On the Semantic Difference of Judicial and Standard Language

Gregor Behnke
g.behnke@uva.nl
Institute for Logic, Language and Computation
University of Amsterdam
Amsterdam, The Netherlands

Niklas Wais
niklas.wais@uni-leipzig.de
Institute for Civil, Information, Data and Media Law
Leipzig University
Leipzig, Germany

ABSTRACT

Legal language is considered to be a key obstacle to the comprehensibility of court decisions for laypeople. While differences between written 'standard' and legal language have already been analysed with regard to syntactic peculiarities, there is still a lack of findings on the influence of divergent word meanings on comprehensibility. We present the course and the preliminary results of a study elaborating such ambiguities on the basis of over half a million German court decisions. As these differences are highly language-dependent, our study consequentially relates (only) to German.

CCS CONCEPTS

• **Applied Computing** → Law.

KEYWORDS

NLP, semantics of legal texts

ACM Reference Format:
Gregor Behnke and Niklas Wais. 2023. On the Semantic Difference of Judicial and Standard Language. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3594536.3595123

1 INTRODUCTION

Acts of scientific communication often rely on subject-specific technical language that is difficult for outsiders to understand. Court decisions, which are closely connected to legal research [13], are no exception. However, court decisions are not primarily intended to enrich professional discourse but to contribute to the accomplishment of a concrete social task – the safeguarding and enforcement of the law. While calls for more comprehensible communication are not unusual in the academic environment, they consequently carry particular weight in relation to court decisions.

Known factors that influence the comprehensibility of texts include sentence and word lengths as well as word frequencies. Linguistic analyses of court decisions that are aligned accordingly have been available for some time [6]. In this work, we go beyond such syntax-focused measures and focus on systematic ambiguity. Court decisions are particularly susceptible to the use of ambiguous words, since the legal terminology extensively differentiates existing words, instead of resorting to technical terms that are clearly identifiable as such. Therefore, laypersons are likely to partially attach a different meaning to the words of a decision than the issuing court did. The resulting systematic ambiguity was identified early on as a central obstacle to comprehensibility [10], but beyond anecdotal fragments had not yet been illuminated.

Our study aims to shed light on those differences with the help of distributional semantics. Based on the hypothesis that the meaning of a word is reflected in the context of its usage, information about semantic relations can be obtained "radically empirical" [2] from a corpus by examining the embedding of words in sentences. In addition to dispensing with third-party sources such as dictionaries or lexical networks, distributional methods offer the advantage of producing directly mathematically usable results. To analyze the emergence of ambiguities in court decisions caused by legal terminology in this way, we initially gather two datasets: one containing court decisions and one containing what is commonly regarded as 'standard' (nontechnical) language. From these datasets, we generate word vectors using word2vec. Since a direct cross-model comparison is not possible, we develop more sophisticated indirect methods. This includes measuring the respective distance to similar words, the D-value of the Kolmogorov-Smirnov test, the formation of word pairs as well as translation learning. Lastly, we present and evaluate the results of our study using graphical representations.

2 DATA COLLECTION AND PROCESSING

As a corpus of 'standard' German representing how laypeople would understand any given word, our study draws on a collection of sentences in German provided by the University of Leipzig [3]. We use their collection of 30 million sentences from randomly selected German websites and 30 million sentences from German news, both from 2019. These sizes were chosen to obtain a corpus that is roughly equal in size to the second corpus comprising German court decisions, the construction of which is described below. We reference these corpora as the standard and decision corpus.

Obtaining the Dataset. Since 2010, the German federal government and 15 of the 16 federal states (Bavaria joined in 2016) have been publishing selected court decisions on individual online platforms. Due to unsolved copyright issues [5], we are not able to publicly release the full dataset we gathered. We have, however, published the tool `gesp` (https://github.com/niklaswais/gesp) that was created to download the decisions. By running `gesp`, anyone can replicate our dataset. The tool is also able to add new court decisions to the dataset in order to update it. Given that `gesp` is free software, it is a significant improvement over prior datasets, which require a licence or are not always up-to-date.

