

# 1 Analyzing Legal Language Using Corpus Analysis

## 1.1 Selecting Terms of Interest

The data for this study comprises two distinct corpora: a corpus with legal documents and a corpus with comments from Reddit, the world’s largest online-forum. The legal corpus (LC) contains court opinions from Court of Appeals for the 1st to 11th regional circuit (without DC and the federal court), based on open data provided by the Free Law Project (2020).<sup>1</sup> For the Reddit corpus (RC), we gathered data using the API for the Pushshift Reddit Data Set provided by Baumgartner et al. (2020).

In our case, the corpus generation is an iterative process. We start off with a list of target adjectives we specified without information about the corpora, and are based on the literature. This initial list contains adjectives in two categories, namely descriptive concepts or concepts which have a high potential to count as thick concepts. Among those thick concepts, we created sub-groups which differ in what non-descriptive information might be conveyed. First, there are ethical thick concepts whose non-descriptive content is ethical in nature. Examples were selected based on the vast literature on thick ethical concepts that are of special interest in the legal domain, as offences in the criminal law are not merely legal offences, but transgress moral norms too. Second, 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. We therefore created a group of thick epistemic concepts which is inspired by the philosophical literature. Finally, it is plausible to believe that some thick concepts are used exclusively or predominantly in the legal context, such as the term “legal” itself.

However, it turns out that some of these adjectives are rarely used in the legal context, while others may occur frequently, yet are most often part of legal phrases, which indicate a different semantic embedding. In order to avoid selection bias and exclude adjectives with predominantly phrasal use, we inductively select a second battery of adjectives that is used complementary to the first. This inductive approach is based on an analysis of part of speech (PoS)-sequences in the legal corpus. PoS-tagging is an unsupervised method to annotate the syntactic structure of text data. For each of the subcorpora (1st to 11th Court of Appeals), we first draw a random sample of 2000 documents which are subsequently PoS-tagged using UDPipe (Straka and Straková, 2017, 2020). Based on these PoS-tags, we isolate all syntactic structures of the form  $(M) * A(, ) * C(M) * A$  ( $M$  = modifier,  $A$  = adjective,  $C$  = conjunction, (...) = optional part). Finally, all adjectives are ranked according to frequency as well lexical diversity in regards to the conjoined adjectives. We use Yule’s  $K$  (Yule, 1944; Tweedie and Baayen, 1998) as a measure for lexical diversity. Based on this ranking, we manually select adjectives that match our concept classes and retrieve documents containing suitable target structures both from the legal corpus and via the Pushshift API. The final list of 24 adjectives is shown in Table 1 below.

---

<sup>1</sup>The Court of Appeals are the intermediate appellate courts of the United States federal judiciary. Each of the so-called regional circuits hears appeals from the district courts within its borders, or from other designated federal courts and administrative agencies. The appeals from the circuit courts are taken to the Supreme Court of the United States, which means that circuit court decisions are quite influential. Moreover, circuit court decisions, establish binding precedents, unlike those of the lower federal courts. The lower federal courts have to follow the guidance of their circuit court in similar cases, irrespective of the trial judge’s opinion.

