Examining Evaluativity in Legal Discourse:

A Comparative Corpus-Linguistic Study of Thick Concepts

Pascale Willemsen, Lucien Baumgartner, Severin Frohofer, Kevin Reuter*

(Department of Philosophy, University of Zurich, Zuerichbergstrasse 43, CH-8044 Zurich)

** To whom correspondence should be addressed*

Preprint: To appear in Advances in Experimental Philosophy of Law,

edited by Stefan Magen & Karolina Prochownik

Abstract

How evaluative are legal texts? Do legal scholars and jurists speak a more descriptive or perhaps a more evaluative language? In this paper, we present the results of a corpus study in which we examined the use of evaluative language in both the legal domain as well as public discourse. For this purpose, we created two corpora. Our *legal professional corpus* is based on court opinions from the U.S. Courts of Appeals. We compared this professional corpus to a *public corpus,* which is based on blog discussions on the internet forum Reddit. While many linguistic phenomena can give insights into evaluativity, we investigated the use of a wide selection of evaluative adjectives (more specifically, thick adjectives) to gain a more comprehensive picture of the degree of evaluativity in the legal domain. Our analysis shows that legal professionals use thick terms less evaluatively than laypeople, which suggests that legal texts are less evaluative than ordinary discussions. This result, more generally, supports the philosophical idea that thick concepts may vary in their evaluative intensity.

1 Introduction

Legal professionals need to be objective in many respects. For instance, each defendant has a constitutional right to be given a fair trial, independent of any personal liking or disliking that the legal professionals involved might have for the defendant. This involves an objective, unbiased treatment of the available evidence, and, especially on the side of the defense, a fair representation of the defendant. Legal professionals must only follow the law and cannot allow their own norms and ideals affect their legal judgments. Whether or not legal professionals object to anyone's personal lifestyle and decisions must not influence their judgment as long as these issues are not in violation of the law. Because of this particularly high need for objectivity, one might suspect that the legal discourse is devoid of verbally expressed evaluations. The legal system is there to reveal the truth and, relatedly, legal processes should be characterized by a strictly regulated, objective, impersonal and unbiased adjudication to not distort the quested facts. Accordingly, so one might assume, this must also be reflected in the language of legal professionals.

However, consider the following case. In 1978, Ted Bundy was convicted of murder, attempted murder, and burglary and sentenced to death. In his statement, Judge Edward Cowart said:

> The court finds that of both these killings were indeed heinous, atrocious, and cruel, and that they were extremely wicked, shockingly evil, vile, and the product of a design to inflict a high degree of pain and utter indifference to human life.

While the statement does not appear to be particularly unusual and many people might sympathize with its general message, the statement demonstrates a very explicit display of, arguably, the judge's contempt of the defendant and, so one might think, reflects the judge's very personal opinion. Given a statement like this, we might wonder whether the legal system is that non-evaluative after all.

The Ted Bundy trial was special not just because of the seriousness of the allegations. It was also the first trial in which a dental impression was used as evidence. One of the victims showed a bite wound on her body, and it was argued by forensic odontologists that the particularities of Ted

Bundy's teeth matched the wound. The defense attorney John Henry Browne objected to the acceptability of this evidence by saying:

> The evidence in this case presents many reasonable doubts. It is a sad day for our system of justice that can put a man's life on the line because they say he has crooked teeth. How tragic it would be if a man's life were to be taken from him because 12 people *thought* that he was probably guilty, but they were *not sure.*

Should we take issue with this statement? If the legal system understands itself as a non-evaluative business in which personal approval and disapproval should have no place, shouldn't then phrases like 'a sad day for our system of justice', and 'how tragic it would be' make us rather uncomfortable?

Perhaps, you might say, we cherrypicked the Ted Bundy case in which Judge Cowart was a bit over the line, and the defense lawyer desperately tried to avoid the inevitable. Perhaps, you might think, evaluative statements in extreme cases like the Ted Bundy trial are hardly avoidable but are less common in more mundane cases.

If these points are correct, we should be able to find evidence that legal discourse is indeed more descriptive and less evaluative than public discourse. Surprisingly, little evidence has been collected in support of that view. Several studies have investigated evaluativity in legal texts (Heffer, 2007; Finegan, 2010; Mazzi, 2010; Goźdź-Roszkowski & Pontrandolfo, 2012). However, there are two reasons why these studies do not allow us to answer our main question directly. First, previous research has so far mostly addressed very particular ways of being evaluative and the linguistic phenomena under investigation are rather limited for drawing more general conclusions. Second, almost all previous studies lack a direct comparison between legal and some form of baseline or control discourse, such as the ordinary public discourse. As a consequence, it is very hard to determine how evaluative legal discourse is and if it is more or less evaluative than other discourses.

The aim of this paper is to provide an analysis of the evaluativity of legal discourse. However, in doing so, we need to avoid both these shortcomings. First, we compared legal texts with public discussions in terms of the evaluative extent and intensity of their contents. For this purpose, we created two corpora. Our *legal professional corpus* is based on court opinions from the U.S. Courts of Appeals; the *public corpus* is based on blog discussions on the internet forum Reddit and serves as our control discourse. Second, in order to reach a rather general and robust conclusion about the evaluativity of legal discourse, we focused on a linguistic phenomenon that is broad, frequent, and at the heart of evaluative judgments: the use of thick adjectives. While thin terms like ‘good’ and ‘bad’ merely express approval or disapproval, thick terms combine evaluative and descriptive content. For instance, ‘generous’ combines a positive evaluation with the descriptive feature ‘willingness to share limited resources with others’, and ‘rude’ combines a negative evaluation with the descriptive features ‘causing offense by violating rules of good manner’. Thin concepts are not ideal items to measure the evaluativity of a text because writers often try to avoid plain terms like ‘good’ and ‘bad’. Thick concepts, in contrast, are more subtle ways of being evaluative. In addition, thick concepts come in a much greater variety due to their descriptive components. While there is only a limited number of terms to express one’s approval or disapproval using a thin term, there are plenty of thick terms that all express (dis-)approval, but are different with respect to their descriptive elements. Consequently, thick concepts are the perfect items to examine the evaluative extent and intensity of legal contexts.

Here is how we will proceed. In Section 2, we review the current state of the literature and develop our own “thick concepts”-approach. Section 3 describes the generation of the legal and ordinary discourse corpus and the selection of suitable terms for the corpus analysis. Section 4 contains the actual corpus study which consists of a global descriptive analysis of the two corpora (Section 4.1) and hypothesis-driven inferential analyses (Sections 4.2). We discuss the limitations of our study, possible alternative interpretations and the philosophical importance of our findings in the General Discussion (Section 5).

2 Evaluative Language – What Should We be Looking for?

Evaluations play an important part in communicative interactions. As members of social groups, we are not merely interested in exchanging factual information with one another, but we also have various normative interests. We wish to establish normative standards for how to behave and try to reinforce or alter one another's behavior to meet these standards (e.g., Sripada, 2007; Schmidt & Tomasello, 2012; Rakoczy & Schmidt, 2012; Feldmann Hall, Son, & Heffner, 2018). Evaluative language contributes to this goal in a unique way (e.g., Hare, 1952, Stevenson, 1937, Williams, 1985).

Philosophers have been interested in evaluative language as part of a larger, metaethical project that aims to understand the meaning of ethical terms, such as 'good' or 'bad'. Also, in the last thirty years, a wider circle of cognitive scientists has become involved in studying evaluative language, including experimental and theoretical linguists, developmental and behavioral psychologists, and experimental philosophers (e.g., Cepollaro, Sulpizio, Bianchi, 2019; Del Pinal & Reuter, 2017; Reuter et al., 2020; Willemsen & Reuter, 2020, 2021). The identification of evaluativity is yet difficult, as evaluation can be realized in various ways. Bednarek (2008) offers an extensive overview of the different approaches to evaluativity, including eight different perspectives that are each connected to various cognitive scientific sub-disciplines.