In creating the corpus, we excluded court decisions provided by the federal states of Saxony, Thuringia, and Bremen. Thuringia

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICAIL 2023, June 19–23, 2023, Braga, Portugal
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0197-9/23/06.
https://doi.org/10.1145/3594536.3595123

	Total	< 1990	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
Federal	130928	4	1	0	1	1	0	3	1	669	3722	5064	5406	5263	5622	6008	6542	6685
Baden-Württemberg	25274	0	0	0	0	0	0	0	1	0	7	106	476	666	1118	1006	1027	1075
Bavaria	34700	0	2	0	0	0	1	4	0	3	10	182	431	686	726	875	818	805
Berlin	21344	0	0	0	0	1	0	1	0	0	0	36	72	161	267	452	545	651
Brandenburg	17035	1	0	0	0	0	0	0	0	0	5	5	16	246	249	184	192	270
Bremen	578	0	1	0	0	0	0	0	1	0	0	0	9	8	41	68	95	83
Hamburg	10557	0	0	0	0	0	0	0	0	1	0	3	23	122	251	215	252	240
Hesse	41740	4126	357	356	347	345	315	356	322	444	346	368	622	869	826	1213	1267	1371
Lower Saxony	31323	0	0	0	0	0	0	3	1	6	4	21	888	1205	1674	1901	1786	1734
Mecklenburg-Vorpommern	4986	0	0	0	0	0	0	0	0	0	0	1	9	24	31	100	159	119
North Rhine-Westphalia	180906	988	198	275	439	524	558	647	1030	1528	1962	2668	3748	4192	5820	6471	6916	8161
Rhineland-Palatinate	20309	1	0	0	0	0	0	1	0	1	16	137	371	437	568	640	1285	1557
Saarland	7267	0	0	0	0	0	0	3	1	0	0	12	21	58	60	187	202	237
Saxony	1988	0	0	0	0	0	0	0	0	0	3	73	64	70	129	190	126	170
Saxony-Anhalt	13419	0	1	0	0	0	0	5	1	0	2	4	22	239	250	319	292	282
Schleswig-Holstein	9007	1	1	0	0	1	0	0	0	0	1	41	145	231	223	217	326	459
Thuringia	1008	0	0	0	0	2	2	2	5	3	6	11	71	78	78	121	112	96
Total	552369	5121	561	631	787	874	876	1025	1363	2655	6084	8732	12394	14555	17933	20167	21942	23995

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
Federal	6529	7141	7278	5459	6285	6212	5956	5358	5081	4701	5013	5006	4483	4584	4357	2493
Baden-Württemberg	1731	1376	1374	1317	1302	1428	1274	1215	1253	1257	1411	1200	993	974	917	770
Bavaria	975	982	981	726	0	0	0	0	0	0	5265	4897	3974	4168	4231	3958
Berlin	635	717	545	435	1387	1421	1551	1350	1421	1412	1329	1351	1507	1601	1475	1021
Brandenburg	487	801	605	365	1275	1186	1141	1107	1002	864	951	1142	1093	1412	1528	908
Bremen	74	81	66	51	0	0	0	0	0	0	0	0	0	0	0	0
Hamburg	255	317	269	178	617	547	722	690	667	749	822	885	783	722	742	485
Hesse	1527	1684	1820	1690	1722	1966	2052	2152	1838	1828	1710	1808	1756	1811	1562	963
Lower Saxony	1805	2102	2319	2149	1544	1211	1289	1224	1096	1111	1049	1187	1073	1049	1077	815
Mecklenburg-Vorpommern	109	135	226	169	262	309	307	295	334	330	343	375	362	362	329	296
North Rhine-Westphalia	8635	9349	9250	9083	9092	9341	9473	8957	9277	8859	8829	8213	7597	6983	7119	4724
Rhineland-Palatinate	1428	1336	1185	1255	1104	1190	1041	820	914	827	911	772	771	653	709	379
Saarland	327	276	221	195	705	582	450	479	437	428	442	393	425	396	420	310
Saxony	198	188	221	556	0	0	0	0	0	0	0	0	0	0	0	0
Saxony-Anhalt	371	345	323	235	1179	1247	1256	1196	934	905	865	799	725	646	574	402
Schleswig-Holstein	611	611	578	463	216	253	197	237	326	320	496	590	591	552	761	559
Thuringia	106	106	110	99	0	0	0	0	0	0	0	0	0	0	0	0
Total	25803	27547	27371	24425	26690	26893	26709	25080	24580	23591	29436	28618	26133	25913	25801	18083

