Tracing Thick and Thin Concepts Through Corpora

Kevin Reuter[*,§], Lucien Baumgartner[*], Pascale Willemsen[*]

Unpublished Manuscript

[*] *University of Zurich*
[§] *To whom correspondence should be addressed: kevin.reuter@uzh.ch*
This research was supported by the Swiss National Science Foundation (100012_169484)

Abstract

Philosophers and linguists currently lack the means to reliably identify evaluative concepts and to measure their evaluative intensity. Using a corpus-based approach, we present a new method to distinguish evaluatively thick adjectives like 'courageous' from descriptive adjectives like 'narrow', and from value-associated adjectives like 'sunny'. Our study reveals 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 the operationalization we suggest may pave the way for a more quantitative approach to the study of thick and thin concepts.

Keywords: Thick Concepts; Modifiers; Truly; Evaluation; Sentiment; Corpus Studies

1 Introduction

The two most prominent kinds of evaluative concepts are thin and thick concepts. Thin terms like 'great' and 'terrible' evaluate without specifying the descriptive aspects that ground their evaluation. Thick terms, such as 'rude' and 'courageous', describe and evaluate at the same time. For instance, by calling a person courageous, we not only evaluate her positively, we also describe the person 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.

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 terms apart from terms such as 'homeless', that are value-associated but do not evaluate in the sense of expressing approval or disapproval of someone or something. Terms 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 illustrate the difference between evaluation and arousal/affect with two examples: Terms 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, words like 'generous', 'insane', and 'ugly', evaluate a person, behavior, or object.[1]

Scholars working in ethics, aesthetics or epistemology usually do not care very much about arousal and affect. Instead, they rather focus their studies on the *evaluative* aspects of honesty, beauty and justification. It is therefore (or rather should be) a central endeavour to identify evaluative terms out of the large group of terms that trigger arousal and affect. Rather surprisingly, no tool for measuring the evaluative intensity of an evaluative concept has so far been developed. Instead, scholars rely 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., Roberts, 2013; Väyrynen, 2013).

It is certainly helpful to have a list of examples of evaluative concepts—this list contains terms 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

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

of a term 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 terms: there is no a priori reason to suppose that all thick terms behave similarly in regards 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.

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 to 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., Kyle, 2013; Kotzee & Wanderer, 2008; 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) and *intention* (Knobe 2003) that play a central role 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, see e.g. Alfano et al., 2018; Cepollaro & Stojanovic, 2016; Cepollaro, 2018; Eklund, 2011; Väyrynen, 2011; and Willemsen & Reuter, 2020).

It seems to us that the lack of consensus in this area is, at least partially, a matter of methodology. If we are right, we are confronted with the task of developing a suitable method to classify thick and thin concepts as well as measure their evaluative intensity. To this end, we present a method that provides us with a way of operationalizing thick and thin terms, and hence, will serve as a first pass at identifying evaluative terms empirically.

Readers familiar with lexical sentiment analysis might wonder whether we are not 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.[2] 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 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 and McKeown (1997), Osgood et al. (1957) of the text data, as well as its (iii) 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, '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. 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

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

our research, and the likely success of quantifying the study of thick concepts in the General Discussion (Section 3).

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.[3] In this paper, we focus on thick and thin *adjectives*, which undoubtedly have received the greatest attention of all types of thick terms.

Our approach takes inspiration from recent research on dual character concepts (Knobe et al., 2013; Leslie, 2015; Del Pinal & Reuter, 2017; Reuter, 2019). Dual character concepts are concepts that are often, perhaps mostly, used descriptively but also encode an independent normative dimension. 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 Knobe et al. (2013) and Del Pinal and Reuter (2017).

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 can be used to stress the normative dimension of dual character concepts, as in ‘true mechanic’ or ‘true scientist’, the modifier ‘truly’ seems to intensify the evaluative aspect of thick and thin adjectives, as in ‘truly courageous’ and ‘truly awful’.

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

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).[4]

Although ‘truly’ seems to intensify the evaluative aspect of thick (as well as thin) adjectives, the 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 ‘too courageous’, and ‘not enough’ as in ‘not rude enough’ have been intensely debated as 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 America 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) 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 (2012, 198).