Class	Target	Polarity	Sentiment Quantiles				Lex. Diversity			
			25%	50%	75%	Avg.	N	TTR	CTTR	K
Descriptive	active	neutral	-0.10	0.08	0.27	0.07	948	0.21	4.64	343.83
Descriptive	ambiguous	neutral	-0.35	-0.18	-0.03	-0.19	1182	0.16	3.93	638.53
Descriptive	complex	neutral	-0.16	0.00	0.16	-0.01	1954	0.14	4.26	292.63
Descriptive	explicit	neutral	0.03	0.21	0.23	0.13	1565	0.11	3.18	982.63
Descriptive	limited	neutral	-0.01	0.08	0.23	0.09	1501	0.19	5.27	375.77
Descriptive	practical	neutral	0.01	0.19	0.29	0.18	1338	0.13	3.31	444.01
Epistemic	illogical	negative	-0.50	-0.39	-0.20	-0.33	3855	0.16	7.13	143.74
Epistemic	inappropriate	negative	-0.57	-0.45	-0.27	-0.39	6161	0.11	6.04	152.70
Epistemic	inconsistent	negative	-0.45	-0.34	-0.05	-0.26	4847	0.16	7.98	100.21
Epistemic	consistent	positive	0.06	0.24	0.50	0.26	7038	0.13	7.59	117.42
Epistemic	logical	positive	0.17	0.27	0.41	0.28	8426	0.10	6.44	201.46
Epistemic	reasonable	positive	0.06	0.18	0.35	0.22	15523	0.05	4.47	260.49
Legal	illegal	negative	-0.55	-0.42	-0.11	-0.34	3633	0.15	6.41	269.85
Legal	unjust	negative	-0.55	-0.42	-0.27	-0.37	2879	0.15	5.53	234.55
Legal	unlawful	negative	-0.42	-0.17	0.11	-0.15	1566	0.17	4.75	235.22
Legal	lawful	positive	-0.08	0.19	0.50	0.16	1699	0.14	4.15	427.35
Legal	legal	positive	0.01	0.22	0.22	0.18	14254	0.04	3.70	1189.78
Legal	legitimate	positive	0.00	0.24	0.41	0.20	6539	0.10	5.91	252.86
TC	dishonest	negative	-0.55	-0.45	-0.34	-0.40	5271	0.13	6.63	122.17
TC	improper	negative	-0.44	-0.42	-0.12	-0.30	1906	0.17	5.38	521.61
TC	unfair	negative	-0.51	-0.39	-0.27	-0.36	7617	0.10	5.95	175.56
TC	careful	positive	0.06	0.29	0.46	0.26	4200	0.12	5.30	211.82
TC	honest	positive	0.21	0.32	0.51	0.32	6624	0.09	5.21	356.66
TC	proper	positive	0.06	0.22	0.45	0.23	7884	0.10	6.01	424.46

Table 1: Final Adjective List

## 1.2 The Data

The full LC has 49'199 entries, whereas RC has 69'211. Both corpora are cleaned, PoS-tagged, lemmatized and the conjoined adjectives are annotated with sentiment values from the SentiWords dictionary based on SENTIWORDNET (Esuli and Sebastiani, 2006; Baccianella et al., 2010; Guerini et al., 2013; Gatti et al., 2016). In the following we present the summary statistics for each corpus. Table 2 shows the measures of sentiment dispersion and lexical diversity by class and polarity for LC. Table 3 contains the same for RC.

The sentiment dispersion in Table 2 indicates more extreme values for negative target adjectives overall. According to Yule's  $K$ , their conjuncts are also more lexically diverse

than the ones of the positive target adjectives. The average observed sentiment is consistent with our assumption that AND-conjunctions pair adjectives of the same polarity. Descriptive concepts have a very neutral average (0.05), which is also consistent with our expectations. However, there seems to be an overlap between descriptive concepts and of the several concept classes, which might prove to be significant later on.

Class	Polarity	Sentiment Quantiles				Lex. Diversity		
		25%	50%	75%	Avg.	TTR	CTTR	K
Descriptive	neutral	-0.08	0.03	0.23	0.05	0.11	6.98	102.40
Epistemic	negative	-0.44	-0.33	-0.10	-0.26	0.16	6.09	124.78
Epistemic	positive	0.04	0.14	0.30	0.17	0.05	3.68	342.11
Legal	negative	-0.40	-0.29	0.00	-0.19	0.13	4.69	249.80
Legal	positive	0.07	0.22	0.22	0.17	0.04	2.93	1530.47
TC	negative	-0.44	-0.35	-0.20	-0.30	0.11	4.81	507.77
TC	positive	0.06	0.16	0.33	0.21	0.06	3.55	640.33

Table 2: Summary Statistics LC

Table ?? indicates a far more polar sentiment dispersion in RC compared to LC, which is also reflected by the more polar averages. Lexical diversity is a lot higher in RC than LC, which was to be expected, due to the more codified nature of legal language use. The difference in homogeneity is most acute for legal concepts and TC, but less so for epistemic concepts.

