Language and Cognition (2024), 16: 2, 263–282
doi:10.1017/langcog.2023.35

ARTICLE

Tracing thick and thin concepts through corpora

Kevin Reuter, Lucien Baumgartner and Pascale Willemsen

Department of Philosophy, University of Zurich, Zurich, Switzerland
Corresponding author: Kevin Reuter; Email: kevin.reuter@uzh.ch

(Received 17 November 2022; Revised 17 May 2023; Accepted 01 July 2023)

Abstract
Philosophers and linguists currently lack the means to reliably identify evaluative concepts and measure their evaluative intensity. Using a corpus-based approach, we present a new method to distinguish evaluatively thick and thin adjectives like 'courageous' and 'awful' from descriptive adjectives like 'narrow,' and from value-associated adjectives like 'sunny.' Our study suggests that the modifiers 'truly' and 'really' frequently highlight the evaluative dimension of thick and thin adjectives, allowing for them to be uniquely classified. Based on these results, we believe our operationalization may pave the way for a more quantitative approach to the study of thick and thin concepts.

Keywords: thick concepts; thin concepts; modifiers; truly; evaluation; sentiment; corpus studies

1. Introduction

The two most prominent kinds of evaluative concepts are thin and thick concepts. Thin concepts like GREAT and TERRIBLE evaluate without specifying the descriptive aspects that ground their evaluation. Thick concepts, such as RUDE and COURAGEOUS, describe and evaluate at the same time. For instance, by calling a woman courageous, we not only evaluate her positively but also describe her as willing to take risks and, thus, reveal the descriptive aspect for our positive evaluation. It is this combination of evaluation and description that makes them very efficient communicative tools.[1]

Both thick and thin concepts are ubiquitous in everyday talk and have an important function in assigning blame and praise. Despite their ubiquity and importance, we lack a reliable means to tell thick and thin concepts apart from concepts such as HOMELESS that are value-associated but do not evaluate in the sense of expressing approval or disapproval of someone or something. Concepts from all three categories (thick, thin, value-associated) evoke positive or negative arousal and affect, but only thick and thin concepts are in the business of evaluation. Let us

[1]For introductory and overview articles on thin and thick concepts, see, e.g., Kirchin (2013), Smith (2013), Väyrynen (2021).

illustrate the difference between evaluation and arousal/affect with two examples: concepts like YOUNG and EMPTY are value-associated, but they are not considered to be evaluative: it is a pleasant thing to be young, and an empty glass of beer can be unpleasant, but saying that the glass is empty or that a person is young does not evaluate in the sense of approving or disapproving of someone or something. In contrast, concepts like GENEROUS, INSANE, and UGLY evaluate a person, behavior or object.[2]

Scholars working in ethics, aesthetics, or epistemology usually do not care very much about arousal and affect simpliciter. Instead, they focus their studies on the *evaluative* aspects of honesty, beauty, and justification. It is therefore (or rather should be) a central endeavor to identify evaluative concepts out of the large group of concepts that trigger arousal and affect. Rather surprisingly, no tool for reliably identifying evaluative concepts has so far been developed.[3] Instead, scholars rely almost exclusively on their own intuitions in order to identify thin and thick concepts, usually sticking with a list of examples that most people agree with (see, e.g., Kirchin, 2013; Roberts, 2013; Väyrynen, 2013). Payne (2005, p. 90), for instance, writes: "We still lack a thorough understanding of their [thick concepts] characteristics. We do not know why a particular concept qualifies as a thick concept, beyond a general sense that specific, traditional concepts are likely to be thick." In part, a lack of a more systematic way of identifying thick concepts can be accounted for by the deep disagreement among scholars on how to characterize the notion of the evaluative.

It is certainly helpful to have a list of examples of evaluative concepts – this list contains concepts like RUDE, FRIENDLY, and FUNNY. Such examples allow us to discuss two questions that have received substantial attention in the literature. First, can the evaluative component of thick concepts be separated from the descriptive component (e.g., Kirchin, 2010; Williams, 1985; for a summary of the various positions, see Väyrynen, 2021), and second, is the evaluative component of a concept part of its semantics or communicated pragmatically (Blackburn, 1992; Elstein & Hurka, 2009; Hare, 1952)? However, having merely a list of such examples imposes severe limitations. For one thing, it raises doubts about whether the given answers to these questions can be generalized to a more comprehensive list of concepts: there is no a priori reason to suppose that all thick concepts behave similarly in regard to how evaluative and descriptive content are entangled (Väyrynen, 2013). For another, just operating with a list of examples might hinder further scholarly debate on evaluative concepts. We might wish to ask questions like 'How can we reliably distinguish evaluative concepts from other concepts?' 'Do thick concepts have differing evaluative intensities?' 'How does the evaluative component of an evaluative concept depend on the context in which it is uttered?' and many more.

[2]We do not rule out that value-associated concepts can be used evaluatively in specific contexts. In contrast to context-dependent evaluative uses of value-associated terms, however, thick and thin terms standardly communicate approval or disapproval.

[3]A rare exception is a recently published paper by Stojanovic and Kaiser (2022) in which they empirically investigate the evaluative character of taste predicates. Stojanovic and Kaiser suggest that taste predicates belong to the broader class of subjective terms that only partly overlap with the evaluative terms with which we are concerned in this paper. They propose two novel linguistic tests to detect those taste predicates that can be used not only evaluatively but also neutrally, namely the Attitude Compatibility Test and the Good/Bad Way Test. Stojanovic and Kaiser substantiate their linguistic analyses with corpus data. A forthcoming paper by Stojanovic and McNally (n.d.) also uses corpus data to analyze differences in the way moral predicates and personal taste predicates differ in regard to the way they combine with 'find' and 'consider.'

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

By limiting ourselves to the same examples, we make it rather difficult to identify problems, see where they tend to arise, and diagnose the reasons when our intuitions become unclear or at least controversial. In order to illustrate the lack of consensus and underline the importance of answering questions like those listed above, let us highlight just a few recent controversies. For instance, no consensus exists on whether legal concepts like CONSTITUTIONAL and LEGAL (see, e.g., Enoch & Toh, 2013; Topham, 2016), epistemic concepts like JUSTIFIED and KNOWLEDGE (see, e.g., Kotzee & Wanderer, 2008; Kyle, 2013; Roberts, 2018; Väyrynen, 2021), emotional concepts like HAPPY and AFRAID (see, e.g., Díaz & Reuter, 2020; Phillips et al., 2017), concepts linked to the domain of purity like DIRTY (see, e.g., Curry et al., 2019); Haidt, 2007), and other concepts like CAUSATION (Sytsma et al., 2019), INTENTION (Knobe, 2003), and CONSPIRACY THEORY (Napolitano & Reuter, 2021) that play a central role in contemporary debates in philosophy and psychology are evaluative concepts. There is also no consensus on whether and (if so) how many thick concepts demonstrate variability with respect to their evaluative component. These so-called objectionable thick concepts include, among others, LEWD, CONSERVATIVE, RELIGIOUS, and BLASPHEMOUS (for a classical dispute, see Blackburn, 1992; Dancy, 1995; Gibbard, 1992) for more recent discussions on some of these concepts, see, e.g., Alfano et al., 2018; Baumgartner et al., 2022; Cepollaro & Stojanovic, 2016; Cepollaro, 2018; Eklund, 2011; Väyrynen, 2011; Willemsen & Reuter, 2021).