While ‘actually’ is rarely used to modify adjectives (ibid, 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 (212) argue that (a) ‘truly’ is more formal than ‘really’ (ibid, 210), (b) ‘really’, and ‘truly’ are often used to modify evaluative adjectives (ibid, 216), and (c) ‘really’ is the most versatile modifier (ibid, 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

[4] 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 form like ‘artist’, and ‘scientist’. Consequently, the truly modifier cannot be applied as straightforwardly to examine dual character concepts.

terms. Importantly, our claim is not that the modifiers 'truly' and 'really' cannot be reasonable applied to highlight other aspects of the adjective they modify, but rather that those modifiers are frequently used to highlight the evaluative content of adjectives, such that patterns of use emerge which reveal differences between the concept classes at stake.[5]

2.1 Stimuli & 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 non-moral thick adjectives: 5 positive (beautiful, delicious, funny, justified, wise) and 5 negative (boring, disgusting, insane, stupid, ugly)[6]
- 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 5000 hits on the Corpus of Contemporary American English (COCA) ('courageous' being the only exception, with 4742 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 had a very similar average absolute sentiment value of 0.56 (SD = 0.20). It is thus unlikely for the sentiment

[5] 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.

[6] 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 applicable not only in one of those domains.

values to have had any confounding effect on our studies. Purely descriptive adjectives have 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. 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 end empirically verified, some reliance on intuitions was unavoidable. That said, some of the analyses we have done dispense with any intuition-based pre-categorization. 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, a comparison between 'truly' and 'really' on the one side, as well as 'very' on the other side, should allow us to determine whether any positive result cannot 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.[7] The value for *eval*—our variable for evaluative intensity—was then calculated by taking the average of

[7] 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.

both normalized ratios. 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. We first queried for 200 instances of ‘truly x’. The query was conducted backwards from t_1 to t_2, where t_1 was 31.08.2020 and t_2 was the date on which we reached the 200th instance of ‘truly x’. In a second step, we queried for ‘really x’ for the time period determined by the truly-query, that is, t_1 to t_2. Finally, we did the same for the adjective without modifiers. Each of the ratios were then standardized as follows: for 200 ‘truly x’, we have *n* ‘really x’ and *m* ‘x’.

2.2 Results

Table 1 below 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.[8] There were also significant differences between thick moral terms, thick non-moral terms, and thin terms. Most thin terms had higher *eval* numbers than moral thick terms, while non-moral thick terms being mostly positioned between thin and moral thick 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’ mod-

[8] 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.

ifier as thin and thick terms like ‘bad’, ‘honest’, ‘rude’, ‘boring’, beautiful’.[9]

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	couragous	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 non-moral	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
$\overline{thick-non-moral}$		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

Table 1: *eval* values and ratios per million 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 1000 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.

[9] A Pearson correlation test between the very-ratios and the combined truly- and really-ratios shows that they do not correlate significantly (t(43) = 0.18128, p = 0.857), on 0.05-alpha level. This indicates that ‘very’ is indeed used differently and does not allow analogous inferences.

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 very similar values (see Table 2 below). With the exception of ‘honest’ (and to a lower extent ‘generous’), the ratios seem very robust. The differences for ‘honest’ might be explained by the rather heavy representation of the news media in COCA; questions of honesty might be more salient in these sources than in everyday talk.

target	really	truly	% really	% truly	Eval	Eval from COCA
bad	21806	200	2.86	0.026	2.89	2.18
empty	1329	200	0.20	0.030	0.23	0.16
generous	5372	200	1.65	0.06	1.71	0.63
honest	706	200	0.24	0.07	0.31	1.15
short	48756	200	0.17	0.00	0.17	0.16
stupid	6350	200	1.63	0.05	1.69	1.63

Table 2: Ratios and *eval* values for 6 adjectives using data from Reddit.

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 results are displayed in the form of a tree diagram (Figure 1). The cluster analysis yielded three main clusters. 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’; colored in light-grey in Figure 1 below). Most terms in the left evaluative cluster are either thin terms and non-moral thick terms. Not a single thin term was assigned to the right evaluative cluster.