Given the complexity of evaluation as a linguistic phenomenon and the variety of approaches one could take, examining evaluative language in the legal discourse requires serious pragmatic decisions on where and how to start. In this section, we lay out the decisions we made for this paper, and we further elaborate on why we believe that they provide a promising starting point.

2.1 Previous Empirical Studies

Several corpus studies have investigated evaluative language in the courtroom by focusing on different linguistic devices by which legal professionals express their stance. Some focus on linguistic patterns rather than particular words (e.g., Mazzi, 2010[1]; Goźdź-Roszkowski &

[1] Mazzi examines the most striking linguistic tools underlying judges' evaluative statements, including straightforwardly evaluative verbal and adjectival items (like 'disagree' and 'incorrect'), and he analyses the pattern 'this/these/that/those + labelling noun'. He finds that this pattern "is characterized by the occurrence of inherently evaluative elements as labelling nouns" (e.g., 'distortion', 'misapplication', 'omission', 'nonsense banner'), while the negative semantic prosody is predominant.

Pontrandolfo, 2012[2]). There are also several studies focusing on individual words (e.g., Heffer, 2007[3]; Finegan, 2010[4]; Goźdź-Roszkowski, 2018[5]). These studies strongly suggest that, thanks to numerous linguistic devices, legal speech, especially in court, is indeed evaluative – although often more subtly. Moreover, a few studies compare the legal discourse with ordinary language (e.g., Marín & Rea, 2014[6]; Wang & Yin, 2020[7]). These studies indicate that the legal discourse is less evaluative than ordinary discourses.

Two features of these studies stand out. First, the empirical evidence is generated by analyzing linguistic corpora consisting of legal texts and by applying corpus-linguistic tools to detect evaluations. Given the insights this method has already provided, we aim to extend this method further. Second, while these studies do provide fascinating insights into the evaluativity of legal language, only very few compare the legal corpus to some form of baseline corpus, thus making it difficult to evaluate how evaluative the legal discourse actually is. Such a comparison is one of the main goals we set ourselves in this paper.

[2] Goźdź-Roszkowski & Pontrandolfo's research focuses on the pattern 'noun + that' (e.g., 'fact that'). They find that certain nouns tend to entail negative polarity in their collocational environment (e.g., 'fact that'), while others are used primarily with terms of positive polarity (e.g., 'view that').

[3] Heffer finds that "the figures [of evaluative terms] on the whole are comparatively low, as would be expected of genres where explicit construal of judgement is proscribed" (Heffer, 2007, p. 159). He also reveals that the intensity of the inscription of judgment in sentencing seems to match the severity of the respective crime.

[4] Finegan investigates the use of stance adverbials, such as 'properly', 'improperly', 'appropriately', and 'correctly'.

[5] Goźdź-Roszkowski argues that 'liberty' and 'dignity' are keywords in the majority and dissenting opinions of two landmark civil rights cases concerning same-sex marriage given by the Supreme Court of the United States. Goźdź-Roszkowski sees these keywords as manifestations of a superordinate (ethical) value – i.e., respect –, towards which the judges' argumentation is orientated; and Goźdź-Roszkowski highlights the "central importance" of the related evaluative language for legal argumentation.

[6] Marín and Rea examine so-called sub-technical terms of the legal discourse, i.e., terms that are shared by the general and the legal field. They either denote the same (legal) concept in both fields (like 'judge' or 'court') or have a special meaning in the legal field that differs from its use in everyday language (like 'trial' or 'relief'). Marín and Rea provide both qualitative and corpus-based approaches to the process of specialization of such sub-technical legal terms.

[7] Wang and Yin compare the top 50 keywords of a legislative Chinese corpus and a general Chinese corpus respectively and assign them to different semantic categories to work out the linguistic features of legislative Chinese. They find that politics and economy are the predominant semantic categories and that the keywords "show strong professionalism" (p. 651).

Going beyond the extant empirical studies, we believe that other linguistic devices might be able to shed light on how evaluative the legal discourse is as well. We suggest the investigation of what philosophers have called thick concepts and, more specifically, the use of thick adjectives.

2.2 Thick Concepts

Thick concepts, so philosophers argue, are special in expressing evaluative and descriptive content at the same time. Although different kinds of thick concepts are discussed in the literature, such as ethical, epistemic, aesthetic thick concepts, most attention has been given to thick *ethical* concepts (see, e.g., Eklund, 2011; Roberts, 2013; Väyrynen, 2021). Typical examples are virtue concepts, such as *rude*, *friendly*, *cruel*, *compassionate*, and, as Williams (1985, p. 144) claims, "*treachery* and *promise* and *brutality* and *courage*"[8].

Some legal scholars have already highlighted the importance of thick concepts for various debates in the legal domain as well. Heidi Li Feldman (1997), for example, in discussing Williams (1995), presupposes the existence of thick *legal* concepts and stresses their significance for common-law reasoning. She claims that not only can the philosophical debate about thick concepts shed light on the mechanisms of legal language, but, conversely, philosophy can "learn more about the nature and workings of thick concepts" (p. 180) by paying closer attention to the way "judges and lawyers apply, deploy, manipulate, exploit, and engineer" (p.185) thick terms. David Enoch and Kevin Toh (2013, p. 264) seem to agree with Feldman when arguing that "many of the crucial legal concepts that our legal judgments deploy are thick concepts". However, Enoch and Toh do not provide an exemplary list of legal concepts they believe to be thick. Instead, they focus on the notion of legality itself and elaborate on why 'legal' communicates both descriptive and evaluative content. In a recent study, Flanagan and Hannikainen (2020) provide empirical data on the evaluative dimension of the folk concept of law. Flanagan and Hannikainen find that a rule's legal

[8] For an introduction to and discussion of thick ethical concepts, see Väyrynen, 2021.

character depends heavily on whether the rule is in alignment with what is considered morally acceptable and that “wickedness diminishes lawfulness” (2020, p. 11).

Chris Heffer (2007) has provided more extensive empirical evidence on the role that thick concepts play in legal discourse – even though he does not use the label ‘thick concepts’ and does not link his findings to the philosophical debate. He finds that in the legal discourse, evaluative judgments are often conveyed by “the use of attitudinal lexis, particularly adjectival epithets (*normal*, *capable*, *reliable*), but also through attitudinal nouns (*liar*, *thief*, *saint*) and verbs (*lie*, *steal*)” (Heffer, 2007, p. 154). With his research, Heffer has identified an excellent starting point for a systematic investigation of evaluative language in legal discourse. We aim to build on this research and connect it more systematically with the philosophical debate on thick concepts.

Thick ethical concepts strike us as particularly relevant in the legal domain. First, the source of the criminal system is the moral convictions of the people within the criminal legal system (e.g., Hart 1963; Devlin 1965; for a discussion see Edwards, 2021). Second, conversely, the criminal system can also change our moral views, “such that neglected values come to be taken seriously by community members” (e.g., Green, 2013). Thus, while certainly not all moral matters are also legally relevant, we should expect at least a significant overlap in the terms that are used.

In addition to thick ethical concepts, thick *epistemic* concepts[9] are highly relevant for legal arguments. One of the central features of the law is to determine whether a crime has been committed. This involves, among other things, the evaluation of whether a piece of evidence can be legally ‘admissible’ in a trial (as the dental imprint in the Bundy trial), whether it can ‘prove’ the defendant’s guilt or innocence ‘beyond reasonable doubt’, whether a witness’s testimony is ‘credible’ or ‘trustworthy’, and whether or not an action was ‘justified’ or ‘careless’.

[9] For a discussion of thick epistemic concepts, see Väyrynen, 2021, and the special issue in Philosophical Papers, edited by Kotzee and Wanderer, 2008.