Table 1: Number of court decisions per year and State/Federal level.

publishes decisions via the `openjur.de` platform, which we are currently not allowed to scrape. Saxony and Bremen provide decisions as PDF files only. All other states and the federal government, fortunately, publish their court decisions as easily machine-readable raw texts. Most of the states do not provide any or only very few decisions from before 2010. Commendable exceptions are Hesse and North Rhine-Westphalia (4,125 decisions from 1949-1989 and 985 decisions from 1957-1989, respectively). Moreover, Lower Saxony, Rhineland-Palatinate, and Schleswig-Holstein have published isolated decisions as of 2000 and Baden-Württemberg as of 2002.

In addition to the federal and state databases, we have also accessed the privately run database of `judicialis.de`. It contains 147,531 court decisions dated up to 31.12.2009 (i.e. until the beginning of the publication by the federal government and the governments of the federal states). The scope of this material is limited though, since it comprises only decisions of the federal, fiscal, higher administrative, state labor, and the higher regional courts.

Our dataset does not contain all court decisions made by German courts since 2010, as the courts themselves select a subset of their decisions to be published. For example, the Federal Court of Justice (BGH) settled 9,786 proceedings in 2020. In stark contrast, the decision database of the federal government contains only 1,662 decisions of the BGH from 2020, i.e., at most 16.98% of the actual decisions. The percentage of published decisions of the lower instance courts is even smaller. For 2019, a total of 841 decisions made by local ordinary courts (the courts at the lowest level) were published, while 2,136,439 decisions were rendered in total (926,514 civil [15], 568,588 family, and 641,337 criminal[1]). For the regional courts, slightly more decisions (2,197) were published in 2020; the number rises to 4,120 for the higher regional courts. This effect is reinforced by the fact that lower court decisions prior to 2010 are only made available by some of the states. Consequently, the data may contain significant selection bias – where we assume that courts are more inclined to publish legally important rulings. For this study, however, we require the data to only be a representative sample of the technical language used in court decisions, not a representative sample of all court decisions with regard to their legal content. Thus, the potential bias is not problematic.

Structure of the Dataset. As of 31 January 2022, the dataset contains 552,369 court decisions. Table 1 shows the decisions by state or federal government and year, as well as the totals for the states, the federal level, and each year. The largest group (180,906) consists of decisions from North Rhine-Westphalia, followed by 130,928 from the federal level. If federal decisions remain out of consideration, one would expect the number of decisions to be distributed among the states according to the size of their population. However, this assumption does not hold: North Rhine-Westphalia accounts for about 21.56% of the population but about 42.93% of the non-federal decisions. On the other hand, Baden-Wuerttemberg represents 13.35% of the population but only 6.00% of the decisions. This points to a highly inconsistent practice in the publication of court decisions across German federal states.

Given the uneven distribution of the decisions, it seems possible that our following observations are special effects of the language used by judges in, e.g., North Rhine-Westphalia. Yet, taking into account the uniform textbooks, commentaries, law journals, etc. used throughout Germany [12], one can safely assume that the style of legal language does not differ between the states.

In terms of jurisdiction, 216,714 decisions (39.23%) originate from civil and criminal jurisdiction, 166,650 from administrative jurisdiction (30.02%), 51,955 from social jurisdiction (9.41%), 51,283 from fiscal jurisdiction (9.23%), 47.479 from labor jurisdiction (8.60%), 11,651 from constitutional jurisdiction (2.11 %), 6057 from the patent jurisdiction (1.10%), 576 from the lawyers' courts (0.10%). There is no obvious uneven distribution between the different jurisdictions.

Decisions in civil and criminal cases are dominated by the decisions of the Higher Regional Courts (110,632) and the Federal Court of Justice (53,515), while the decisions of the administrative jurisdiction are predominantly from the administrative courts (79,122) and higher administrative courts or tribunals (76,155) and only to a small extent from the Federal Administrative Court (11,373).

Preparation. Before creating our models, both corpora had to be preprocessed. To summarise, we tokenised both corpora using the Natural Language Toolkit (NLTK) [9], then removed stop words, lemmatised all words, and split compound nouns via CharSplit [16].