It seems to us that the lack of consensus in this area is, at least partially, a matter of methodology. Thus, in this paper, we set ourselves two goals. Our first and primary aim consists in developing a suitable method to identify thick and thin concepts. Our second – and admittedly more demanding – aim is to provide a method to measure the evaluative intensity of concepts. If praise and blame serve as a close-enough proxy for evaluation, then we expect some evaluative concepts to be more evaluative than others. At least intuitively, demonstrating *good* behavior is likely to be less praiseworthy than displaying *exemplary* behavior. Especially for thin concepts, it seems that our stock of English terms (substandard < inadequate < bad < horrible) allows people to easily express different degrees of evaluation. In the next section, we will present a method that provides us with a way of operationalizing the evaluative intensity of thick, thin, and value-associated concepts that aims to reflect the varying evaluative intensities of concepts. That said, our primary goal is the identification of evaluative concepts through empirical, nonintuitive means.

Before we present our method, we need to deal with an objection that readers familiar with lexical sentiment analysis might have, namely, that we are already in possession of a classification and measurement device for evaluative concepts. Lexical sentiment analysis is a flourishing and growing research area with the central aim of determining the sentiment values of terms, phrases, sentences, and whole texts.[4] In computational linguistics, the term 'sentiment' is often referred to as an aspect or indicator of the broader concept of subjectivity (Benamara et al., 2012; Mohammad, 2016; Taboada et al., 2011). Taboada (2016, p. 326), for instance, defines 'sentiment' as "the expression of subjectivity as either a positive or negative opinion." With regard to sentiment *analysis*, that is, the real-world application of sentiment

[4]Sentiment analyses often rely on sentiment dictionaries like sentiWords and senticNet that contain both the polarity and the intensity of words, i.e., whether a word out of context evokes something positive or negative.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

annotation procedures, Esuli and Sebastiani (2006) specify three core aspects: (i) determining whether the text data is factual or an opinion, and, in the case where it expresses an opinion, (ii) the polarity, semantic orientation, or valence (e.g., Hatzivassiloglou & McKeown, 1997; Osgood et al., 1957) of the text data, as well as (iii) its intensity. Importantly, 'subjectivity' and 'sentiment' are often used as umbrella terms, covering appraisal, subjective belief, emotion, evaluation, stance, and attitude. Thus, both terms appear to be too coarse-grained, ultimately raising the question of what we are actually measuring with lexical sentiment values.

Given this coarse-grainedness of *sentiment*, lexical sentiment analysis is not up to the task we set ourselves. Evaluative terms only form a proper subset of all terms that receive high values in sentiment analysis. Take the examples 'young' and 'empty' from above, as well as the terms 'sunny' and 'moldy.' Using sentiment scores, we are simply not able to distinguish evaluative from value-associated words based on sentiment dictionaries. Where '–1' is the most negative value and '+1' the most positive in the dictionary sentiWords (Gatti et al., 2016), 'sunny' and 'honest' have the same score of +0.76, 'young' and 'diligent' have the same score of +0.32, 'empty' and 'careless' have the same score of −0.33, and 'moldy' and 'rude' have the same score of −0.71.

Given the limitations of both intuitive classification and sentiment analysis, we propose to approach the identification and measurement of thick and thin concepts using tools from corpus linguistics.[5] We present the results of a corpus-linguistic study for a wide range of thick, thin, value-associated, and descriptive adjectives in Section 2. Our study reveals that the modifiers 'truly' and 'really' highlight the evaluative dimension of thick and thin concepts, allowing for them to be reliably classified. We discuss the limitations of our methodology, some implications of our research, and the likely success of quantifying the study of thick and thin concepts in Section 3. For the most part of the next section, we change terminology and rather talk about terms and adjectives and how they are used instead of the concepts that are expressed with such terms. Such a terminological shift is appropriate given the corpus-related aspects of our study. In Section 3, we will discuss more generally whether our results should be more carefully interpreted to reveal aspects of uses of terms in certain contexts, or whether we are indeed allowed to draw conclusions at the level of concepts.

2. Corpus-linguistic study

The linguistic method we present in this section does not represent a *unified* approach to investigate all word classes. Thick nouns like 'filth' and 'champion,' thick adjectives like 'honest' and 'rude,' as well as thick verbs like 'insult' and 'brag' can hardly be investigated by the same means given their different functions in a sentence.[6] In this paper, we focus on thick and thin *adjectives*, which undoubtedly have received the greatest attention of all types of thick terms.

[5]In recent years, the use of corpus-linguistic methods has been steadily on the rise in epistemology (e.g., Hansen et al., 2021; Nichols & Pinillos, 2018; Reuter & Baumgartner, n.d.b), history of philosophy (e.g., Alfano, 2018), metaphilosophy (Andow, 2015; Mizrahi, 2020), philosophy of mind and language (e.g., Bordonaba, n.d.; Fischer et al., 2015; Reuter, 2011; Sytsma et al., 2019). For an overview, see e.g., Chartrand (2022) and Reuter & Baumgartner. n.d.a.

[6]Of course, 'filth' has an adjectival form, and 'honest' a noun form, which allows at least for some extended interpretation of our studies.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Our approach takes inspiration from recent research on dual character concepts (Del Pinal & Reuter, 2017; Knobe et al., 2013; Leslie, 2015; Reuter, 2019). Dual character concepts are concepts that are often, perhaps mostly, used descriptively but also encode an independent normative dimension.[7] For example, Julie will be considered a mechanic (descriptively) if she works at a garage fixing cars for customers. This holds regardless of whether she is committed to and enjoys what she is doing. We might also think of people as mechanics (normatively) if they have a passion for fixing things. And while this is probably most often the case when they fix things professionally, this need not be the case. For example, we might say of Andre, the philosopher, that he is a 'true' mechanic because he spends all his spare time fixing things instead of reading philosophy; here, the true modifier operates on the normative dimension of dual character concepts, as suggested and empirically investigated by Del Pinal and Reuter (2017) and Knobe et al. (2013).

Dual character concepts are a class of evaluative concepts apart from thick concepts. In contrast to dual character concepts, the descriptive and evaluative content of thick concepts is not doubly dissociable. If we say of Julie that she is courageous, we cannot choose to use the term 'courageous' merely normatively and without its descriptive meaning. But while thick and dual character concepts are different kinds of concepts, it seems the normative dimension of thick and thin concepts can be highlighted in a similar fashion. Whereas the true modifier has been argued to stress the normative dimension of dual character concepts, as in 'true mechanic' or 'true scientist,' the modifier 'truly' might well intensify the evaluative aspect of thick and thin adjectives, as in 'truly courageous' and 'truly awful.' A possible connection between the use of 'truly' (and 'really') with evaluative adjectives has also been posited by Liu and Espino (2012) as well as Simon-Vandenbergen and Taverniers (2014), although it needs to be said that our view on what counts as an evaluative adjective differs from their perspectives.

If 'truly' indeed highlights or intensifies the evaluative aspect of thick adjectives, then 'truly x' should sound more acceptable for adjectives that have an evaluative component, like 'truly honest,' compared to value-associated adjectives, like 'truly sunny,' or 'truly large.' Translating this into a hypothesis for corpus-analytic studies, we predict that 'truly x' is more common for thick and thin adjectives compared to descriptive and value-associated adjectives. In other words, the 'truly' modifier allows us to distinguish those classes of concepts that have an evaluative dimension (thick and thin concepts) from those that do not (descriptive and value-associated concepts).[8]

Although 'truly' seems to intensify the evaluative aspect of thick (as well as thin) adjectives, the philosophical literature on thick concepts does not feature any discussion of the role of the modifier 'truly' to raise the evaluative aspect of thick terms. This might be surprising, especially given that other modifiers like 'too' as in

[7]Within the literature on dual character concepts, scholars typically employ the terminology 'normative' as opposed to 'evaluative' when addressing the nondescriptive facets of such concepts. We abide by this convention, but see Reuter (2019) for a discussion of the interplay between the normative and the evaluative aspects in dual character concepts.