Given the importance of thick ethical, legal, as well as epistemic concepts, we decided to investigate the evaluativity of legal discourse by studying the use of a wide selection of thick ethical, epistemic, and legal adjectives.[10]

2.3 Conjunctions of Adjectives and Sentiment Analysis

Having selected thick adjectives as our linguistic phenomena to study evaluative language, we still lack a way to determine how legal professionals and ordinary people use these terms. Simply counting the number of thick terms in both corpora will, of course, not do. We are not interested in how often people use thick terms, but instead we want to know how evaluative these terms are *when* they are used.

We decided to investigate the extent to which thick terms are used evaluatively by looking at those occurrences in which thick terms are conjoined with adjectives through the connective 'and'. The connective 'and' is a simple means to connect two adjectives, and its use is restricted in that the two conjoined adjectives usually share the same polarity (Elhadad & McKeown 1990; Hatzivassiloglou & McKeown 1997). Thus, while 'dishonest and unfair' seems to be a standard way to use the connective 'and', 'dishonest and fair' seems to be very unusual and requiring a specific context, as we would expect a person to rather say 'dishonest *but* fair' to highlight the different polarities that are conjoined together.[11]

Given the way the 'and' connective works, we propose a rather simple operationalization to measure the extent to which thick adjectives are used evaluatively. If a thick term like 'unfair' is conjoined with another evaluative term like 'deplorable' or 'rude', we can infer that the term is used

[10] Thick concepts can be nouns (such as 'liar', 'traitor', 'honesty', or 'brutality'), adverbs ('shamelessly', 'rudely', or 'bravely'), verbs ('to lie', 'to torture', 'to care'), and adjectives ('cruel', 'friendly', etc.). While all of these word classes would be interesting to investigate, adjectives provide the most straightforward starting point. First, philosophers usually discuss thick nouns or adjectives, with verbs and adverbs hardly ever being mentioned. Second, the use of thick nouns seems rather uncommon in ordinary conversations – we hardly ever speak of an 'act of kindness' or 'cruelty being inflicted on a person', but we would rather express the same thought by using the corresponding adjective.

[11] Focusing on conjunctions allows us to hold pragmatic effects constant (i.e., to connective-based effects), and, as a consequence, to use sentiment propagation patterns within connectives as a metric for evaluativeness.

evaluatively. In contrast, if 'unfair' is conjoined with more descriptive adjectives like 'difficult' or 'ambiguous', then the term is also used more descriptively. Such an interpretation of the use of thick terms is not deductively valid but an inference to the best explanation. In the General Discussion (Section 5) we will discuss other possible interpretations of our data. Importantly, we cannot make any inferences from a single or a small number of occurrences. Thus, we need to analyze a huge number of uses of a term like 'unfair' to determine whether that term is used more descriptively or more strongly evaluatively. In other words, we need large corpora.

An important question remains: How do we *measure* the evaluative intensity of the conjoined adjective like 'rude' or 'ambiguous' to determine the way the target adjective like 'unfair' is used? Due to the absence of a good metric of evaluativeness, we worked with a proxy, namely *sentiment values*. Sentiment dictionaries, such as SentiWords encode both the polarity (positive vs. negative) as well as the intensity for an enormous number of adjectives. These sentiment values range from '-1' meaning highly negative to '+1' meaning highly positive. In order to give our readers a better "feel" for the sentiment values that are attributed to terms in SentiWords, Figure 1 shows some exemplary adjectives for various different sentiment scores.

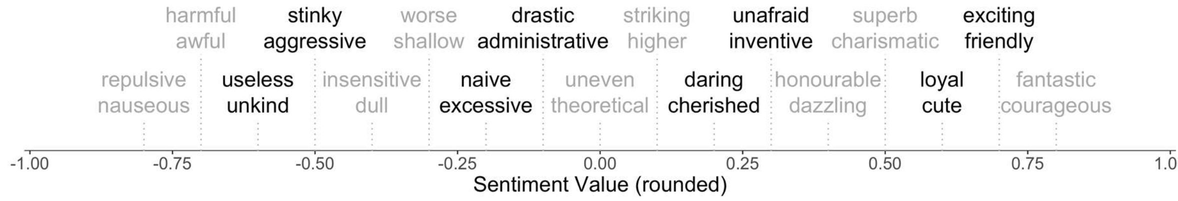

Figure 1: Sentiment values of various adjectives from the dictionary SentiWords.

Unfortunately, sentiment values not only represent the level of evaluativeness but also encode other aspects like subjectivity, emotion, value-association (for discussions of what sentiment represents in lexical sentiment analysis see, e.g., Benamara et al., 2012; Mohammad, 2020; Taboada et al., 2011). However, recording the sentiment values of hundreds of conjoined adjectives and subsequently calculating the average of those sentiment values, gives us a reliable picture of the way the target adjective, e.g., ‘unfair’ is used, especially given the high number of occurrences we aimed to collect and analyze. Having described our approach to examine evaluativity of both legal and ordinary discourse, we are now ready to discuss and analyze the data of our empirical study.[12]

3 Data

To address the question of how evaluative the legal discourse is in comparison to everyday talk and writing, we created two new text corpora. In this section, we describe the generation of the corpora and the selection of terms we used for our analysis.

3.1 Data Sources

Two corpora were generated. First, we created a corpus with legal documents (henceforth: ‘legal corpus’ or LC), based on the Free Law Project 2020, which provides open data from court opinions of the U.S. Courts of Appeals for the 1st to 11th regional circuit (without Washington DC and the federal court).[13] The courts of appeals are considered among the most powerful and influential courts in the United States, as they often set a legal precedent that guides subsequent legal rulings. Court opinions – our text data – are announced after the case is tried. They usually include a summary of facts, the applicable law and how it relates to the facts, the rationale for the decision, and a judgment.

Second, we created a corpus that includes non-legal language, based on comments on the world’s largest online forum Reddit (henceforth: ‘Reddit Corpus’ or RC).[14] Reddit seems to be a suitable control corpus for contrasting legal writings with ordinary texts. Other corpora, e.g.,

[12] In Reuter, Baumgartner, Willemsen (ms), we use the ‘and’-connective to determine the evaluative extent of thick concepts with the goal to differentiate thick adjectives from thin, descriptive, and value-associated adjectives.

[13] The courts of appeals are the intermediate appellate courts of the federal judiciary of the United States. Their task is to determine whether the law was applied correctly in the actual trial court. The courts of appeals sit below the Supreme Court, which is the last judicial instance to be consulted. In the vast majority of federal cases, these courts of appeals constitute the final legal instance.

[14] For the Reddit corpus (RC), we gathered data using the API for the Pushshift Reddit Data Set provided by J. Baumgartner et al. (2020).

corpora containing newspaper articles, would certainly allow us to make further interesting comparisons. Unfortunately, the inclusion of further corpora is beyond the scope of this article.

3.2 Corpus Generation and Adjective Selection

Both the court opinions from the Free Law Project 2020 and the Pushshift Reddit Data Set (J. Baumgartner et al., 2020) contain a wealth of information, not all of which are necessary for our purposes. As a first step, we needed to select adjectives that are suitable to measure the evaluativeness of the two corpora. We started with those thick adjectives that are often discussed in the philosophical literature as prototypical, agreed-upon thick terms, such as 'cruel' or 'honest'. We further required that all thick terms be regularly used in ordinary conversations. For instance, the terms 'lewd' or 'chaste' belong to some of the paradigmatic examples of thick terms, but we did not expect these terms to be part and parcel of the vocabulary of most laypeople or relevant in the legal context.