Class	Polarity	Sentiment Quantiles				Lex. Diversity		
		25%	50%	75%	Avg.	TTR	CTTR	K
Epistemic	negative	-0.52	-0.42	-0.18	-0.34	0.10	7.65	73.19
Epistemic	positive	0.16	0.29	0.50	0.30	0.08	7.26	94.69
Legal	negative	-0.55	-0.44	-0.24	-0.38	0.13	6.65	198.62
Legal	positive	0.00	0.22	0.45	0.21	0.09	6.43	132.81
TC	negative	-0.55	-0.45	-0.29	-0.39	0.09	6.93	80.10
TC	positive	0.16	0.32	0.52	0.31	0.09	7.06	125.27

Table 3: Summary Statistics RC

The sentiment dispersion on the level of the target adjectives shared by LC and RC are shown in Figure 1. As we can see, the polarity of the target adjective is a good indicator for the polarity of the conjoined adjective, and vice versa. With the exception of *lawful* and *unlawful*, the sentiment spread is mostly limited to either the positive or the negative region of the scale and does not include the midpoint. The differences between the corpora we noted above are also present on the level of the target adjectives: LC has lower averages (i.e. dots) than RC across the board, with the exception of *dishonest* and *improper*.

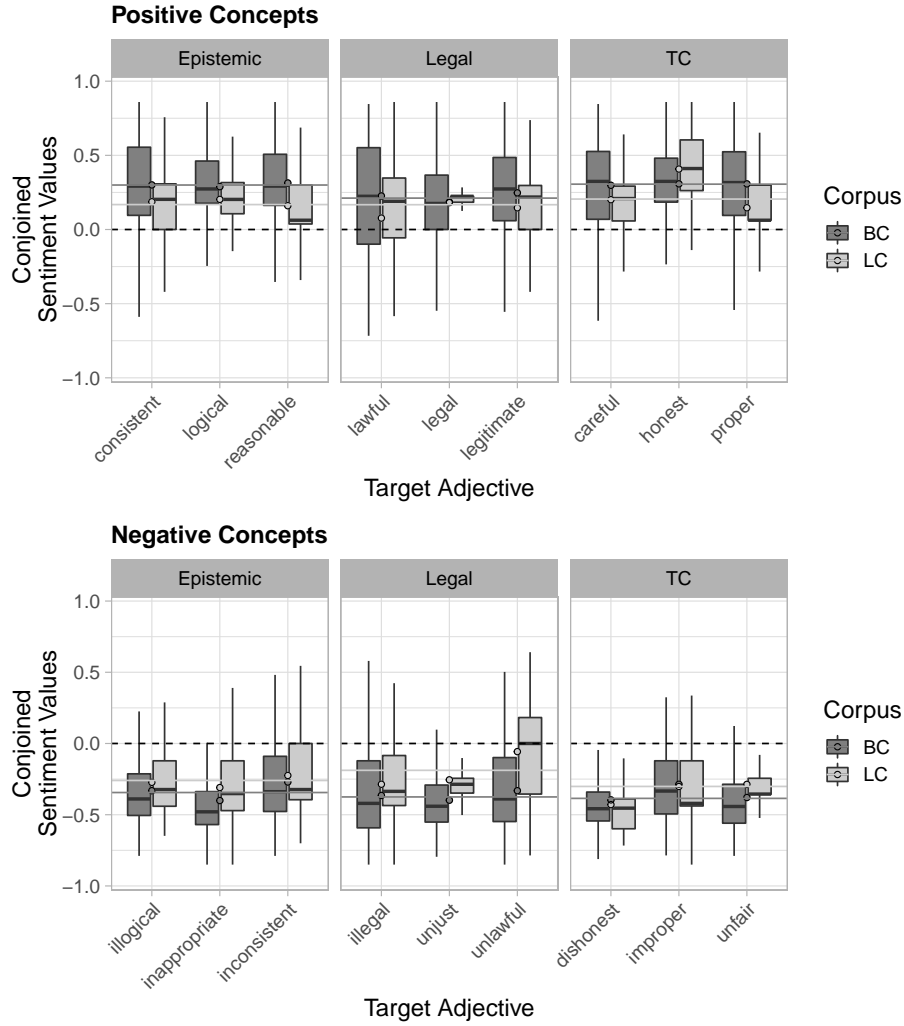


Figure 1: Sentiment Dispersion by Target Adjective

### 1.3 Study 1

#### 1.4 Proceeding

In the first study, we assess the average context effects for both corpora. The first hypothesis concerns the average absolute sentiment intensity of conjoined adjectives in each corpus: we expect the legal corpus to contain conjunctions with more neutral sentiment values than the Reddit corpus ( $H_1$ ). In order to test  $H_1$ , we use a linear model with the absolute conjoined sentiment values as dependent variable and the corpus-dummy (RC/LC) as independent variable. Based on this model, we compute the estimated marginal means (EMM) for the corpora. This gives us an overall estimate of differences in sentiment intensity between the corpora, irrespective of sentiment polarity (pos/neg).