[8]Although our approach is motivated by research on dual character concepts, we do not investigate dual character concepts in this paper. Dual character concepts usually do not come in adjectival form but rather in noun forms like 'artist' and 'scientist.' Consequently, the truly modifier cannot be applied as straightforwardly to examine dual character concepts.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

'too courageous,' and 'not enough' as in 'not rude enough,' have been intensely debated as a putative means to change the polarity of the thick term in question. One reason for this omission could be the rather infrequent use of the term 'truly.' The relatively scarce use (67,683 hits on the Corpus of Contemporary American English (COCA)) of 'truly' might also be a problem for our purposes, because it makes a corpus analytical study less robust to artifacts. We therefore decided to explore other modifiers that might have a similar function. Liu and Espino (2012, p. 198) argue that the modifiers 'actually' (353,908 hits on COCA), 'genuinely' (9,061 hits on COCA), 'really' (896,050 hits on COCA), and 'truly' are near synonymous but also have important semantic and usage differences.

While 'actually' is rarely used to modify adjectives (Liu and Espino, 2012, p. 214), 'genuinely' often modifies adjectives and thus might be a good additional modifier for our study. Unfortunately, 'genuinely' is far less frequent than 'truly.' In contrast, 'really' is over 10 times more frequent than 'truly' and is also commonly applied to modify adjectives. Based on their analysis, Liu and Espino (2012, p. 212) argue that (a) 'truly' is more formal than 'really' (210), and (b) 'really' is the most versatile modifier (217). Given the strong semantic similarities between 'really' and 'truly,' as well as the very common use of 'really' as a modifier for adjectives, we extended our investigation to also cover 'really' as a possible means to distinguish truly evaluative terms from mere value-associated terms as well as descriptive terms. Importantly, our claim is not that the modifiers 'truly' and 'really' cannot be reasonably applied to highlight other aspects of the adjective they modify, but rather that those modifiers are frequently used with evaluative adjectives, such that patterns of use emerge that reveal differences between the concept classes at stake.[9]

2.1. Stimuli and methods

A wide selection of adjectives – an assortment of thick, thin, merely descriptive, and value-associated – is needed to provide the data for achieving the two desiderata mentioned above. We therefore selected 45 adjectives to be investigated in our study:

- 6 thin adjectives: 3 positive (good, great, terrific) and 3 negative (awful, bad, terrible)
- 10 moral thick adjectives: 5 positive (compassionate, courageous, friendly, generous, honest) and 5 negative (cruel, reckless, rude, selfish, vicious)
- 10 nonmoral thick adjectives: 5 positive (beautiful, delicious, funny, justified, wise) and 5 negative (boring, disgusting, insane, stupid, ugly)[10]

[9]There certainly are many other differences between the use of 'truly' and 'really'. For example, we often say 'Really?' (but not 'Truly?') in order to express our surprise. These differences have no bearing, though, on the question at stake, as we only investigate these terms in their function to modify subsequent adjectives.

[10]While most philosophical discussions on thick concepts focus on terms from the moral domain, thick terms are also frequent and increasingly discussed in the epistemic domain (insane, justified, stupid, wise), the aesthetic domain (beautiful, ugly), the culinary domain (delicious, disgusting), and the entertainment domain (boring, funny). Some terms are not restricted to a single domain only but can be applied across several domains, e.g., food, a comedian, a proof, a building might all be considered boring.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

- 10 value-associated adjectives: 5 positive (quiet, rich, tall, shiny, sunny) and 5 negative (bloody, broken, closed, empty, homeless)
- 9 purely descriptive adjectives: dry, large, loud, narrow, permanent, rainy, short, wooden, yellow

We selected adjectives that are fairly common English words. Each of the terms has at least 5,000 hits on the COCA ('courageous' being the only exception, with 4,742 hits). All value-associated adjectives had a sentiment value of at least 0.25 (absolute number) with an average absolute value of 0.49 (SD = 0.16). Thick terms and thin terms had similar average absolute sentiment values of 0.56 (SD = 0.20) and 0.63 (SD = 0.11). It is thus unlikely for the sentiment values to have had any confounding effect on our studies. Purely descriptive adjectives had an average absolute sentiment value of 0.06 (SD = 0.07).

The grouping of adjectives into descriptive and value-associated concepts was based on sentiment scores in the dictionary sentiWords (Gatti et. al, 2016). The classification into thick, thin, and value-associated adjectives was based on the authors' intuitions as well as claims from the thick concepts literature. As no comprehensive classification has so far been theoretically defined and empirically verified, some reliance on intuitions was unavoidable. That said, some of the analyses we have done dispense with any intuition-based precategorization. But most importantly, our aim is to operationalize the evaluative dimension of thick adjectives and test how well our methods categorize those adjectives into one of the sets we started with (thin, thick, value-associated, descriptive).

Investigating how often an adjective x is modified by 'truly' and 'really' is comparatively easy. With a sufficiently large corpus, we can simply record the number of hits for 'truly x' and 'really x' and divide this number by the number of hits for 'x.' This will give us the respective ratios. We decided to use the Corpus of Contemporary American English (COCA) for this task. The advantage of using such a simple, pre-existing corpus is that anybody can (a) directly replicate our results with the concepts we used and also (b) investigate whether concepts we did not include behave similarly or differently, thereby supporting or challenging our main conclusions.

We also included a control condition. The modifier 'very' is generally used to indicate high levels of a certain property, e.g., when saying "Her behavior is very courageous," or "Today, it is very sunny." Descriptive and value-associated terms should be just as much open to intensification by the modifier 'very' as are thin and thick terms. Of course, absolute adjectives like 'perfect' and 'permanent,' as well as extreme adjectives like 'great' and 'insane,' are susceptible to the use of 'very' to little or no extent. Thus, we need to factor in whether or not an adjective is gradable. At the same time, a comparison between 'truly' and 'really' on the one side, and 'very' on the other side, should allow us to determine whether any positive result can be accounted for by other features of modification.

For all 45 concepts, we recorded the amount of hits for 'truly x,' 'really x,' and 'x' on COCA. We then calculated the ratios (e.g., # 'truly rude' divided by # 'rude') and normalized them for all 45 concepts.[11] The value for $eval$ – our variable for evaluative

[11]The normalization value for 'truly' (i.e., average value of all 45 truly ratios) was 0.0836; the normalization value for 'really' (i.e., the average value of all 45 really ratios) was 0.560.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

intensity – was then calculated by taking the average of both normalized ratios. Thus, the following equation was used to determine *eval*-values:

$$eval_x = \left(\frac{trulyratio_x}{\frac{\sum_{n=1}^{45} trulyratio_x}{45}} + \frac{reallyratio_x}{\frac{\sum_{n=1}^{45} reallyratio_x}{45}} \right) \cdot \frac{1}{2}$$

Thus, despite having many more hits for 'really x,' the data for both 'truly x' and 'really x' are represented equally strong in our study. Given the relatively infrequent use of 'truly' with some terms on COCA, as well as some general worries that COCA is not a representative corpus for everyday talk, we decided to run a robustness check with a small selection of terms ('bad,' 'empty,' 'generous,' 'honest,' 'short,' 'stupid') using Reddit data (see also Supplementary Material), collected via the Pushshift API (Baumgartner et. al., 2020). For the data collection, we first queried 200 instances of 'truly x,' starting on August 31, 2020 (t_1). The query is going back in time and stops as soon as we hit the 200th mention of 'truly x' (t_2). This gives us a different time period for different target phrases. For example, the time period for 'truly bad' is a lot smaller than for 'truly empty,' since the latter is much more rare. In a second step, we queried for 'really x' for the time period determined by the truly query ($t_1 - t_2$). Finally, we did the same for the adjective without modifiers (i.e., the total number of occurrences). This means that the counts for 'really x' and the adjective without modifiers are indexed to $t_1 - t_2$: for 200 'truly x,' we have *n* 'really x' and *m* 'x.'