Within the group of thick terms, we created three sub-groups. The first group comprised thick *ethical* terms that are related to issues of moral relevance, such as 'honest' and 'rude'. We expected thick *ethical* terms to show up frequently in the legal discourse, as offenses in the criminal law are not merely legal offenses but transgress moral norms, too. Good candidates seemed to be terms related to physical harm, violations of someone else's property or dignity, and terms with which we may describe the defendant's character.[15]

The legal system operates within a set of epistemic norms – norms of what we should believe and may conclude from a given set of premises. Therefore, we created a second group consisting of thick *epistemic* concepts including, among others, terms like 'logical' or 'reasonable'. Finally, it is plausible to believe that some terms are used predominantly in the legal context and are likely to be thick, such as the term 'legal' itself, but also 'legitimate' or 'unlawful'. Thus, the third group

[15] Examples were selected based on the vast literature on thick ethical concepts. We specifically selected terms that seem to be relevant for the legal domain as well, excluding archaic terms or those connected to objectionable moral values. Again, 'blasphemy', 'lewdness', and 'chastity' might play no or at least a negligible role in legal discourses.

comprised such thick *legal* terms. While the philosophical literature hardly discusses thick legal terms, we selected legal terms based on Merriam-Webster's Law Dictionary as well as the authors' intuitions on which of the adjectives in the dictionary provide good examples of thick concepts.

In a second step, we examined the validity of our selection of terms based on our legal corpus. Since the legal corpus did not provide enough occurrences of all the terms, we had initially selected for us to run a sufficiently robust analysis, we had to adapt our list (e.g., the adjectives 'brutal' and 'crude' are rarely used in legal contexts).[16] Other terms like 'cruel' occur frequently, yet are often part of legal phrases, e.g., 'cruel and unusual punishment'. Infrequently occurring adjectives and those with a predominantly phrasal use were subsequently dropped.[17] The final list contained the following 24 target adjectives:

- *Descriptive*: 'active', 'ambiguous', 'complex', 'explicit', 'limited', 'practical'
- *Negative thick ethical*: 'dishonest', 'improper', 'unfair'
- *Positive thick ethical*: 'careful', 'honest', 'proper'
- *Negative thick epistemic*: 'illogical', 'inappropriate', 'inconsistent'
- *Positive thick epistemic*: 'consistent', 'logical', 'reasonable'
- *Negative thick legal*: 'illegal', 'unjust', 'unlawful'
- *Positive thick legal*: 'lawful', 'legal', 'legitimate'

In a third step, we reduced the legal and Reddit corpus to conjunctions that contain our target adjectives. All observations in the respective corpora were 'and'-conjunctions of two adjectives, one of which was the pre-defined *target adjective* while the other was what we henceforth call the *conjoined adjective*. The final LC contains 49'199 entries, whereas RC has 69'211. Both corpora were

[16] In order for a term to be analyzable, we needed a sufficiently high number of occurrences of that term in both corpora. As LC is a much smaller corpus, the number of occurrences within LC was therefore our main limiting factor.

[17] To avoid selection bias, we inductively selected a second battery of adjectives. We computed the co-occurrences of all adjectives and ranked each adjective based on their frequency and lexical diversity. We then selected promising adjectives that also matched our pre-defined concept classes. This inductive selection process was based on an analysis of part of speech tags (PoS-tags) in the legal corpus. PoS-tagging is an unsupervised method to annotate the syntactic structure of text data. For each of the legal corpus' sub-corpora (1st to 11th court circuits), we first drew a random sample of 2000 documents which were subsequently PoS-tagged using UDPipe (Straka & Strakovà, 2017; Straka & Strakovà, 2020). Based on these PoS-tags, we isolated all adjective pairs in 'and'-conjunctions. As a measure for lexical diversity, we used Yule's K (Yule, 1944; Tweedie & Baayen, 1998).

cleaned, PoS-tagged, lemmatized, and the conjoined adjectives were annotated with sentiment values from the SentiWords dictionary (Baccianella et al., 2010; Esuli & Sebastiani, 2006; Gatti et al., 2016; Guerini & Turchi, 2013).

4 Empirical Study

The goal of our corpus study was to measure the use of evaluative language by legal professionals. A rather straightforward approach would be to simply measure the overall sentiment score of the two corpora. Therefore, we did this for both the legal and the public corpus. In other words, we calculated the average sentiment score of the 49'199 entries of LC as well as of the 69'211 entries of RC. The legal corpus has an average absolute sentiment value of 0.2569, whereas the Reddit corpus is significantly higher with 0.3083. From this, one might infer that legal texts are less evaluative compared to ordinary discussions. This conclusion would yet be premature. Legal texts are different from ordinary writing, as they are often more technical and their claims are often more specific. Let us illustrate this with an example: While the sentence 'Voting is important' has a sentiment score of 0.28 (average sentiment value of 'voting' (0.12) and 'important' (0.45)), 'Cats are important' has a SentiScore of 0.44 (average sentiment value of 'cat' (0.43) and 'important' (0.45)). However, just because legal professionals are more likely to talk about voting and laypeople more about cats, this does not mean that legal scholars speak and write less evaluatively.

As we have argued above, to investigate whether legal language is indeed more descriptive than everyday conversations, we need to focus on the very use of evaluative terms (that legal scholars undoubtedly use a lot) even if looking at conjoined adjectives presents a much more limited phenomenon to study evaluative language. Hence, we deem this more limited approach to yield much more promising results.

We start by providing the basic descriptive statistics for our corpora, to gain insights into the data distributions (4.1). We then present the main results of our comparative analysis between legal language and public discourse (4.2).

4.1 Summary Statistics

In the following, we present the summary statistics for the key variables in each corpus: the sentiment values of conjoined adjectives (on a [-1,1] interval-scale), the sentiment polarity of the target adjective (pos/neg/neutral), and the concept classes of the target adjectives (Descriptive/Epistemic/Ethical/Legal). Table 1 shows the average sentiment dispersion and lexical diversity (*K*-values) by class and polarity for the legal as well as the Reddit corpus.[18]

Table 1: Summary Statistics comparing the Legal Corpus with the Reddit Corpus for three classes of concepts (descriptive, epistemic, legal, and ethical) as well as their polarity (neutral, negative, or positive). The lexical diversity of the conjuncts is higher the lower the *K*-value.

		Legal Corpus		Reddit Corpus	
Class	Polarity	Avg.	K	Avg.	K
Descriptive	neutral	0.05	102.40	-	-
Epistemic	negative	-0.26	124.78	-0.34	73.19
	positive	0.17	342.11	0.30	94.69
Legal	negative	-0.19	249.80	-0.38	198.62
	positive	0.17	1530.47	0.21	132.81
Ethical	negative	-0.30	507.77	-0.39	80.10
	positive	0.21	640.33	0.31	125.27

Let us look at the *legal corpus* first. There are two main takeaways. First, we have more extreme sentiment values for negative target adjectives than for positive or neutral ones. Negative conjuncts also have a higher diversity than positive target adjectives (the lower *K*, the more diverse). Second, irrespective of the concept class, positive target adjectives are conjoined with other positive adjectives, on average. Negative target adjectives, on the other hand, have a distinctly negative average, and the ones have a more neutral average. At first glance, the average observed sentiment seems consistent with our assumption that 'and'-conjunctions pair adjectives of the same polarity.

Compared to the legal corpus, the *Reddit corpus* shows a far more polar sentiment dispersion, which indicates that laypeople use the same adjectives more evaluatively than legal professionals.

[18] A more detailed version including the sentiment dispersion of the concept classes of Table 1 can be found at https://osf.io/8vx29/?view_only=336a7d04340a4c62a00a2f9bf4c2a44a

Lexical diversity is also a lot higher in RC than LC. The difference is most acute for legal and ethical thick concepts, and less so for thick epistemic concepts. Otherwise, RC exhibits the same patterns we noted above for LC.

Figure 2 below shows the sentiment dispersion on the level of the target adjectives shared by LC and RC (excl. descriptive concepts). The polarity of the target adjective indeed looks like a good indicator for the polarity of the conjoined adjective and vice versa: the sentiment spread (i.e., the whiskers) is mostly limited to either the positive or the negative region of the scale. In addition, the differences between the corpora we noted above are also present at the level of the target adjectives: LC has lower averages (i.e., dots) than RC across the board, except for 'dishonest' and 'improper'.