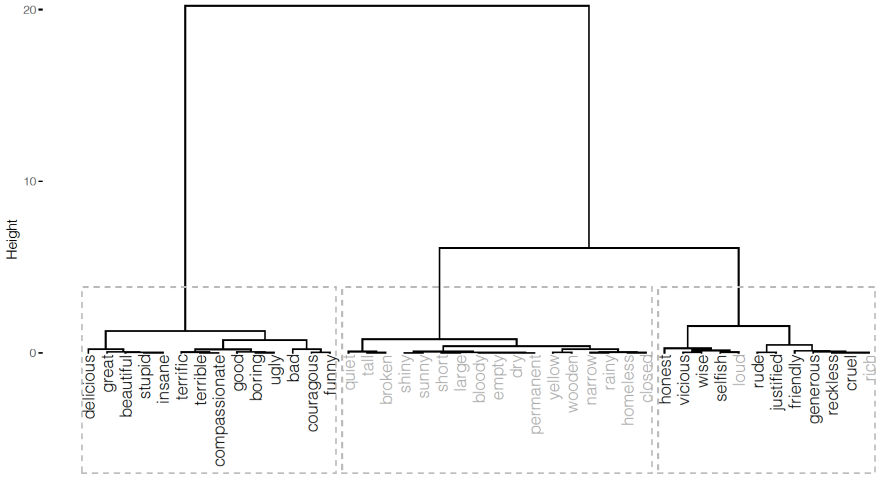

Figure 1: Tree diagram displaying the clusters using hierarchical cluster analysis. Three main clusters emerge, two of which (left and right) feature almost only evaluative terms, and one cluster (middle) which only contains non-evaluative adjectives. Terms in light-grey are the selected non-evaluative adjectives, whereas the ones in black are the evaluative adjectives.

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. As can be seen in Figure 2, the predicted probabilities for the different classes exhibit different progression patterns along the average of normalized ratios for truly and really.[10]

Interestingly, the accuracy by class for value-associated concepts (74.4%) is relatively high, similar to thick non-moral (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 thick non-moral concepts, and descriptive concepts get misclassified (66.6%) as value-associated concepts.

[10] 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. Cohen's Kappa is moderate with 40.7%, which is particularly interesting as we have a slight imbalance in the classes.

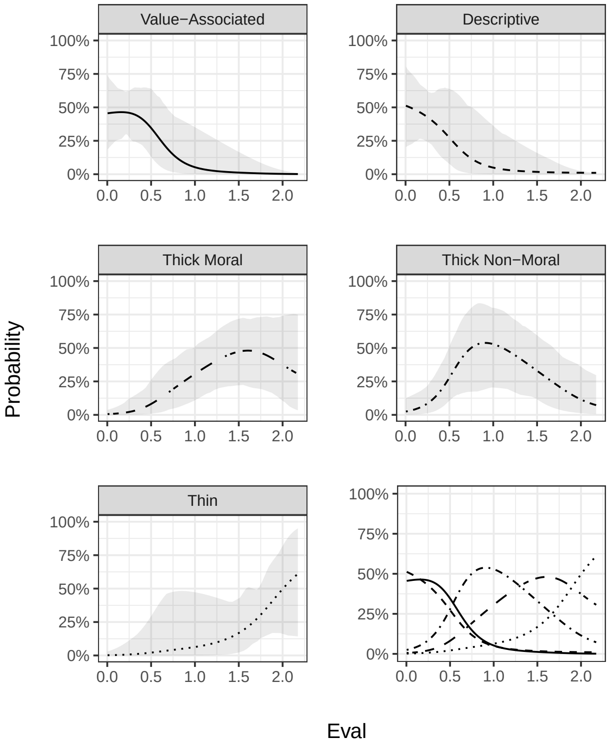

Figure 2: Predicted probabilities for class membership. The first five charts are single plots for the 5 pre-defined categories, the sixth is a combined plot without confidence intervals. The x-axis indicates the *eval* number.

In this paper, we are primarily interested in the binary classification of inherently evaluative (thin and thick concepts) and non-evaluative concepts (value-associated and descriptive concepts). The multinominal model above had its difficulties distinguishing thin concepts from non-moral thick concept. In the binary approach we no longer need to discriminate between the two, since both are evaluative concept classes. Figure 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 that is greater than 0.75 are quite likely to be thick or thin adjectives. 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: non-eval.), and ‘loud’ (true: non-eval.). While the classification of ‘cruel’ as non-evaluative 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.