2.2. Results

Table 1 displays the values for *eval* as well as 'truly' and 'really' uses per thousand hits for all 45 adjectives, grouped according to which class they were originally assigned to. The values for *eval* show that almost all value-associated and descriptive terms had lower values than thick and thin terms. Only for 'rich' and 'loud' did the modifier approach yield results that put them above some thick terms. All other descriptive and value-associated terms were well below the lowest-ranked thick terms.[12] There were also substantial differences between thick moral terms, thick nonmoral terms, and thin terms. Most thin terms had higher *eval* numbers than moral thick terms, with nonmoral thick terms being mostly positioned between thin and thick moral terms.

The results for the 'very' modifier show a markedly different pattern, according to which many descriptive and value-associated terms are used roughly as frequently compared to thin and thick terms. For example, gradable descriptive adjectives like 'loud,' 'rainy,' 'narrow,' as well as gradable value-associated adjectives like 'quiet,' 'shiny,' and 'tall' are used as commonly with the 'very' modifier as thin and thick

[12]We also did a pairwise comparison for *eval* numbers between positive and negative terms (based on sentiment values from sentiWords). Recent research has shown that the evaluative component can be more easily cancelled for positive thick terms compared to negative thick terms (Willemsen & Reuter, 2021). A *t*-test revealed that the average value for *positive* terms (M = 0.817, SD = 0.72) marginally failed to be significantly lower than the rating for *negative* terms (M = 1.27, SD = 1.25), ($t(43) = 1.51$, $p = 0.069$). Further investigations using a greater number of values are necessary to find out whether positive and negative terms differ from each other.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Table 1. *eval* values, as well as 'truly,' 'really,' and 'very' uses per thousand (per mill) for all 45 adjectives using data from COCA, as well as the average values for each predefined category. For example, take the values for the adjective 'courageous': for every 1,000 uses of the term 'courageous,' we find that it is modified with 'truly' 2.74 times, with 'really' 4.22 times, and with 'very' 50.19 times.

Class	Adjective	Eval	'truly' per mill	'really' per mill	'very' per mill (control)
thin	awful	4.66	5.97	12.16	1.54
	bad	2.18	0.33	22.17	21.21
	terrific	1.87	0.08	15.43	0.75
	terrible	1.84	1.96	7.46	2.53
	good	1.72	0.17	18.14	38.07
	great	1.58	1.20	9.63	2.25
$\overline{thin}$		2.31	1.74	14.17	11.06
thick moral	courageous	2.02	2.74	4.22	50.19
	compassionate	1.84	2.72	2.36	24.82
	honest	1.14	1.34	3.88	16.02
	selfish	0.97	0.99	4.26	18.73
	vicious	0.92	1.11	2.84	6.92
	rude	0.78	0.17	7.57	30.87
	generous	0.63	0.56	3.29	77.64
	reckless	0.58	0.86	0.69	5.52
	cruel	0.56	0.40	3.57	17.78
	friendly	0.49	0.18	4.25	35.79
$\overline{thick - moral}$		0.99	1.11	3.69	28.43
thick nonmoral	disgusting	4.04	4.91	12.37	3.40
	funny	1.98	0.49	18.95	54.33
	boring	1.78	0.45	16.96	18.19
	ugly	1.77	0.96	13.39	20.23
	insane	1.64	2.24	3.30	0.31
	stupid	1.63	0.40	15.52	7.45
	beautiful	1.61	1.42	8.50	17.08
	delicious	1.46	1.22	8.18	5.54
	wise	0.89	1.17	2.17	27.48
	justified	0.74	0.94	1.98	1.20
$thick - \bar{n}onmoral$		1.75	1.42	10.01	15.52
value-assoc	rich	0.57	0.43	3.48	21.28
	quiet	0.39	0.05	4.00	29.23
	broken	0.33	0.29	1.74	0.95
	tall	0.31	0.00	3.46	19.55
	shiny	0.20	0.00	2.18	7.35
	sunny	0.20	0.17	1.04	3.98
	bloody	0.18	0.10	1.31	5.85
	empty	0.16	0.17	0.70	1.64
	homeless	0.11	0.15	0.20	0.20
	closed	0.09	0.07	0.53	1.27
$\overline{value - assoc}$		0.25	0.14	1.86	9.13
descriptive	loud	1.02	0.00	11.45	23.23
	dry	0.17	0.04	1.65	9.93
	permanent	0.17	0.25	0.25	0.39
	large	0.16	0.13	0.89	30.35
	short	0.16	0.01	1.68	25.54
	narrow	0.14	0.03	1.34	33.05
	rainy	0.12	0.00	1.40	5.80
	yellow	0.03	0.00	0.34	0.48
	wooden	0.00	0.00	0.00	0.22
$\overline{descriptive}$		0.22	0.05	2.12	14.11

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Table 2. Ratios and *eval* values for six adjectives using data from Reddit.

Target	'really' (*n*)	'truly' (*n*)	'really' per mill	'truly' per mill	Eval_{reddit}	Eval_{COCA}
bad	21,806	200	28.59	0.26	1.60	2.18
empty	1,329	200	1.99	0.30	0.48	0.16
generous	5,372	200	16.52	0.62	1.51	0.63
honest	706	200	2.41	0.68	0.97	1.15
short	48,756	200	1.71	0.01	0.09	0.16
stupid	6,350	200	16.34	0.52	1.37	1.63

terms like 'bad,' 'honest,' 'rude,' 'boring,' 'beautiful.' Of course, we do not deny that within the different groups of adjectives, there are substantial differences in the frequencies with which adjectives are modified by 'very.'[13]

We also calculated the evaluation values for six adjectives using Reddit to check for the robustness of the data from COCA. The calculation performed on Reddit data delivered similar values (see Table 2). Especially the 'truly' ratios seem to be very robust, while we observe a bit more variation for the 'really' ratios.

As one of the central aims of this study is to develop a method that will help us assign terms to categories without relying on people's intuitions, we wanted to know how well a cluster analysis would perform on the given terms. We performed a hierarchical cluster analysis using squared distance (Ward's method) to identify the inherent structure of the data. The cluster analysis is univariate, based solely on the eval values, and thus only intended as a sanity check. We excluded 'awful' and 'disgusting' as their *eval*-values are outliers. The results are displayed in the form of a tree diagram (Fig. 1). We specified three clusters for the cluster analysis, which map quite well onto our distinction between evaluative and nonevaluative terms. In the 'descriptive' cluster (middle), only descriptive as well as value-associated concepts were included (not a single thick or thin term). The two 'evaluative' clusters (left and right) included only two terms that were originally classified as descriptive or value-associated ('loud' and 'rich'; shown in light grey in Fig. 1). Most terms in the left evaluative cluster are either thin terms or nonmoral thick terms. Not a single thin term was assigned to the right evaluative cluster.

Based on the *eval* numbers, we calculated the predicted class membership for an adjective using a multinomial logit model, providing additional support for the separability of evaluative adjectives. We decided to run the model without the terms 'awful' and 'disgusting,' as they are outliers in the data (see Table 1). As can be seen in Fig. 2, the predicted probabilities for the different classes exhibit different progression patterns along the average of normalized ratios for truly and really.[14]

Interestingly, the accuracy by class for value-associated concepts (74.4%) is relatively high, similar to thick nonmoral (76.0%) and thick moral concepts (78.9%). Descriptive (58.2%) and thin concepts (57.4%), on the other hand, are classified much less accurately – thin concepts get mostly (80.0%) misclassified as

[13] A Pearson correlation test between the very-ratios and the eval numbers shows that they do not correlate significantly ($\rho = 0.0276$, $t(43) = 0.18128$, $p = 0.857$) on 0.05 alpha level. This suggests that 'very' is indeed used differently and does not allow analogous inferences.