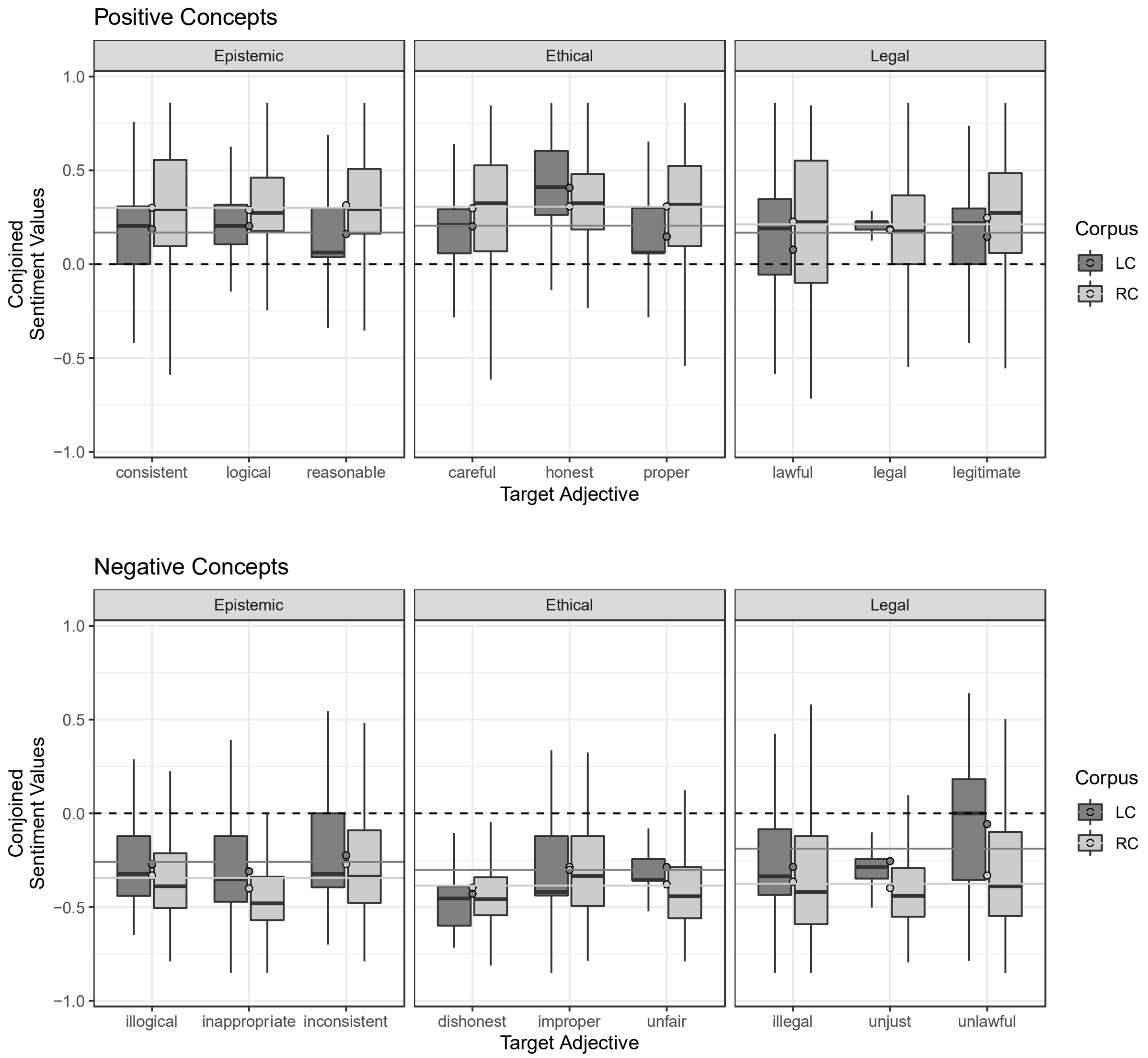

Figure 2: Sentiment Dispersion by Target Adjectives. The upper part of the figure displays sentiment dispersion of positive concepts in the three concept classes (epistemic, ethical, and legal) for three target items per class. The lower part of the figure displays the same for negative concepts. Example: the average conjoined sentiment for 'careful' is 0.297 in RC and 0.201 in LC; the lower and upper quantiles are both above zero, indicating a generally positive distribution.

4.2 Comparing the Use of Evaluative Adjectives by Legal Professionals and Laypeople

In our comparative analysis, we assess the average context effects for both corpora. First, we are interested in whether there is a *difference in intensity* of evaluative language between legal professionals and laypeople. If the legal context is indeed less evaluative, we should find that those potentially evaluative terms that do appear in legal conversations are communicated with less evaluative intensity compared to ordinary conversations.

4.2.1 Hypotheses

Our main hypotheses concern differences in sentiment intensity of conjoined adjectives between both corpora:

> **(H_1)**: The legal corpus contains conjunctions with more neutral sentiment values than the Reddit-based corpus.
>
> **(H_2)**: The legal corpus has more neutral conjoined sentiment values than the Reddit-based corpus for both positive and negative target adjectives.
>
> **(H_3)**. Differences in sentiment values between both corpora persist across all concept classes.

We also aimed to measure whether our concept classes, i.e., thick ethical vs. thick epistemic vs. thick legal, are distinguishable in legal language, i.e., whether they form distinct strata along the evaluative sentiment continuum. Hence, we also investigated the following hypothesis:

> **(H_4)**: All concept classes have significantly different sentiment averages for their respective adjective conjunctions in the legal corpus.

4.2.2 Results

To assess the overall evaluative intensity of the target adjective, we disregarded the polarity of the sentiment scores. To test **H_1**, we used a linear model with the absolute sentiment values as dependent variable (DV) and the corpora (LC vs. RC) as independent variable (IV). Based on this model, we computed the estimated marginal means (EMMs) for both corpora. Table 2 presents the EMMs based on the first model. The EMM for RC is 0.3622, the one for LC is 0.2360 on the absolute sentiment scale. LC has an average context-effect of β=-0.1262 compared to RC, t-value:-99.63, Pr(>|t|)=<2e^(-16), all other things equal. Hence, the sentiment values of the conjoined adjectives are significantly less intense for LC than for RC.

Table 2: Absolute Estimated Mean Sentiment Difference Between Corpora. Results are given on an absolute scale. Confidence level used: 0.95.

Corpus	EMM	SE	df	lower CL	upper CL
Reddit	0.3622	0.0008	109920	0.3607	0.3637
Legal	0.2360	0.0010	109920	0.2341	0.2380

To test $\mathbf{H}_2$, we further discriminated the *polarity* of the target adjectives. We, therefore, used the polarity-discriminator (positive vs. negative) as part of an interaction term (IV) with the corpus-dummy (LC vs. RC), and performed pairwise contrasts between the EMMs of the sentiment values for each corpus by target polarity. The effect of positive polarity compared to negative polarity was $\beta_1 = 0.6456$, t-value: 324.43, whereas the effect of LC compared to RC dropped slightly to $\beta_2 = 0.1085$, t-value: 35.31. The interaction of positive polarity and LC compared to the intercept has an effect of $\beta_3 = -0.2130$, t-value: -58.68. All effects were highly significant. ($Pr(>|t|) =< 2e^{-16}$). Table 3 contains the EMMs by sentiment polarity for this model. The pairwise contrasts are all significant, which supports that LC has more neutral values than RC on both sides of the sentiment scale. In other words, the difference between both corpora persisted when we took polarity into account.

Table 3: Estimated Mean Sentiment Difference Between Corpora by Target Polarity. Confidence level used: 0.95.