For the second hypothesis, we further discriminate between positive and negative target adjectives. As with  $H_1$ , we expect the legal corpus to have more neutral conjoined sentiment values on average, and that the effect is significant for both positive and negative target adjectives ( $H_2$ ). On the one hand,  $H_{2a}$  posits that conjoined sentiment values for negative target adjectives are on average significantly closer to 0 in the case of the legal corpus compared to the Reddit corpus. On the other hand,  $H_{2b}$  assumes the same for positive target adjectives. In contrast to  $H_1$  we are not only interested in sentiment intensity, but also also in polarity. Thus the second model uses the non-transformed sentiment values as dependent variable, instead of absolute sentiment values used to test  $H_1$ . The polarity-discriminator (pos/neg) is added as an interaction term between the corpus-dummy and the polarity-dummy for the target adjectives. This allows us to assess the estimated marginal means of the conjoined sentiment values for each corpus by target polarity.

The third and final model hypothesis is that the difference between the corpora further persist across all concept classes. We expect more neutral average sentiment values for each concept class in the legal corpus compared to the same concept class in the Reddit corpus ( $H_3$ ). Because we are mostly interested in differences in terms of intensity, the third model once again uses absolute conjoined sentiment values, as for testing for  $H_1$ . As estimator we use an interaction term between the corpus dummy (BC/LC) and the concept class factor (Epistemic/Legal/TC). The partial hypotheses are defined as follows:  $H_{3a}$  assumes that epistemic concepts have lower average conjoined sentiment values in LC compared to BC.  $H_{3b}$  expects that the same holds for legal concepts, while  $H_{3c}$  does so for TC.

#### 1.4.1 Results

The first model of study 1 is presented in Table 4 and shows the effect of context (LC v. RC) on the absolute sentiment of conjoined adjectives. According to the linear model, LC has an average effect of  $\beta = -0.1262$  compared to RC,  $t$ -value:  $-99.63$ ,  $Pr(> |t|) = < 2e^{-16}$ , all other things equal. The estimated marginal mean for RC is 0.3622, the one for LC is 0.2360. Hence, the sentiment values of the conjoined adjectives are indeed less intense for LC than for RC.

Corpus	Est. Mean	SE	df	lower CL	upper CL
RC	0.3622	0.0008	109920	0.3607	0.3637
LC	0.2360	0.0010	109920	0.2341	0.2380

Results are given on the absolute scale. Confidence level used: 0.95

Table 4: Absolute Estimated Mean Sentiment Difference Between Corpora

The second model shows similar results, even if we take polarity into account. The effect of positive polarity compared to negative polarity is  $\beta_1 = 0.6456$ ,  $t$ -value: 324.43, whereas the effect of LC compared to RC drops 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 are highly significant on a 0.05 alpha-level ( $Pr(> |t|) = < 2e^{-16}$ ). Table 5 contains the estimated marginal means for this model. In summary, LC has significantly more neutral sentiment values than RC on both sides of the sentiment scale.

Corpus	Est. Mean	SE	df	lower CL	upper CL
Polarity = negative					
RC	-0.3659	0.0015	109918	-0.3689	-0.3629
LC	-0.2574	0.0027	109918	-0.2627	-0.2522
Polarity = positive					
RC	0.2797	0.0013	109918	0.2772	0.2822
LC	0.1752	0.0015	109918	0.1723	0.1780
Confidence level used: 0.95					

Table 5: Estimated Mean Difference Between Corpora by Target Polarity

Table 6 shows the contrasts between the absolute estimates for each concept class and corpus. The differences indicate higher estimated values for RC compared to LC across the board, which is consistent with the findings of the previous models. Interestingly, legal concepts show the smallest difference of all concept classes. All contrasts are significant on a 0.05 alpha-level. The effect sizes for all estimators can be found in Table 7 in the [Appendix](#). This shows that the differences between legal and everyday use of concepts are robust across concept classes.