[14] Overall the model has an accuracy of 53.5% (CI 37.7%; 68.8%) at a no-information rate of 23.3%. This means the model is significantly more accurate than just picking the most prevalent observed class.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

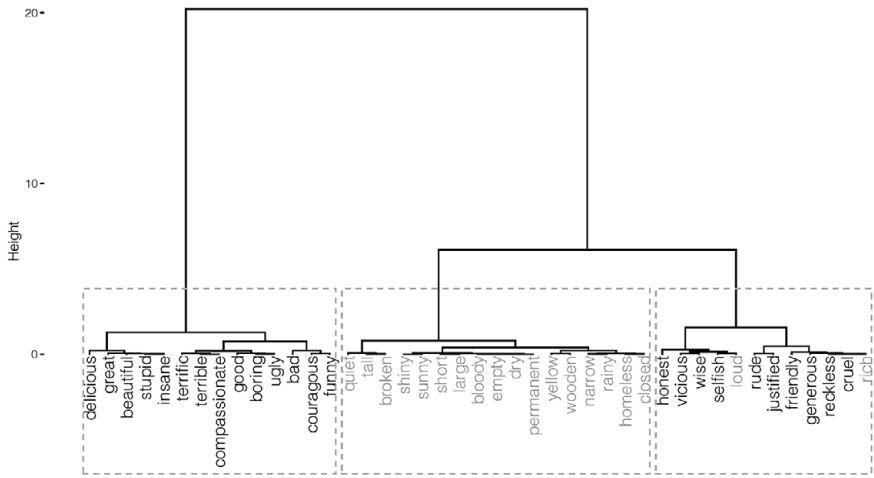

Figure 1. Tree diagram displaying the clusters using hierarchical cluster analysis. We specified three clusters, two of which (left and right) feature almost only evaluative terms, and one cluster (middle) only contains nonevaluative adjectives. Terms in light grey are the selected nonevaluative adjectives, whereas the ones in black are evaluative adjectives.

thick nonmoral concepts, and descriptive concepts get misclassified (66.6%) as value-associated concepts.

In this paper, we are primarily interested in the binary classification of inherently evaluative (thin and thick concepts) and nonevaluative concepts (value-associated and descriptive concepts). The multinominal model above had its difficulties distinguishing thin concepts from nonmoral thick concepts. In the binary approach, we no longer need to discriminate between the two, since both are evaluative concept classes. Fig. 3 shows the probability that an adjective is evaluative only based on its *eval* value using a logistic regression model. This graph shows that adjectives that have an *eval* value >0.75 are quite likely to be thick or thin. Below an eval value of 0.3, adjectives are far more likely to be descriptive or value-associated. The logistic regression model has an accuracy of 90.7% (CI 77.9%; 97.4%), at a no-information rate of 55.8%. The only false classifications were 'cruel' (true: eval.), 'friendly' (true: eval.), 'rich' (true: noneval.), and 'loud' (true: noneval.). While the classification of 'cruel' as a nonevaluative term is surprising, 'friendly' can arguably be expected to be used with a low evaluative intensity on a regular basis. The value-associated terms 'rich' and 'loud' received high *eval* numbers primarily because of their frequent combination with the modifier 'really.' So, arguably, the more varied use of the modifier 'really' creates some confounding noise in the data.

2.3. Discussion

Inspired by recent research on measuring the evaluative component of dual character concepts, we examined the use of intensifiers 'truly' and 'really' for thin, thick, descriptive, and value-associated adjectives. A cluster analysis as well as a multinomial logit model to predict class membership that we performed over all 45 adjectives yielded very promising results, showing that the intensifier method can be

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

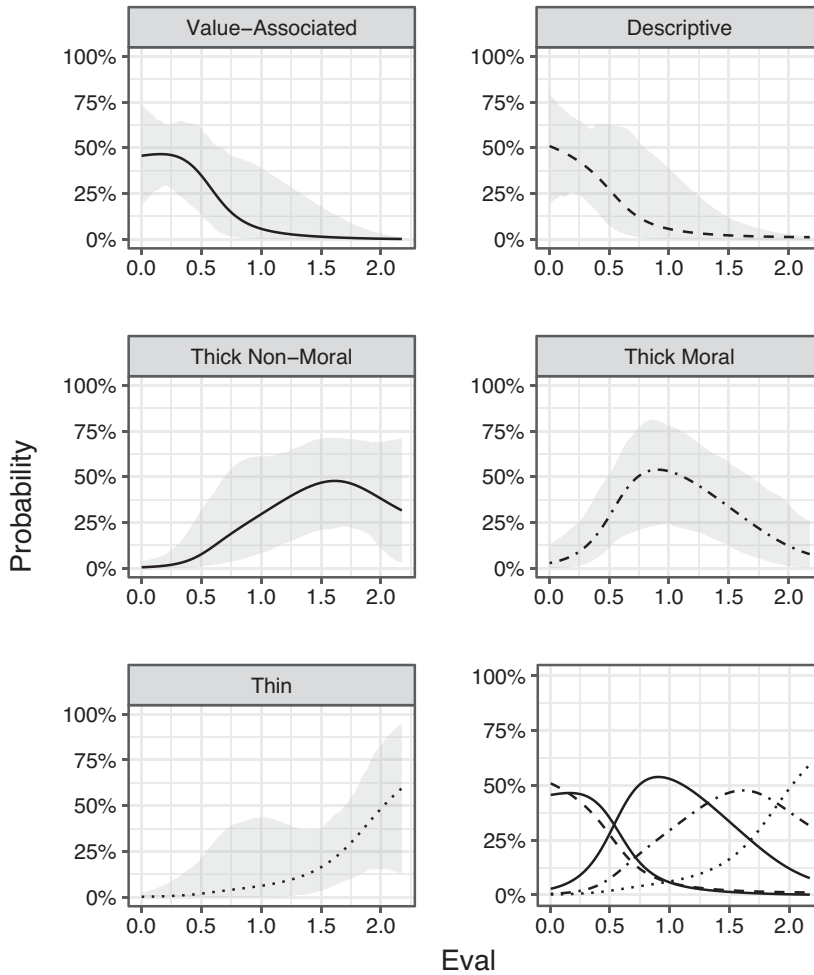

Figure 2. Predicted probabilities for class membership. The first five charts are single plots for the 5 predefined categories, the sixth is a combined plot without confidence intervals. The x-axis indicates the $eval$ number. The outliers 'awful' and 'disgusting' were not included.

utilized to identify evaluative adjectives (both thin and thick) and separate them from both descriptive and value-associated adjectives. We consider the results that demonstrate a separation of evaluative adjectives from value-associated adjectives particularly encouraging.