Corpus	EMM	SE	df	lower CL	upper CL
Polarity = negative					
Reddit	-0.3659	0.0015	109918	-0.3689	-0.3629
Legal	-0.2574	0.0027	109918	-0.2627	-0.2522
Polarity = positive					
Reddit	0.2797	0.0013	109918	0.2772	0.2822
Legal	0.1752	0.0015	109918	0.1723	0.1780

Our last comparative hypothesis concerns the question of whether different *concept classes* yield different results in their legal use compared to their everyday use. To test $\mathbf{H}_3$, we again used absolute sentiment values as DV. As IV, we used an interaction term between the corpus dummy (LC vs. RC) and the concept class factor (Ethical vs. Epistemic vs. Legal). Subsequently, we used pairwise contrasts between LC and RC for the EMMs of each concept class. Table 4 shows pairwise contrasts between the absolute estimates for each concept class and corpus on the absolute scale. The differences show significantly higher estimated values for RC compared to LC across all classes, which is consistent with the findings of the previous models.

Table 4: Planned Absolute Contrasts by Concept Class. Note: contrasts are on absolute scale.

Contrast	Estimate	SE	df	*t*-ratio	*p*-value
Class = Epistem. Concepts					
Reddit – Legal	0.1426	0.0020	109916	71.640	<.0001
Class = Legal Concepts					
Reddit – Legal	0.0985	0.0023	109916	42.669	<.0001
Class = Ethical Concepts					
Reddit – Legal	0.1153	0.0024	109916	48.231	<.0001

Besides analyzing differences between both corpora, we were interested to find out whether different concept classes can be distinguished in LC (**H**$_4$). As it is not possible to perform polarity contrasts for descriptive concepts, we dropped the descriptive concept class to investigate **H**$_4$. The linear model contains untransformed sentiment values as DV and an interaction of polarity (positive vs. negative) and concept class (Ethical vs. Epistemic vs. Legal) as IV. We further conducted pairwise contrast between concept classes of the same polarity. Table 5 shows the contrasts for all evaluative concept classes as a function of polarity. Among negative concepts, all classes have significantly distinct sentiment averages. The same was true for positive concepts, with the following exception: the difference between epistemic concepts ('consistent', 'logical', 'reasonable') and legal concepts ('lawful', 'legal', 'legitimate') was *not* significant. The estimated between-class differences are overall much smaller for positive than for negative target adjectives.

Table 5: Pairwise Contrasts between Concept Classes of the Same Polarity within Legal Corpus. P-value adjustment: tukey method for comparing a family of 4 estimates.

Contrast	Estimate	SE	df	*t*-ratio	*p*-value
Polarity = negative					
Desc. – Epistemic	0.1784	0.0058	49191	30.613	<.0001
Desc. – Legal	0.1073	0.0059	49191	18.032	<.0001
Desc. – Ethical	0.2204	0.0054	49191	41.052	<.0001
Epistemic – Legal	-0.0712	0.0061	49191	-11.653	<.0001
Epistemic – Ethical	0.0420	0.0055	49191	7.574	<.0001
Legal – Ethical	0.1132	0.0057	49191	19.956	<.0001
Polarity = positive					
Desc. – Epistemic	-0.0467	0.0037	49191	-12.792	<.0001
Desc. – Legal	-0.0459	0.0037	49191	-12.495	<.0001
Desc. – Ethical	-0.0838	0.0042	49191	-20.118	<.0001
Epistemic – Legal	0.0009	0.0028	49191	0.316	0.9891
Epistemic – Ethical	-0.0371	0.0035	49191	-10.739	<.0001
Legal – Ethical	-0.0383	0.0035	49191	-10.948	<.0001

The results support our hypothesis (**H**$_4$) that legal professionals use certain concept classes in a more evaluative manner than others. On average, ethical concepts are much thicker (i.e., more evaluative) than epistemic concepts, and epistemic concepts are thicker than legal concepts. It is important to stress, however, that the differences are overall very small and that their significance is positively impacted by the high number of observations. We are nonetheless confident that the comparative design and the within-context design establish a precedent for a sentiment-based analysis of concepts' thickness.

5 General Discussion

In recent years, evaluative language has attracted much attention from a variety of scholars in philosophy, linguistics, psychology, and other domains. There has also been significant progress in examining various aspects of evaluative language in the *legal domain* using corpus analytic tools. However, we have also observed that those previous studies are not sufficiently general and usually not comparative, which makes it difficult to address the question of how evaluative legal texts are.

Building on these two observations, the first step in our research was the identification of a linguistic approach to examine evaluative language in the legal domain more generally. We decided to investigate the arguably largest class of evaluative terms, namely thick terms. Thick terms are ubiquitous and highly frequent vehicles to make evaluative claims. Of course, by focusing on thick terms, we are not able to provide a complete picture of the evaluativity of legal discourse, but our approach allows for a *more* comprehensive view than any of the approaches we have found in the literature. The second step was to build corpora that allow for a comparative investigation of evaluative language. Thus, we decided to investigate both a legal corpus as well as an ordinary discourse corpus that includes comments and discussions drawn from Reddit. Our main empirical study was then guided by two questions:

1. How *strongly* evaluative do legal scholars use thick terms like 'illegal' and 'dishonest'?
2. Are there differences between legal and ordinary discourse in terms of how those terms are used?

To answer these questions, we investigated the sentiment values of adjectives that are conjoined with our target thick terms through the modifier 'and'. If legal discourse is indeed less evaluative, claims such as 'cruel and calculating' (SentiScore of 'calculating' is 0.15793) should be more representative of legal discussions, and claims such as 'cruel and mean' (SentiScore of 'mean' is -0.66570) more representative of ordinary discussions.

Examining the use of 49'199 occurrences of thick adjectives in the legal corpus and 69'211 occurrences of thick adjectives in the Reddit corpus, we were able to provide evidence for the claim that legal texts are less evaluative than ordinary writing. For all 6 different classes of thick terms that we investigated (negative epistemic, positive epistemic, negative legal, positive legal, negative ethical, positive ethical), legal scholars use thick terms more often in conjunction with adjectives that are more descriptive according to their sentiment values from the dictionary SentiWords.

In the remaining parts of this paper, we first tackle two objections against our interpretation of the empirical results (5.1), and then discuss some philosophical implications in regards to the evaluative variability of thick concepts (5.2).

5.1 Two objections

In this paper we have made the following inference:

1. Our empirical studies show that thick terms are conjoined more often with descriptive terms in legal texts compared to ordinary texts.
2. Thus: Thick terms are used more descriptively in legal discourse compared to ordinary discourse.

Obviously, this inference is not deductively valid but what we take to be an inference to the best explanation. However, other explanations for the empirical results are possible. Let us now discuss what we consider to be the two most plausible objections to our studies and our interpretation.

5.1.1 The 'Technical and Sophisticated Use' Objection

We justified our focus on thick terms instead of doing a global sentiment analysis of the two corpora by arguing that legal statements and conversations are often very technical, leading to the use of terms that are likely to be less evaluative. However, it seems a similar objection can be made against the approach we used in this paper: legal scholars often need to engage with more technical matters and, additionally, are trained to use vocabulary that is especially suited to engage with those matters. We can distinguish two versions of this objection. First, legal discourse is shaped by language games that distort the aggregated sentiment values of the target adjectives. Second, legal scholars may use less common and, hence, less evaluative words – be they technical terms or more sophisticated Latin or Greek-based vocabulary – in conjunction with thick adjectives, e.g., a legal person might say things like 'honest and pellucid'. Such uses would similarly push down the sentiment scores for the selected terms.

In response to the first objection, let us repeat once more that we only selected adjectives for our study that are not parts of common legal phrases, which indicate a different semantic embedding. Of course, the technical uses of thick terms might be more widespread and harder to detect, and, as such, would continue to influence our results. Nonetheless, we believe we took reasonable precautionary measures to escape this objection.

