being compassionate, can either meet our expectations or they can exceed our expectations. The results of a series of studies indicate that people can use positive terms in two ways: first, a proper evaluative way in which speakers intend to praise the agent and, second, in an evaluatively deflated manner to refer to actions that only meet our expectations.

Applying this account to the example of 'honest' above, we can easily see why people might interpret 'honest' in an evaluatively deflated way. In order for communication, in particular, and cooperation, more generally, to work, people need to be honest.[10] Thus, following Willemsen et al.'s suggestion, we should expect that when people call a person's behaviour honest, they often do not want to praise the agent for having exceeded our moral standards. Rather, all they wish to communicate is that the agent meets a certain standard, necessary for people to cooperate. Whether the **social norm explanation**, or the **evaluative deflation explanation**, or an altogether different explanation will prevail is a matter for future research to determine.

Acknowledgments

The research of Lucien Baumgartner, Pascale Willemsen, and Kevin Reuter was funded by the Swiss National Science Foundation (SNSF), grant number PCEFP1_181082. Pascale Willemsen also received generous support by the SNSF, grant number PZ00P1_201737. We would like to thank Bianca Cepollaro, Severin Keller, Ethan Landes, as well as the participants at conferences in for their comments on previous versions of the manuscript. We are grateful for feedback we have received at several conferences and workshops, e.g. the First European XPhi Conference, the 10th Annual Conference of the Sociedad de Lógica, Methodologia y Filosofia de la Ciencia en España, and the XPhi Lab Meeting at the University of Zurich.

References

Anderson, R., Crockett, M., & Pizarro, D. (2020). A theory of moral praise. *Trends in Cognitive Sciences*, 24(9), 694–703.

Beebe, J., & Buckwalter, W. (2010). The epistemic sideeffect-effect. *Mind & Language*, 25(4), 474–498.

Blackburn, S. (1992). Through thick and thin. *Proceedings of the Aristotelian Society*, 66, 284–299.

Carston, R. (2004). Truth-Conditional Content and Conversational Implicature, in *The Semantics/Pragmatics Distinction*, Claudia Bianchi (ed.), Stanford, CA: CSLI Publications, 65–81.

Cepollaro, B. (2020). *Slurs and thick terms. when language encodes values.* Roman & Littlefield.

[10]Truthfulness is one of the central maxims in Gricean and neo-Gricean frameworks (Carston, 2004, Horn, 2004). Also, truthfulness is a key element in many discussions on the norm of assertion (Kneer, 2018, Marsili & Wiegmann, 2021, Reuter & Brössel, 2019).

Cepollaro, B., & Stojanovic, I. (2016). Hybrid evaluatives: In defense of a presuppositional account. *Grazer Philosophische Studien*, 93(3), 458–488.

Del Pinal, G., & Reuter, K. (2017). Dual character concepts in social cognition: Commitments and the normative dimension of conceptual representation. *Cognitive Science*, 41(3), 477-501.

Eklund, M. (2011). What are thick concepts? *Canadian Journal of Philosophy*, 41(1), 25–49.

Elstein, D., & Hurka, T. (2009). From thick to thin: Two moral reduction plans. *Canadian Journal of Philosophy*, 39(4), 515–535.

Grice, H. (1989). Logic and conversation. In H. Grice (Ed.), *Studies in the way of words*, pp. 22-40, Harvard University Press.

Guglielmo, S., Malle, B. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PloS One*, 14(3), e0213544.

Hare, R. (1952). *The language of morals.* Clarendon Press.

Hare, R. (1963). Freedom and reason (ed.). Clarendon Press.

Horn, L. (2004), Implicature, in *The Handbook of Pragmatics*, Horn and Ward (ed.), 2–28.

Kirchin, S. (2010). The shapelessness hypothesis. *Philosophers' Imprint*, 10(4), 1–28.

Kneer, M. (2018). The norm of assertion: Empirical data. *Cognition*, 177, 165-171.

Knobe, J., Prasada, S., & Newman, G. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127(2), 242-257.

Kyle, B. (2020). The expansion view of thick concepts. *Noûs*, 54(4), 914–944.

Marsili, N., & Wiegmann, A. (2021). Should I say that? An experimental investigation of the norm of assertion. *Cognition*, 212, 104657.

Putnam, H. (2002). *The collapse of the fact/value dichotomy and other essays.* Harvard University Press

Reuter, K. (2019). Dual character concepts. *Philosophy Compass*, 14(1), e12557.

Reuter, K., Brössel, P. (2019). No knowledge required. *Episteme*, 16(3), 303-321.

Reuter, K., Löschke, J., & Betzler, M. (2020). What is a colleague? The descriptive and normative dimension of a dual character concept. *Philosophical Psychology*, 33(7), 997-1017.

Roberts, D. (2011). Shapelessness and the thick. *Ethics*, 121(3), 489–520.

Sterken, R. (2017). The meaning of generics. *Philosophy Compass*, 12(8). 1-13.

Sullivan, A. (2017). Evaluating the cancellability test. *Journal of Pragmatics*, 121, 162–174.

Sytsma, J., Bluhm, R., Willemsen, P., & Reuter, K. (2019). Causal attributions and corpus linguistics. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy.* Bloomsbury Academic.

Thakral, R (2018). Generics and weak necessity. *Inquiry*, 1-28.

Väyrynen, P. (2013). *The lewd, the rude and the nasty.* Oxford University Press.

Väyrynen, P. (2021). Thick ethical concepts. In E. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy.* https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts

Wiggins, D. (1993). Cognitivism, Naturalism and Normativity. In Haldane and Wright (Eds.) (1993), 279–300.

Willemsen, P., Baumgartner, L., Frohofer, S., & Reuter, K. (2022). Examining evaluativity in legal discourse: A comparative corpus-linguistic study of thick concepts. in Magen, S., Prochownik, K. (Eds.) *Advances in Experimental Philosophy of Law.*

Willemsen, P., & Kirfel, L. (2019). Recent empirical work on the relationship between causal judgments and norms. *Philosophy Compass*, 14(1), e12562.

Willemsen, P., & Reuter, K. (2020). Separability and the effect of valence. In M. Denison & A. Xu (Eds.), *Proceedings of the 42th Annual Conference of the Cognitive Science Cociety* 2020 (794–800).

Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought*, 10(2), 135-146.

Williams, B. (1985). *Ethics and the limits of philosophy.* Harvard University Press.

Zakkou, J. (2018). The cancellability test for conversational implicatures. *Philosophy Compass*, 13(12), e12552.

Zakkou, J. (2021). Conventional evaluativity. *Australasian Journal of Philosophy*, 1-15.