A further observation concerns differences between thick moral terms, thick nonmoral terms, as well as thin terms. Almost all thick moral terms, except 'courageous' and 'compassionate,' had lower $eval$ numbers compared to thin terms. This is not a very surprising result. Given that thin terms only have an evaluative but no descriptive content, the modifiers 'truly' and 'really' are likely to be applied more frequently to highlight how bad or good something is. In contrast, descriptively rich thick terms have the function to describe aspects of the world with a more subtle communicative, evaluative purpose.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

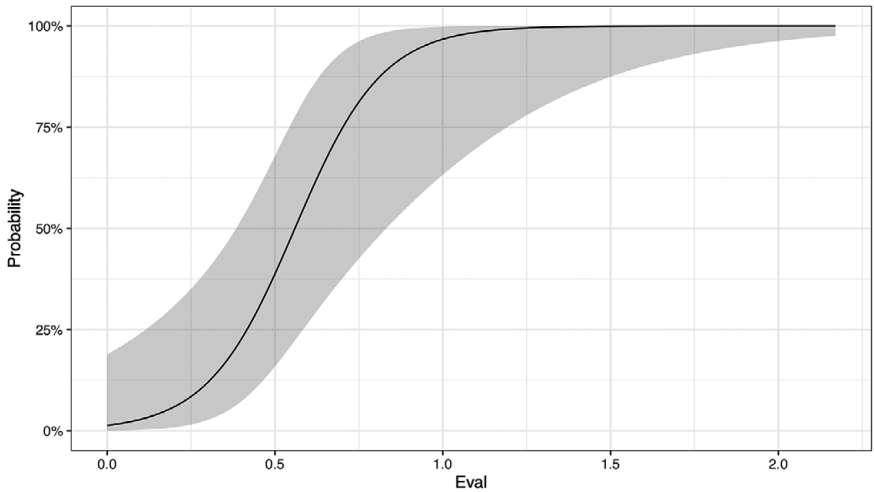

Figure 3. Predicted probability for being an evaluative concept, with confidence intervals.

Most thick *nonmoral* terms had higher *eval* numbers than the investigated thick *moral* terms: thick nonmoral adjectives have rates comparable to thin adjectives (i.e., similarly high).[15] This might be explained by the fact that some of the thick nonmoral terms are descriptively thinner than the moral terms. In fact, some philosophers have argued that aesthetic terms like 'beautiful' and 'ugly' and epistemic terms like 'knowledge' are thinner than moral terms (Kirchin, 2013; Väyrynen, 2008). Zangwill (2013), for instance, argues for the thinness of the term 'beautiful' because descriptively richer terms like 'elegant,' 'delicate,' 'balanced,' and so forth are merely "*ways* – ways of being beautiful" (Zangwill, 2013, p. 317). Chappell (2013), who generally questions the existence of thin concepts, similarly claims that "if there are any thin concepts in aesthetics, perhaps beautiful is one of them" (Chappell, 2013, p. 187). Others have been more skeptical defending the thickness of aesthetic and epistemic terms (Kyle, 2013; Roberts, 2018). Our results do provide some support for authors like Zangwill and Chappell, at least in suggesting that most of the nonmoral thick terms we investigated are thinner than moral thick terms. An additional interpretation may be that evaluative words related to areas like aesthetics or taste have the ability to elicit more intense emotional reactions, thereby making them more likely to be utilized in constructions typically associated with evaluation.[16]

3. General discussion

Empirical work on thick concepts is in its infancy. Research on thick concepts has been mostly theoretical (but see Reuter et al., 2020; Willemsen & Reuter, 2021) for some very recent experimental studies). Consequently, many claims that have been made with regard to the nature and structure of thick concepts are based on the

[15]This explains the low accuracy for classifying thin concepts, as thin adjectives were mostly misclassified (80%) as nonmoral adjectives.

[16]We would like to thank one of the reviewers of this journal for this suggestion.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

linguistic intuitions of a small group of individuals. Such overreliance on individual intuitions places severe limitations on current projects on thick and thin concepts, including efforts to answer questions about which concepts are evaluative (see also the current controversies we listed in the introduction) and efforts to expand the scope of questions that scholars can meaningfully address.

3.1. Summary of the results

In this paper, we have introduced a new corpus-based tool for identifying evaluative concepts and measuring the extent to which thick concepts are used evaluatively. We recorded the frequencies with which thin, thick, descriptive, and value-associated adjectives combine with the intensifiers 'truly' and 'really.' Our principal findings are as follows:

- Thick and thin adjectives are more frequently used with the intensifiers 'truly' and 'really' compared to descriptive and value-associated adjectives. Subsequently, thick and thin adjectives can be differentiated from descriptive adjectives and value-associated adjectives.
- Thin adjectives are more often modified with 'truly' and 'really' compared to thick moral adjectives, with more varied results for thick *nonmoral* adjectives.
- Descriptive and value-associated adjectives are used roughly as frequently with the modifier 'very' compared to thin and thick nonmoral adjectives.

Philosophers often assume that thin and thick concepts form unique classes with features that set them apart from other classes of concepts. While this assumption is a matter of long-standing tradition and enjoys some prima facie plausibility, no empirical evidence has so far been presented in its favor. Our study suggests that the evaluative component of thick and thin concepts can be emphasized by using modifiers such as 'truly' and 'really.' Most descriptive and value-associated concepts do not work in the same way and cannot be as easily combined with these intensifiers. Our results therefore present the first empirical evidence of its kind that thick concepts might indeed form a unique class of concepts.

In the final sections of this paper, we first address some limitations of the proposed methodology. We then discuss whether we have succeeded in operationalizing and measuring the evaluative component of thick and thin adjectives.

3.2. Limitations and moving forward

Using large-scale corpora to examine linguistic hypotheses has some well-known advantages and disadvantages. In contrast to conducting vignette studies, not directly manipulating the stimuli means less control over the actual phenomena to be examined. On the positive side, corpus analysis provides relatively unbiased access to the way linguistic entities work. And importantly, the large corpora we assembled make us confident that the results are reliable and robust.

Some limitations to our studies are structural and could not have been avoided. We can only make claims regarding the class of *adjectives*, because our operationalization targeted only this class of words. Finding out whether thick and descriptive nouns, verbs, adverbs, and so on behave similarly is beyond the scope of this paper.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

We aim to direct our attention to other classes of words in follow-up studies. For example, it seems a reasonable assumption that when people use evaluative words, they like to specify the intensity of the evaluation. Thus, assuming 'kitsch' and 'filth' to be thick nouns compared to the descriptive nouns 'ornament' and 'dust,' we do expect composites like 'terrible kitsch' and 'disgusting filth' to appear more frequently than 'terrible ornament' and 'disgusting dust.'

But even if the focus is on adjectives only, we can certainly do a more fine-grained analysis. Let us quickly highlight two areas in which such an analysis seems promising. First, the evaluative force of many words is likely to vary with context. In fact, some scholars on evaluative concepts have even argued that both thin and thick concepts can be used nonevaluatively (see, e.g., Stojanovic & Kaiser, 2022; Willemsen et al., n.d.).[17] This raises the immediate question of whether we should soften our conclusions, limiting them to specific uses of thick and thin terms. We believe that such a limitation would be overly restrictive and is not supported by our data. Looking at our data *qualitatively*, we find that the selected thin and thick terms are used in a wide variety of contexts. In other words, the effects we found do not appear to be based on specific uses of evaluative terms (using COCA, readers can easily check this for themselves). That said, we do not claim that evaluative terms cannot be used with very different evaluative purposes.

Second and relatedly, the evaluative intensity of adjectives might be influenced by whether they describe animate objects or inanimate and abstract objects. We have not controlled for either of these two factors in our analyses. In future studies, we plan to run structural topic models (STMs) to inductively annotate topic labels (see, e.g., Egami et al., 2018) and use automatic animacy classification (Bjerva, 2014; Bowman & Chopra, 2012; Jahan et al., 2018) to investigate how our results change once these aspects are factored in.