To tackle the second strand of the objection, we computed the lexical diversity of the legal corpus in relation to the Reddit corpus, i.e., we calculated, how many conjoined adjectives appeared in LC but not in RC, and vice versa. The results reveal that for all 6 different classes examined, the percentage of adjectives that are used in LC but not in RC is relatively low (see Table 6 below). Perhaps not surprisingly, the percentage of adjectives found in LC but not RC that are combined with our target *legal* terms is highest. Table 6 also lists the average sentiment values for those adjectives that appear in both corpora as well as in only one of them. The average sentiment values are consistently more polar in the intersection than in the complementary sets.

Table 6: Overlap between Corpora. The different measures of Jaccard's Distance show the overlap-ratio between RC and LC (RC ∩ LC), the ratio of unique tokens in RC (RC \ LC), and the ratio of unique tokens in LC (LC \ RC). The right hand side of the table contains the average sentiment within the subsets.

		Jaccard's Distance			Avg. Sentiment		
Class	Polarity	RC ∩ LC	RC \ LC	LC \ RC	RC ∩ LC	RC \ LC	LC \ RC
Epistemic	negative	0.177	0.705	0.117	-0.278	-0.181	-0.1150
	positive	0.226	0.663	0.110	0.213	0.133	0.0369
Legal	negative	0.179	0.586	0.236	-0.286	-0.214	-0.0968
	positive	0.223	0.565	0.211	0.195	0.113	0.0876
Ethical	negative	0.144	0.579	0.277	-0.327	-0.235	-0.1230
	positive	0.198	0.700	0.102	0.258	0.137	0.0744

While we need to be cautious in not overinterpreting these results, we can fairly safely conclude that legal professionals, by and large, do not conjoin a great number of adjectives with our target terms that laypeople would not also use. In other words, they speak the 'same' language.

5.1.2 The 'Two Scales' Objection

Our task at hand was to measure the evaluativeness of our thick *target* adjectives. The central auxiliary tool for this task is the SentiWords dictionary, which assigns each *conjoined* adjective a sentiment value on a scale [-1,1]. Using the power of high numbers (almost 120'000 uses of thick adjectives in total), we believe we get a robust and representative estimate of the evaluative force of our target adjectives. Crucially, we use the same sentiment values of the SentiWords dictionary for both corpora LC and RC. One might object, however, that such a general application of SentiWords to different corpora is not warranted. To illustrate the potential problem, consider how differently the term 'good' is used in various contexts. On the one hand, imagine a piano teacher who seldom makes a positive comment about the skills of her students. Then, one day, she listens to one of her students, and states 'That was good'. Well, you bet this was not just good, but most likely excellent. On the other hand, imagine you read a reference letter, in which the referee writes that her student's 'analytic skills are good'. This is a damning verdict: the student almost certainly will not get the job.

Applied to our case at hand, it seems we would need two different sentiment dictionaries for both contexts that reflect the true nature of the evaluative force of words. In the piano teacher context, the term 'good' would be assigned a sentiment value that is much higher compared to the referee context. If legal discourse is more like the piano teacher context and ordinary discourse is more like the referee context, then it seems that we cannot conclude that legal scholars use thick terms more descriptively, because we did not use sentiment scores for the value assignment that are suited to the respective corpora.

An important thing to note is that SentiWords contains prior polarities, i.e., the sentiment values of words out of context. In that respect, SentiScores are indicative of evaluativeness, regardless of context, and provide a measure of lexical evaluativeness. According to this measure, our study suggests that legal professionals use adjectives more neutrally than laypeople. Importantly, our approach does indeed take context into account, insofar as we assume that sentiment propagation patterns in conjunctions are a better estimate of evaluativeness than raw SentiScores. The objection, however, states that this is not yet sufficient: not only should context matter for measuring, but also for the basis of measurement. In response to this objection, we believe our method can account for context-dependent scaling. Instead of generating two different scales before measurement, we can model shifts in evaluativeness as part of the measurement itself.

To do so, we would need to consider the corpus as a set of variables that are specified in the broader embedding of our conjunctions. Such variables include negation, conditionals, modifiers, intensifiers, animacy, etc. Taking into account these variables would allow us to rescale the lexical SentiScore-values. In a perfect world, these variables account for the totality of contextual differences between legal and public discourse. In other words, we can operate with a single scale, while still accounting for pragmatic shifts in scale.

While we believe the objection does not undermine our approach to measuring the evaluativity of thick terms, we have not yet determined how those variables like negation, intensifier and animacy vary across the two corpora. So, why have we not done so yet? The short answer is: because it is a lot of work. Our long-term goal is to build a classifier for evaluative concept classes. Such a classifier will ultimately tell us, how the results of our study shift with a more fine-grained evaluative looking-glass. While we are already underway to address pragmatic mechanisms of evaluation by studying propagation patterns of evaluative information across conjunctions, presenting a more comprehensive picture will take some time.

5.2 Philosophical Implications

A central project in moral philosophy is to provide an adequate characterization and, eventually, a definition of thick ethical concepts. As we have already argued elsewhere (Willemsen & Reuter, 2020, 2021), we believe that this philosophical project can benefit greatly from empirical research on how thick concepts are used in different discourses. We, therefore, strongly agree with Heidi Li Feldman that by investigating and understanding the use of thick concepts in the legal domain, philosophers will be able to understand thick concepts more generally. Based on the empirical study presented in this chapter, what philosophical conclusions can we draw?

A major point of disagreement in the philosophical debate concerns the question of whether and to what extent thick concepts are variable in their evaluative content. Does a thick concept like *dishonest* always communicate an evaluative attitude with a fixed polarity and robust intensity, or are thick terms variable in their intensity and may even change their polarity (Blackburn, 1992; Hare, 1952; Väyrynen, 2021)?

Our results suggest that thick concepts can indeed vary in their evaluative intensity in two ways. First, our data indicates that the evaluative dimension of a thick term is more intense in ordinary conversational contexts compared to the legal context. Thus, thick concepts are variable depending on the conversational domain in which they are used. Second, even within a conversational domain, thick concepts can be more or less evaluatively intense. This is demonstrated by the relatively large variance we found when recording uses of thick adjectives in our corpora. However, looking more closely at the ratings for all 18 adjectives we studied, our data does not suggest that the polarity of a thick concept is likely to be very flexible: almost all thick adjectives ('lawful' and 'unlawful' being the only exceptions) predominantly co-occur with adjectives of the same polarity.

We would like to emphasize that these results are at best indicative and cannot suffice to decide the variability question. Our selection of terms is limited and a more comprehensive list of terms needs to be tested to allow for any general claims. Nevertheless, the methodological approach we presented in this paper motivates such follow-up studies to give a more precise picture of the variability of thick terms.

6 Conclusion

This paper provides a novel approach to the question of whether and how evaluative legal language is, using the tools from corpus linguistics. We created a legal corpus as well as an ordinary, Reddit-based corpus to examine thick adjectives as they are a frequently used linguistic means to communicate evaluation. Our analysis revealed that legal professionals use thick terms more often in conjunction with descriptive terms, suggesting that legal texts are less evaluative than ordinary discussions.

7 Funding Information and Acknowledgment

This research was funded by the Swiss National Science Foundation (SNSF), grant number PCEFP1_181082. We would like to express our gratitude to Davide Mazzi and Justin Sytsma for providing feedback on an earlier draft of this manuscript.

8 References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of the 7th International Conference on Language Resources and Evaluation*, 10, 2200–2204. https://www.aclweb.org/anthology/L10-1531/.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1), 830-839. https://ojs.aaai.org//index.php/ICWSM/article/view/7347.

Bednarek, M. (2008). *Emotion Talk Across Corpora*. New York: Palgrave Macmillan. doi: https://doi.org/10.1057/9780230285712.