3 CREATING WORD VECTORS

We created two word2vec models – one for each of the two corpora – using Gensim's [14] implementation of word2vec. We learned word vectors of dimension 300, using a sliding window of size 5 and 50 iterations. While technically possible, word vectors for words that occur only seldom in the corpus are usually of bad quality, i.e., of low to non-existent informative meaningfulness. For a precise and reliable determination of the word vectors, we want to have words/lemmas to occur in as many different contexts of use as possible. Hence, for most of the following analysis, we will limit

[1]https://www.destatis.de/DE/Themen/Staat/Justiz-Rechtspflege/_inhalt.html, retrieved 25.01.2023.

ourselves to those lemmata that occur at least thrice in the corpus. The trained word vectors as well as all of our code are publicly available at `galvusdamor.github.io/judicialSemantics`.

4 QUANTITATIVE ANALYSIS OF THE DIFFERENCES

Given the sheer size of the learned models – two 300-dimensional vectors for each of the 217,324 lemmata occurring in both corpora at least thrice – the question arises as to how a meaningful comparison can be performed. As a starting point for an in-depth analysis, we need to compute which words are worth manual consideration when trying to understand the differences in associated meaning.

The direct comparison of the two word vectors for each lemma in both models is not possible. Word vectors describe the semantics of a word only relative to other word vectors from the same model and are thus only meaningful w.r.t. that model. Even two semantically identical models (e.g. two models trained on the same corpus) might have different word vectors. For a cross-model comparison, we therefore need to develop a more sophisticated indirect method.

Comparison of Similar Words. Word2vec is based on the distributional hypothesis. It assumes that the meaning of a word is determined by the context of its usage. This leads to the assumption that the distance of two word vectors represents their semantic distance [4], where distance is defined as the cosine similarity of the two word vectors [1]. Given two word vectors $\underline{w}$ and $\underline{v}$, their cosine similarity $\sigma(\underline{w},\underline{v})$ is $\frac{\underline{w}\cdot\underline{v}}{||\underline{w}||\cdot||\underline{v}||}$. $\sigma(\underline{w},\underline{v})$ is a real number between 1 (vectors are identical) and -1 (vectors are exact opposites).

With this metric, we can now attempt to quantify the difference between the two models. If the two models would express similar semantics for a word w, the cosine similarity of w to any other v would be roughly identical in both models. Kim et al. [7] use this idea for measuring the relative change in meaning of a word over time. Provided that the cosine similarity $\sigma(\underline{w},\underline{v})$ is actually the true semantic difference in both models, the difference between the cosine similarity of the two models $\Delta(w,v) = |\sigma_d(\underline{w},\underline{v}) - \sigma_s(\underline{w},\underline{v})|$ measures the difference in semantic distance between w and v. Fixing w, we can average over v to compute a score $S(w) = \frac{1}{|W|}\Sigma_{v\in W}\Delta(w,v)$, quantifying the average shift in meaning between the two models (similar to Kim et al. [7]). We propose $S(w)$ as a measure for the semantical difference w between the two models.

For our models, $S(w)$ ranges from 0.055 (“wovon”/“whereof”) to 0.117 (“bezüglich”/“with respect to”) with a mean and median of 0.07. Since we average over 217,323 other words v, the variation of the values of $S(w)$ is small. However, the variations are strongly centered in the range of 0.06 to 0.08: 210,692 of the lemmata fall in this range, while only 286 of the lemmata (0.13%) have an average distance greater than 0.09. Striking is the strong accumulation of terms that are considered to be typical for legal terminology (e.g. "antragsteller"/"plaintiff" with $S(w)$ of 0.113).

D-value of the Kolmogorov-Smirnov Test. Another way to quantify the differences between the models is via a statistical approach. If we assume that both models assign the same meaning to words, the similarities to all other words should be the same in both models. Since the learning procedure is stochastic, minor deviations will occur which in the limit will be normally distributed at the level of the learned word vectors. Consequently, for a fixed word w, the differences in similarities $\sigma_d(\underline{w},\underline{v}) - \sigma_s(\underline{w},\underline{v})$ should also be normally distributed. Due to the large sample size, traditional tests like Kolmogorov-Smirnov (KS) will not be able to detect a normal distribution, as the empirical distribution will not match a normal distribution closely enough. We, therefore, move to a relative comparison using the D-value of the KS test. The value of D lies between 0 and 1, with 0 indicating a perfectly normal distribution. For a significance level of $\alpha = 0.05$, the critical D-value for rejecting that the distribution is normal is $d_{crit} = \frac{1.36}{\sqrt{217,322}} = 0.0029$ [11]. For our model, the D-values per word w range from 0.000683 to 0.058218 (mean 0.007667 and median 0.006218). 24,909 lemmata have $D \leq 0.0029$, i.e., are likely normally distributed and thus have the same associated meaning in both models. In contrast, there are 49,843 lemmata for which $D \geq 0.01$ and only 24 lemmata with $D \geq 0.05$. Lemmata with a high D-value, e.g. “beibringen”, are most likely ambiguous and should be investigated qualitatively.