3.3. A new tool for measuring evaluative intensity?

Our study was designed to measure the evaluative dimension of concepts. In this paper we have sketched what an operationalization and measurement of evaluative intensity could look like. More specifically, we proposed that a good indicator or proxy for the evaluative intensity of a term is the extent to which the intensifiers 'truly' and 'really' can be reasonably applied to that term. We then operationalized that proxy through the ratio between the frequencies with which a term is intensified by 'truly' and 'really' and the overall frequency of that term, leading us to the variable $eval$. The results revealed a rather differentiated picture, according to which 'truly' and 'really' are most reasonably applied to thin concepts and thick concepts, and not very reasonably applied to value-associated and descriptive concepts. Unfortunately, our design does not allow us to say whether adjectives that receive low $eval$ numbers are evaluative to a very low degree (or in very few contexts), or whether there is a threshold that distinguishes value-associated from evaluative terms.

Now, the crucial question is: are we justified in saying that $eval$ tells us the evaluative intensity of a term? Here are five reasons to answer this question in the affirmative, if tentatively:

[17]For more traditional arguments for and discussions of the evaluative flexibility of thick concepts, see Blackburn (1992), Gibbard (1992), Hare (1963), Richard (2008), Väyrynen (2013, 2021).

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

1. We have motivated the operationalization of evaluative intensity through the modifiers 'truly' and 'really,' independently of the results we collected: first, researchers on dual character concepts have stressed the importance of the modifiers 'true' and 'real.' Second, linguists have suggested that the functionality of the modifiers 'truly' and 'really' includes the highlighting of evaluative aspects of adjectives (Liu & Espino, 2012; Simon-Vandenbergen & Taverniers, 2014).
2. A cluster analysis demonstrated that *eval* allows us to match most pretheoretic intuitions on the level of classes of concepts.
3. Almost all value-associated concepts received very low *eval* values.
4. Investigating the use of the 'very' modifier as a control condition reveals that not all modifiers allow for a neat categorization of evaluative and nonevaluative concepts.
5. For thin concepts, *eval* values matched the semantic meanings of the terms: 'terrific' is more evaluative than 'good,' 'awful' is more evaluative than 'bad,' and, correspondingly, 'terrific' and 'awful' received higher *eval* values.

We also have some reasons to be skeptical that *eval* uniquely encodes evaluative intensity. First, some results we collected do not match our pretheoretic intuitions about these terms (e.g., 'reckless' has a similar *eval* value as 'rich'). Second, the two modifiers, especially the modifier 'really,' are certainly not exclusively used to highlight the normative dimension of evaluative adjectives. They can also be used for standard-raising. For instance, a truly rich person is not a person who is particularly praiseworthy or blameworthy for being rich, but that person's wealth satisfies an incredibly high standard (she is a billionaire and not merely a millionaire). One might, however, argue that there is a sense in which 'rich' is indeed an evaluative term. A philosophical paper might be (epistemically) better for being rich, and an argument might be very poor. In those contexts, 'rich' and 'poor' are not the endpoints of a wealthiness scale but are likely to express thin evaluations synonymous to 'good' and 'bad.' If 'rich' is indeed polysemous in this way, then it is not the empirical method that has its shortcomings, but our pretheoretic intuitions.[18] Third, language is complex and full of standardized phrases and idioms. Thus, the high values for 'disgusting' might simply reflect that 'truly disgusting' is a popular phrase and not that the term 'disgusting' is evaluatively powerful.

We would like to close by applying the binary logistic regression model – even if only to a few terms. In the introduction, we mentioned several terms for which disagreement looms large with regard to whether they are indeed evaluative concepts. Those included AFRAID, CONSERVATIVE, CONSTITUTIONAL, DIRTY, HAPPY, LEWD, LEGAL, LIBERAL, and RELIGIOUS. Table 3 lists their sentiment values (from SentiWords), their *eval* number, and the probability of belonging to the class of evaluative concepts. As can be seen from the table, most of the analyzed terms received pretty low *eval* numbers, suggesting that they are not evaluative: legal and political concepts are likely merely value-associated, not evaluative. Emotion concepts, on the other hand, received higher ratings. This is in line with recent empirical research finding

[18]We would like to thank an anonymous reviewer of this journal for bringing our attention to this possibility.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Table 3. *Eval* values (based on COCA) for disputed concepts and predicted probabilities for formerly unobserved adjectives based on the logistic regression model. If the predicted probability is >0.5, the adjectives would be considered evaluative.

Adjective	Sentiment value	Eval	Prob. for evaluative class	Standard error
afraid	−0.66	0.71	0.76	0.15
conservative	−0.13	0.45	0.30	0.13
constitutional	0.18	0.02	0.02	0.02
dirty	−0.10	0.42	0.26	0.12
happy	0.85	2.26	1.00	0.00
lewd	−0.39	0.25	0.08	0.07
(il-)legal	0.01	0.04	0.02	0.02
liberal	0.38	0.27	0.10	0.08
religious	0.04	0.19	0.06	0.05

that normative considerations have a strong impact on the applications of emotion terms (Díaz & Reuter, 2020; Phillips et al., 2017).

In sum: We have derived stable and plausible results using data that reveal how the 'truly' and 'really' modifiers work. In this paper, we believe we have sketched a clear path for making a Carnapian transition (see Carnap, 1950) for thick adjectives specifically, and evaluative concepts more generally.

Supplementary material. The supplementary material for this article (including the R script we used for the data analyses) can be found at http://doi.org/10.1017/langcog.2023.35. The corpus data from COCA is available through their website https://www.english-corpora.org/coca/. The data we used for analyzing data from Reddit can be made available on request.

Acknowledgments. We would like to thank Ethan Landes, Isidora Stojanovic, as well as the participants at workshops and conferences in Prague, Salamanca, Zurich, at Kloster Kappel, and at the Online Corpus Analysis Workshop for their helpful comments on previous versions of the manuscript. We are especially grateful to two anonymous reviewers for their excellent comments.

Funding statement. The research of Lucien Baumgartner 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.

References

Alfano, M. (2018). Digital humanities for history of philosophy: A case study on Nietzsche. In T. Neilson, I. Ievenberg, & D. Rheams (Eds.), *Research methods for the digital humanities* (pp. 85–101). Palgrave Macmillan.

Alfano, M., Higgins, A., & Levernier, J. (2018). Identifying virtues and values through obituary data-mining. *The Journal of Value Inquiry*, 52, 59–79.

Andow, J. (2015). How "intuition" exploded. *Metaphilosophy*, 46, 189–212.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J., & Io, P. (2020). The pushshift Reddit dataset. Technical report.

Baumgartner, L., Willemsen, P., & Reuter, K. (2022). *The polarity effect of evaluative language* (pp. 1–18). Philosophical Psychology.

Benamara, F., Chardon, B., Mathieu, Y., Popescu, V., & Asher, N. (2012). How do negation and modality impact on opinions? In *Proceedings of the ACL-2012 workshop on extra-propositional aspects of meaning in computational linguistics (ExProM-2012)*, Jeju, Republic of Korea: Association for Computational Linguistics. (pp. 10–18).

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Bjerva, J. (2014). Multi-class animacy classification with semantic features. In *Proceedings of the student research workshop at the 14th conference of the European chapter of the Association for Computational Linguistics*, Gothenburg, Sweden: Association for Computational Linguistics, (pp. 65–75).

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

Bordonaba, D. (n.d.) Dual character concepts, concrete features and abstract values: A corpus-based study of the internal structure of SCIENTIST.

Bowman, S. R., & Chopra, H. (2012). Automatic animacy classification. In *Proceedings of the NAACL HLT 2012 student research workshop*, Montreal, Canada: Association for Computational Linguistics (pp. 7–10).

Carnap, R. (1950). *Logical foundations of probability* (2nd edition 1962). The University of Chicago Press.

Cepollaro, B. (2018). Negative or positive? Three theories of evaluation reversal. *Croatian Journal of Philosophy*, 18, 363–374.

