1 What Has to be Mentioned Before

2 Data

The data for this study comprises two distinct corpora: a corpus with legal documents and a baseline corpus with comments from Reddit, the world's largest online-forum. The legal corpus contains court opinions from Court of Appeals for 1st to 11th circuit, based on open data provided by the Free Law Project (2020). For the baseline corpus, 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. This initial list, L_1, includes 2x5 adjectives often discussed by the TC-literature and we deem rather uncontroversial examples of epistemic concepts, thick and thin concepts as well as legal concepts. However, 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 any sort of selection bias and exclude adjectives with predominantly phrasal use, we inductively select a second battery of adjectives. 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). The conjuncts are then pooled across all subcorpora and only AND-conjunctions are retained. 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 add antonyms to form list L_2. In a third step, we combine L_1 and L_2, and use them to retrieve documents containing suitable target structures both from the legal corpus and via the Pushshift API. The combined list is shown in **??** in the 5.

The full legal corpus has XXX entries, the baseline corpus XXX. In order to keep the computational resources low while keeping a high enough n, we reduce the legal subcorpora each randomly by -40%, resulting in XXX legal documents overall. Both corpora are subsequently 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).

3 Method

In the following, we will discuss the methods used in our two studies.

3.1 Study 1

3.2 Study 2

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: . Instead of comparing context effects (Study 1), we want to inquire whether the concept classes cluster differently within the legal context. We use a combination of K-Means Clustering and Principal Component Analysis (PCA), which is informed by the sentiment values of the conjoined adjectives and a measure for lexical diversity of the target adjective, i.e. Yule's K (Yule, 1944; Tweedie and Baayen, 1998). The cluster analysis further includes information based on the vector space of the corpus, which is created using the *Semantic Vectors* package (Widdows and Ferraro, 2008; Widdows and Cohen, 2010, 2016). For every pair of target adjective and conjoined adjective, we compute the cosine similarities. Cosine similarities are essentially high dimensional representations of co-occurence measures, and inform the cluster analysis about the semantic similarity of the conjuncts by taking the whole corpus into account. In the final ANOVA model, the cosine similarities are added as weights for the sentiment values of the conjoined adjectives.

4 Results

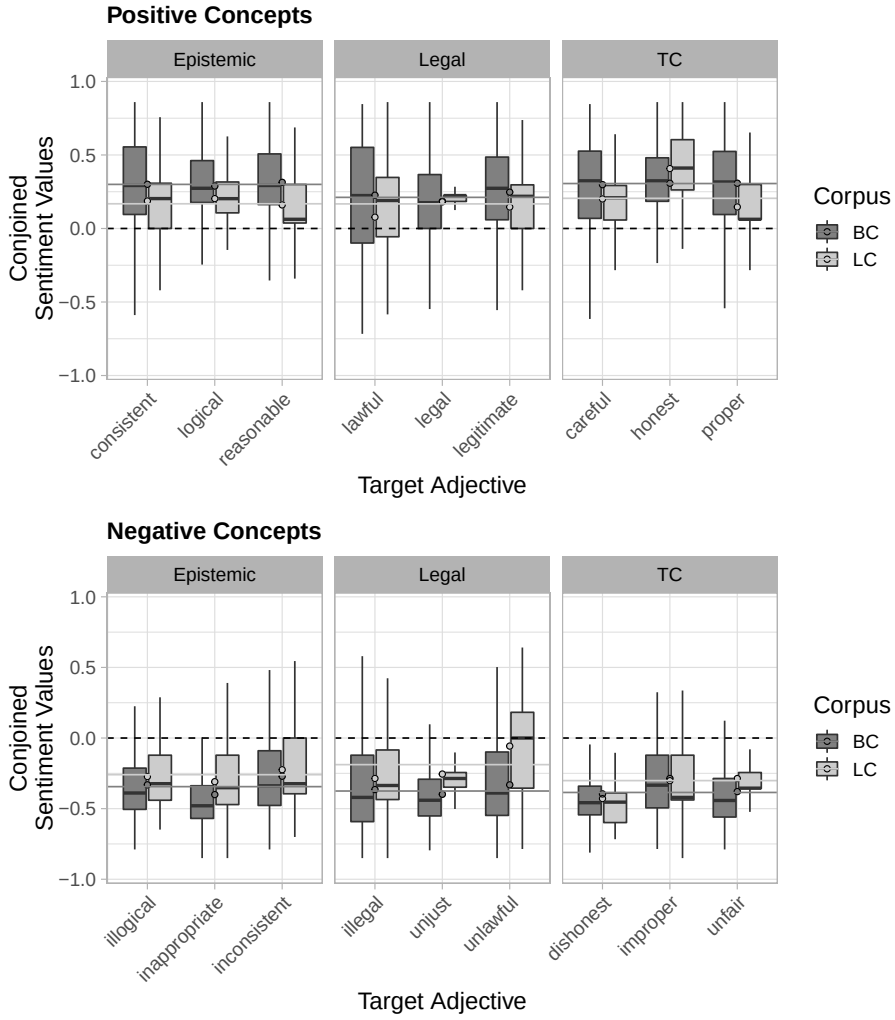
4.1 Descriptive Statistics

		Sentiment Quantiles				Lex. Diversity		
Class	Polarity	25%	50%	75%	Avg.	TTR	CTTR	K
Descriptive	neutral	-0.18	0.00	0.22	0.02	0.15	6.17	168.36
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 1: Summary Statistics Legal Corpus

		Sentiment Quantiles				Lex. Diversity		
Class	Polarity	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 2: Summary Statistics Baseline Corpus



4.2 Study 1

4.3 Study 2

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.

Widdows, D. and Cohen, T. (2010). The semantic vectors package: New algorithms and public tools for distributional semantics. In *Proceedings - 2010 IEEE 4th International Conference on Semantic Computing, ICSC 2010*, pages 9–15.

Widdows, D. and Cohen, T. (2016). Graded semantic vectors: An approach to representing graded quantities in generalized quantum models. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9535, pages 231–244. Springer Verlag.

Widdows, D. and Ferraro, K. (2008). Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*.

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

5 Appendix