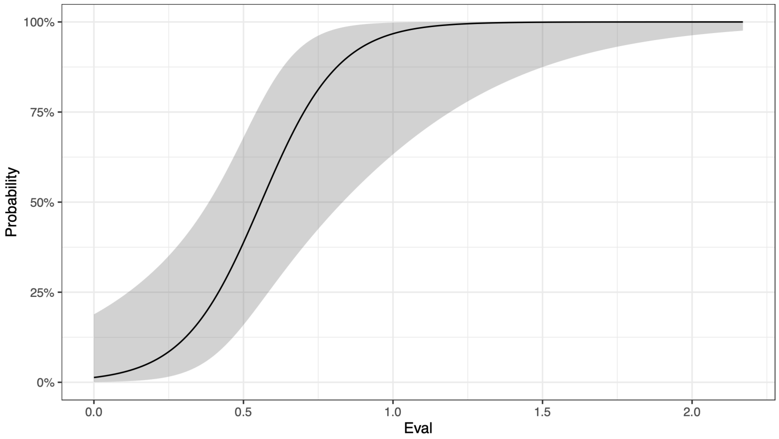

Figure 3: Predicted probability for being an evaluative concept, with confidence intervals.

2.3 Discussion

Inspired by recent research on measuring the evaluative component of dual character concepts, we examined the use of the 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 utilized to identify evaluative adjectives (both thin and thick) and to 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 non-moral 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.

Most thick *non-moral* terms had higher *eval* numbers than the investigated thick *moral* terms: thick non-moral adjectives have rates comparable to those of thin adjectives (i.e., similarly high).[11] This might be explained by the fact that some of the thick non-moral 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 (Väyrynen, 2008; Kirchin 2013; Zangwill, 2013), for instance, argues for the thinness of the term 'beautiful' because descriptively richer terms like 'elegant', 'delicate', 'balanced', etc. are merely "*ways*—ways of being beautiful." (2013, 317). Chappell, 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" (2013, 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 non-moral terms we investigated are thinner than moral thick terms.

3 General Discussion

Empirical work on thick concepts is in its infancy. Research on thick concepts has been mostly theoretical (but see Reuter, Löschke, & Betzler, 2020, and 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 linguistic intuitions of a small group of individuals. Such overreliance on individual intuitions places severe limitations on current projects on thick 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 measuring the extent to which thick concepts are used evaluatively. We recorded the frequencies with

[11] This explains the low accuracy for classifying thin concept, as thin adjectives were mostly misclassified (80%) as non-moral adjectives.

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 *non-moral* concepts.

- Descriptive and value-associated adjectives are not used less frequently with the modifier ‘very’ compared to evaluative adjectives.

Philosophers often assume that thick concepts form a unique class 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 provides evidence that thick concepts are indeed evaluative and that this evaluative component 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 their kind that thick terms 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 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, etc., behave similarly is beyond the scope of this paper. 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. Second, 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 of 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:

1. We have motivated the operationalization of evaluative intensity through *eval* independently of the results we collected.

2. A cluster analysis demonstrated that *eval* allows us to match most pre-theoretic 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 non-evaluative concepts.

5. For thin concepts, *eval* values matched the semantic meanings of the terms: 'terrific' and 'awful' are more evaluative than 'good' and 'bad', and correspondingly 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 pre-theoretic intuitions about these terms (e.g., 'reckless' received the same value as 'rich'). Second, the intensifiers, especially the intensifier '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). 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 our 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 thick terms. Those included 'afraid', 'conservative', 'constitutional', 'dirty', 'happy', 'lewd', 'legal', 'liberal', and 'religious'. Table 3 below 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 terms are likely merely value-associated, not evaluative. Emotion terms, on the other hand, received high ratings. This is in line with recent empirical research finding that normative considerations have a strong impact on applications of emotion terms (Díaz & Reuter, 2020; Phillips et al., 2017).

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

Table 3: *Eval* values for disputed terms and predicted probabilities for formerly unobserved adjectives based on the logistic regression model. If the predicted probability is above 0.5, the adjectives would be considered evaluative.

In sum: We have derived stable and plausible results using data that reveals how the ‘truly’ and ‘really’ intensifiers work. In this paper, we believe we have sketched a clear path for making a Carnapian transition for thick adjectives, specifically, and evaluative concepts, more generally.

References

Alfano, M., Higgins, A, Levernier, J. (2018). Identifying Virtues and Values Through Obituary Data-Mining. *The Journal of Value Inquiry*, 52(1), pp. 59–79.

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

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)*, pp. 10–18.

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*, pp. 65–75.

Blackburn, S. (1992). Through Thick and Thin. *Proceedings of the Aristotelian Society, supplementary volume*, 66, pp. 284–99.

Bowman, S. R. & Chopra, H. (2012). Automatic Animacy Classification. In *Proceedings of the NAACL HLT 2012 Student Research Workshop*, pp. 7–10.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press; second edition 1962.