Contrast	Estimate	SE	df	<i>t</i> -ratio	<i>p</i> -value
Class = Epistemic					
RC - LC	0.1426	0.0020	109916	71.640	<.0001
Class = Legal					
RC - LC	0.0985	0.0023	109916	42.669	<.0001
Class = TC					
RC - LC	0.1153	0.0024	109916	48.231	<.0001
Note: contrasts are still on the abs scale					

Table 6: Planned Absolute Contrasts by Concept Class

## 1.5 Study 2

### 1.6 Proceeding

In the second study, we focus on the legal corpus only. In order to be able to compare the results of the evaluative concepts classes to a baseline, we added corpus entries for the following descriptive target adjectives: *active*, *ambiguous*, *complex*, *explicit*, *limited*, and *practical*. Instead of comparing context effects (Study 1), we want to inquire whether the concept classes can be distinguished from each other within the legal context. The linear model includes the absolute conjoined sentiment as dependent variable and the concept classes as independent variable, followed by pairwise contrasts between the estimated marginal means of the concept classes. This will allow us to assess differences in sentiment intensity. We cannot perform planned contrasts by target polarity, because the descriptive concept only have a neutral polarity, which leads to empty interaction levels and contrasts.

### 1.6.1 Results

In study 2, we focus on LC and add descriptive concepts as a baseline. The aim of this study is to assess the possibility of distinguishing concept classes within the legal discourse, based on sentiment values. Table ?? presents the absolute contrasts between the concept classes in LC. As we can see, all concept classes have significantly different sentiment intensity (0.05 alpha-level), which indicates, that a there is no need to further discriminate by polarity in order to distinguish the classes from each other. The smallest differences are between epistemic and legal concepts ( $\Delta\bar{y} = -0.0128$ ), descriptive and epistemic concepts ( $\Delta\bar{y} = -0.0140$ ), as well as descriptive and legal concepts ( $\Delta\bar{y} = -0.0268$ ). The contrasts involving TC, on the other hand, have a much wider spread.

Contrast	Estimate	SE	df	t-ratio	p-value
Desc. - Epistemic	-0.0140	0.0024	49195	-5.873	<.0001
Desc. - Legal	-0.0268	0.0024	49195	-11.147	<.0001
Desc. - TC	-0.0674	0.0026	49195	-25.942	<.0001
Epistemic - Legal	-0.0128	0.0020	49195	-6.298	<.0001
Epistemic - TC	-0.0534	0.0023	49195	-23.678	<.0001
Legal - TC	-0.0406	0.0023	49195	-17.880	<.0001

Note: contrasts are still on the abs scale

P value adjustment: tukey method for comparing a family of 4 estimates

## References

- Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 2200–2204.
- Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J., and Io, P. (2020). The Pushshift Reddit Dataset. Technical report.
- Esuli, A. and Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 417–422.
- Free Law Project (2020). Bulk Data 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.
- Guerini, M., Gatti, L., and Turchi, M. (2013). Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1259–1269.

- Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.
- Straka, M. and Straková, J. (2020). UDPipe at EvaLatin 2020: Contextualized Embeddings and Treebank Embeddings. In *Proceedings of Language Resources and Evaluation*. arXiv.
- Tweedie, F. J. and Baayen, R. H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. Technical report.
- Yule, G. U. (1944). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

## 2 Appendix



	<i>Dependent variable:</i>
	abs(sentiment)
corpusLC	−0.143*** (0.002)
classLegal	−0.031*** (0.002)
classTC	0.026*** (0.002)
corpusLC:classLegal	0.044*** (0.003)
corpusLC:classTC	0.027*** (0.003)
Constant	0.361*** (0.001)
Observations	109,922
R <sup>2</sup>	0.093
Adjusted R <sup>2</sup>	0.093
Residual Std. Error	0.202 (df = 109916)
F Statistic	2,249.365*** (df = 5; 109916)
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01

Table 7: Linear Model for Planned Absolute Contrasts by Concept Class