Benamara, F., Chardon, B., Mathieu, Y., Popescu, V., and Asher, N. (2012). How do Negation and Modality Impact on Opinions? *Proceedings of the ACL-2012 Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (ExProM-2012), 10–18. https://halshs.archives-ouvertes.fr/halshs-00751065.

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

Cepollaro, Bianca, Sulpizio, Simone and Bianchi, Claudia, (2019), How bad is it to report a slur? An empirical investigation, *Journal of Pragmatics*, 146, 32–42. doi: doi.org/10.1016/j.pragma.2019.03.012.

Devlin, P., (1965). *The Enforcement of Morals*, Oxford: Oxford University Press.

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. https://doi.org/10.1111/cogs.12456.

Edwards, J. (2021). Theories of Criminal Law. *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition), Edward N. Zalta (ed.), https://plato.stanford.edu/archives/fall2021/entries/criminal-law/.

Enoch, D. and 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, 257–278. doi: 10.1093/acprof:oso/9780199675517.001.0001

Elhadad, M., McKeown, K. (1990). Generating Connectives. *Proceedings of the 13th conference on Computational linguistics*, 3, 97–101. https://doi.org/10.3115/991146.991164.

Eklund, M. (2011). What are Thick Concepts?. *Canadian Journal of Philosophy*, 41(1), 25–49. doi: https://doi.org/10.1353/cjp.2011.0007.

Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 417–422. https://www.aclweb.org/anthology/L06-1225/.

Feldman, H. L. (1997). 'Blending Fields: Tort Law, Philosophy, and Legal Theory,' *South Carolina Law Review*, 49(1), 167–185.

Feldman Hall, O., Son, J., & Heffner, J. (2018). Norms and the Flexibility of Moral Action. *Personality Neuroscience*, 1, E15. https://doi.org/10.1017/pen.2018.13.

Finegan, E. (2010). Corpus linguistic approaches to "legal language": Adverbial expression of attitude and emphasis in Supreme Court opinions, in M. Coulthard and A. Johnson (eds.) The Routledge Handbook of Forensic Linguistics, 65–77.

Flanagan, B., & Hannikainen, I. R. (2020). The Folk Concept of Law: Law Is Intrinsically Moral. *Australasian Journal of Philosophy*, 1–15. doi:10.1080/00048402.2020.1833953.

Free Law Project (2020). Court Listener: Bulk Data. https://www.CourtListener.com.

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. doi: https://doi.org/10.1109/TAFFC.2015.2476456.

Guerini, M., Gatti, L., and Turchi, M. (2013). Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. *Proceedings in EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing*, 1259-1269. https://www.aclweb.org/anthology/D13-1125.

Goźdź-Roszkowski, S. (2018). Values and Valuations in Judicial Discourse. A Corpus-Assisted Study of (Dis)Respect in US Supreme Court Decisions on Same-Sex Marriage. *Studies in Logic, Grammar and Rhetoric*, 53:1 (66), 61–79. doi: https://doi.org/10.2478/slgr-2018-0004.

Goźdź-Roszkowski, S. and Pontrandolfo, G. (2012). 'Evaluative Patterns in Judicial Discourse: A Corpus-based phraseological perspective on American and Italian criminal judgments,' *International Journal of Law, Language & Discourse*, 3 (2), 9–69.

Green, L. (2013). Should Law Improve Morality?. *Criminal Law and Philosophy*, 7: 473–494. doi: https://doi.org/10.1007/s11572-013-9248-3.

Hare, R.M. (1952), *The Language of Morals*, Oxford: Clarendon Press. ISBN: 9780198810773.

Hart, H. L. A. (1963). *Law, Liberty and Morality*, New York: Random House.

Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the Semantic Orientation of Adjectives. In *35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 174–181. doi: https://doi.org/10.3115/976909.979640.

Heffer, C. (2007). Judgment in Court: Evaluating Participants in Courtroom Discourse, in K. Kredens and S. Goźdź-Roszkowski (eds.), *Language and the Law: International Outlooks*. Frankfurt am Main: Peter Lang, 145–179. ISBN: 9783631574478.

Kotzee, B. & Wanderer, J. (eds.) (2008). Epistemology Through Thick and Thin. *Philosophical Papers*, 37(3).

Marin, M. J. & Rea, C. (2014). Researching legal terminology: a corpus-based proposal for the analysis of sub-technical legal terms. Asp. La revue du GERAS, 66, 61–82. doi: 10.4000/asp.4572.

Mazzi, D. (2010). 'This Argument Fails for Two Reasons …': A Linguistic Analysis of Judicial Evaluation Strategies in US Supreme Court Judgments. *International Journal for the Semiotics of Law*, 23, 373–385. doi: https://doi.org/10.1007/s11196-010-9162-0.

Mohammad, S. M. (2020). Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text. *Emotion Measurement*, 201–237. doi: https://doi.org/10.1016/B978-0-08-100508-8.00009-6.

Rakoczy, H. & Schmidt, M.F.H. (2012). The Early Ontogeny of Social Norms. *Child Development Perspectives*, 7(1), 17-21. https://doi.org/10.1111/cdep.12010.

Reuter, K., Baumgartner, L., Willemsen, P. (ms). Tracing Thick Concepts Through Corpora.

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

Roberts, D. (2013). Thick Concepts. *Philosophy Compass*, 8 (8), 677–88. doi: https://doi.org/10.1111/phc3.12055.

Schmidt, M.F.H & Tomasello, M. (2012). Young Children Enforce Social Norms. *Current Directions in Psychological Science*, 21(4), 232-236. https://doi.org/10.1177/0963721412448659.

Sripada, C. S. (2007). Nativism and Moral Psychology: Three Models of the Innate Structure That Shapes the Contents of Moral Norms. In *Moral Psychology: The Evolution of Morality: Adaptations and Innateness*, edit by W. Sinnott-Armstrong. Bradford Books (MIT Press), 319-344.

Stevenson, C. L. (1937), The Emotive Meaning of Ethical Terms. *Mind*, 46(181), 14–31. https://doi.org/10.1093/mind/XLVI.181.14.

Stevenson, C. L. (1938). Persuasive Definitions. *Mind, 47(187),* 331–50.

Straka, M. and Strakovà, J. (2020). UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings. *Proceedings of Language Resources and Evaluation*. https://arxiv.org/abs/2006.03687.

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. doi: https://doi.org/10.18653/v1/K17-3009.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267–307. doi: https://doi.org/10.1162/COLI_a_00049.

Tweedie, F.J., Baayen, R.H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities,* 32, 323–352. doi: https://doi.org/10.1023/A:1001749303137.

Väyrynen, P. (2021). Thick Ethical Concepts. In: *The Stanford Encyclopedia of Philosophy* (Spring 2021 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/.

Wang, S., & Yin, J. (2020). A Corpus-Based Study of Keywords in Legislative Chinese and General Chinese. In *Workshop on Chinese Lexical Semantics*, Springer, Cham. 639-653. doi: 10.1007/978-3-030-38189-9_65.

Williams, B. (1985). *Ethics and the Limits of Philosophy*, Cambridge, MA: Harvard University Press. ISBN: 9780674268586.

Williams, B (1995). What Has Philosophy to Learn from Tort Law? in D. G. Owen (ed.), *Philosophical Foundations of Tort Law*. Oxford: Oxford University Press, 487–497. ISBN: 9780198265795.

Willemsen, P., Reuter, K. (2020). Separability and the Effect of Valence. In Denison, Mack, Xu, Armstrong (Eds.), *Proceedings of the 42th Annual Conference of the Cognitive Science Society 2020*, 794-800.

Willemsen, P., Reuter, K. (2021). Separating the Evaluative from the Descriptive: An Empirical Study of Thick Concepts. *Thought: A Journal of Philosophy*. doi: https://doi.org/10.1002/tht3.488.

Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press. ISBN: 9781107633711.