Cepollaro, B. (2018). Negative or Positive? Three Theories of Evaluation Reversal. *Croatian Journal of Philosophy*, 18(3), pp. 363–374.

Cepollaro, B., Stojanovic, I. (2016). Hybrid Evaluatives. *Grazer Philosophische Studien* 93, pp. 458-488.

Chappell, S.-G. (2013). There are no thick concepts. in S. Kirchin (ed.) *Thick Concepts*. Oxford University Press, pp. 182–196.

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, pp. 106-124.

Dancy, J. (1995). In Defense of Thick Concepts. *Midwest Studies in Philosophy*. XX: 263–79.

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*, pp. 451-470.

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2018). *How to Make Causal Inferences Using Texts*. Preprint available online: https://arxiv.org/pdf/1802.02163.pdf.

Eklund, M., (2011). What Are Thick Concepts? *Canadian Journal of Philosophy*, 41(1), pp. 25–49.

Elstein, D., Hurka, T. (2009). From Thick to Thin: Two Moral Reduction Plans. *Canadian Journal of Philosophy*, 39(4), pp. 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*. Oxford: Oxford University Press, pp. 257–278.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pp. 417–422.

Gatti, L., Guerini, M., and Turchi, M. (2016). SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis. *IEEE Transactions on Affective Computing*, 7(4):409-421.

Gibbard, A. (1992). Thick Concepts and Warrant for Feelings. *Proceedings of the Aristotelian Society, Supplementary Volume*, 66, pp. 267–83.

Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, 316(5827), 998-1002.

Hare, R. (1952). *The Language of Morals*. Oxford: Clarendon Press.

Hatzivassiloglou, V., McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. *Association for Computational Linguistics (ACL)*, pp. 174–181.

Jahan, L., Chauhan, G., and 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(4), pp. 1-28.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63(3), 190-194.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual Character Concepts and the Normative Dimension of Conceptual Representation. *Cognition*, 127(2), pp. 242-257. Kotzee, B., & Wanderer, J. (2008). Introduction: A Thicker Epistemology?. *Philosophical Papers*, 37(3), pp. 337-343.

Kyle, B. (2013). Knowledge as a Thick Concept: Explaining Why the Gettier Problem Arises. *Philosophical Studies*, 165(1), pp. 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(2), 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(2), pp. 198-228.

Mohammad, S. M. (2016). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. in H. L. Meiselman (ed.) *Emotion Measurement*. Duxford: Woodhead Publishing, pp. 201-237.

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.

Phillips, J., De Freitas, J., Mott, C., Gruber, J., and Knobe, J. (2017). True Happiness: The Role of Morality in the Folk Concept of Happiness. *J. Exp. Psychol. Gen.*, 146(2), pp. 165–181.

Reuter, K. (2019). Dual Character Concepts. *Philosophy Compass*, 14(1), e12557.

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), pp. 997-1017.

Roberts, D. (2013). Thick Concepts. *Philosophy Compass*, 8(8), pp. 677-688.

Roberts, D. (2018). Thick Epistemic Concepts, in Conor McHugh, in J. Way, and D. Whiting (eds.), *Metaepistemology*, Oxford: Oxford University Press, pp. 159–78.

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*, 209-238.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. *Annual Review of Linguistics*, 2, pp. 325-347.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), pp. 267–307.

Topham, E. (2016). Thick and Thin Concepts in Law. Doctor of Philosophy (PhD) thesis, University of Kent, (KAR id:69464).

Väyrynen, Pekka (2008). Slim Epistemology with a Thick Skin. *Philosophical Papers*, 37(3), pp. 389-412.

Väyrynen, P. (2011). Thick Concepts and Variability, *Philosophers' Imprint*, pp. 11(1).

Väyrynen, P. (2013). *The Lewd, the Rude and the Nasty*. New York: Oxford University Press.

Väyrynen, P. (2021). Thick Ethical Concepts. in E. Zalta (ed.). *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition).

Willemsen, P., & Reuter, K. (2021). Separating the Evaluative from the Descriptive: An Empirical Study of Thick Concepts. *Thought: A Journal of Philosophy*, 10(2), pp. 135-146.

Williams, B. (1985). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

Zangwill, N. (2013). Moral Metaphor and Thick Concepts: What Moral Philosophy Can Learn from Aesthetics. in S. Kirchin (ed.) *Thick Concepts*. Oxford University Press, p. 197 - 209.