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

Chappell, S.-G. (2013). There are no thin concepts. In S. Kirchin (Ed.), *Thick concepts* (pp. 182–196). Oxford University Press.

Chartrand, L. (2022). Modeling and corpus methods in experimental philosophy. *Philosophy Compass*, 17, e12837.

Curry, O., Chesters, M., & Van Lissa, C. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, 78, 106–124.

Dancy, J. (1995). In defense of thick concepts. *Midwest Studies in Philosophy*, XX, 263–279.

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

Díaz, R., & Reuter, K. (2020). Feeling the right way: Normative influences on people's use of emotion concepts. *Mind & Language*, 36, 451–470.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., & Stewart, B. M. (2018). How to make causal inferences using texts. *Science Advances*, 8, 1–13.

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

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

Enoch, D., & Toh, K. (2013). Legal as a thick concept. In W. Waluchow & S. Sciaraffa (Eds.), *Philosophical Foundations of the Nature of Law* (pp. 256–278). Oxford University Press.

Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of the fifth international conference on language resources and evaluation*, Genoa, Italy: European Language Resources Association (ELRA). (pp. 417–422).

Fischer, E., Engelhardt, P. E., & Herbelot, A. (2015). Intuitions and illusions: From explanation and experiment to assessment. In E. Fischer & J. Collins (Eds.), *Experimental philosophy, rationalism, and naturalism* (pp. 259–292). Routledge.

Gatti, L., Guerini, M., & Turchi, M. (2016). SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7, 409–421.

Gibbard, A. (1992). Thick concepts and warrant for feelings. *Proceedings of the Aristotelian Society, Supplementary*, 66, 267–283.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, 316, 998–1002.

Hansen, N., Porter, J. D., & Francis, K. (2021). A corpus study of "know": On the verification of philosophers' frequency claims about language. *Episteme*, 18, 242–268.

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

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

Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Association for Computational Linguistics (ACL)*, (pp. 174–181).

Jahan, L., Chauhan, G., & Finlayson, M. A. (2018). A new approach to animacy detection. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1–12).

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

Kirchin, S. (2013). Introduction: Thick and thin concepts. In S. Kirchin (Ed.), *Thick concepts* (pp. 1–19). Oxford Academic.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, 63, 190–194.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition*, 127, 242–257.
Kotzee, B., & Wanderer, J. (2008). Introduction: A thicker epistemology? *Philosophical Papers*, 37, 337–343.
Kyle, B. G. (2013). Knowledge as a thick concept: Explaining why the Gettier problem arises. *Philosophical Studies*, 165, 1–27.
Leslie, S.-J. (2015). Hillary Clinton is the only man in the Obama administration": Dual character concepts, generics, and gender. *Analytic Philosophy*, 56, 111–141.
Liu, D., & Espino, M. (2012). Actually, genuinely, really, and truly: A corpus-based behavioral profile study of near-synonymous adverbs. *International Journal of Corpus Linguistics*, 17, 198–228.
Mizrahi, M. (2020). The case study method in philosophy of science: An empirical study. *Perspectives on Science*, 28, 63–88.
Mohammad, S. M. (2016). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201–237). Woodhead Publishing.
Napolitano, M. G., & Reuter, K. (2021). What is a conspiracy theory? *Erkenntnis*, 88, 2035–2062.
Nichols, S., & Pinillos, N. Á. (2018). Skepticism and the acquisition of "knowledge". *Mind & Language*, 33, 397–414.
Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.
Payne, A. (2005). A new account of thick concepts. *The Journal of Value Inquiry*, 39, 89–103.
Phillips, J., De Freitas, J., Mott, C., Gruber, J., & Knobe, J. (2017). True happiness: The role of morality in the folk concept of happiness. *Journal of Experimental Psychology: General*, 146, 165–181.
Reuter, K. (2011). Distinguishing the appearance from the reality of pain. *Journal of Consciousness Studies*, 18, 94–109.
Reuter, K. (2019). Dual character concepts. *Philosophy Compass* 14, e12557.
Reuter, K., & Baumgartner, L. (n.d.a). In M. Bauer & S. Kornmesser (Eds.), *Corpus analysis: A case study on the use of 'conspiracy theory'*. Experimental Philosophy for Beginners.
Reuter, K., & Baumgartner, L. (n.d.b). Conspiracy theories are not theories: Time to rename conspiracy theories.
Reuter, K., Löschke, J., & Betzler, M. (2020). What is a colleague? The descriptive and normative dimension of a dual character concept. *Philosophical Psychology*, 33, 997–1017.
Richard, M. (2008). *When truth gives out*. Oxford University Press.
Roberts, D. (2013). Thick concepts. *Philosophy Compass*, 8, 677–688.
Roberts, D. (2018). Thick epistemic concepts. In C. McHugh, J. Way, & D. Whiting (Eds.), *Metaepistemology* (pp. 159–178). Oxford University Press.
Simon-Vandenbergen, A.-M., & Taverniers, M. (2014). The adverb truly in present-day English. In M. de los Ángeles Gómez González, F.J. Ruiz de Mendoza Ibáñez, F. Gonzálvez-García, & A. Downing (Eds.), *The functional perspective on language and discourse* (pp. 169–186). Benjamins.
Smith, M. (2013). On the nature and significance of the distinction between thick and thin ethical concepts. In S. Kirchin (Ed.), *Thick concepts* (pp. 97–120). Oxford Academic.
Stojanovic, I., & Kaiser, E. (2022). Exploring valence in judgments of taste. In J. Wyatt, J. Zakkou, & D. Zeman (Eds.), *Perspectives on taste aesthetics, language, metaphysics, and experimental philosophy. Stojanovic and Kaiser-London and* (pp. 231–259). Routledge.
Stojanovic, I., & McNally, L. (forthcoming). Are moral predicates subjective? A corpus study. In D. Bordonabo (Ed.), *Experimental philosophy of language: Perspectives, methods and prospects*. Springer.
Sytsma, J., Bluhm, R., Willemsen, P., & Reuter, K. (2019). Causal attributions and corpus analysis. In E. Fischer & M. Curtis (Eds.), *Methodological advances in experimental philosophy* (pp. 209–238). Bloomsbury Academic.
Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2, 325–347.
Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37, 267–307.
Topham, E. (2016) Thick and thin concepts in law. Doctor of Philosophy (PhD) thesis, University of Kent, (KAR id:69464).
Väyrynen, P. (2008). Slim epistemology with a thick skin. *Philosophical Papers*, 37, 389–412.
Väyrynen, P. (2011). Thick concepts and variability. *Philosophers' Imprint*, 11, 1–17.

https://doi.org/10.1017/langcog.2023.35 Published online by Cambridge University Press

Väyrynen, P. (2013). *The Lewd, the Rude and the Nasty*. Oxford University Press.
Väyrynen, P. (2021). Thick ethical concepts. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2021 edition).
Willemsen, P., Baumgartner, L., Cepollaro, B., & Reuter, K. (n.d.) Evaluative Deflation, Social Expectations and the Zone of Moral Indifference.
Willemsen, P., & Reuter, K. (2021). Separating the evaluative from the descriptive: An empirical study of thick concepts. *Thought: A Journal of Philosophy*, 10, 135–146.
Williams, B. (1985). *Ethics and the limits of philosophy*. Harvard University Press.
Zangwill, N. (2013). Moral metaphor and thick concepts: What moral philosophy can learn from aesthetics. In S. Kirchin (Ed.), *Thick concepts* (pp. 197–209). Oxford Academic.

Cite this article: Reuter, K., Baumgartner, L., & Willemsen, P. (2024). Tracing thick and thin concepts through corpora, *Language and Cognition* 16: 263–282. https://doi.org/10.1017/langcog.2023.35