Formation of Word Pairs. Next, we focus on the similarity of word pairs. Our aim is to determine whether there are pairs of words w and v that are semantically similar (i.e. synonymous) in one model only. This would produce problems in understanding – as words that are intended to mean different concepts in one corpus may not be read as such in the other. Our idea is similar to the work of Kim et al. [7], who looked for pairs of words that have a high cosine similarity at one point in time but a low one at another.

We computed, for all pairs of lemmata, the absolute difference of the cosine similarities $\Delta(w,v)$. For only 13,196 pairs, $\Delta(w,v)$ is greater than 0.5. In 5,319 cases the cosine similarity is higher in the decision model, i.e., legal language does not differentiate their associated meaning, while standard language does. A notable part of these pairs are abbreviations. For, e.g., “SG”/“Sozialgericht” (EN: social court) and “TOP”/“Tagesordnungspunkt” (EN: item on the agenda), the abbreviations and full terms are similar in the decision model only. This indicates that the abbreviations are particular to legal language and might not be fully understood by laypersons. In 7,877 cases the cosine similarity is higher in the standard model, i.e., only legal language differentiates between the terms.

In the 13,196 pairs, a total of 9,107 words occur, 5,364 of them only in a single pair. We suppose that words occurring often in pairs with high $\Delta(w,v)$ carry some ambiguity in one model only. Most of these words are legal terms, with the words involved in most contrast pairs being “kläger” (215, EN: plaintiff), “klägerin” (159, EN: female plaintiff), “antragssteller” (155, EN: plaintiff in family/administr. procedure), and “unstreitig” (136, EN: not in doubt).

Translation Learning. If we assume that both models represent an internally consistent but different language, we can try to find a word-to-word translation between them. Technically, we are looking for functions $d(w)$ and $s(w)$ that map words from the standard to the decision model and vice versa. We are then interested in whether $d(w) = w$ or $s(w) = w$, i.e., whether words are translated to themselves. If so, it is plausible that these words have the same associated meaning in both models. If not, we can analyse the translation to discover patterns. We use the method proposed by Lample et al. [8] to perform unsupervised machine translation without the need for seed words, which would defeat our exploratory goal. It

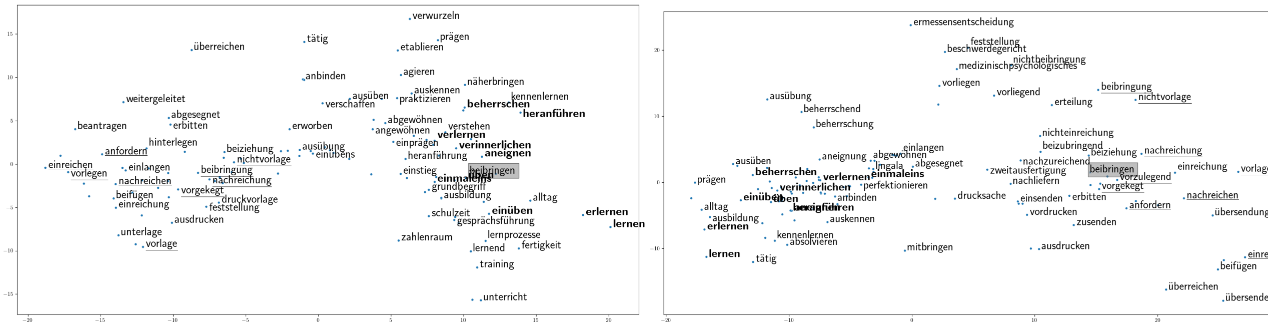

Figure 1: Embedding of the word "beibringen" in the standard (left) and decision (right) model (using PCA).

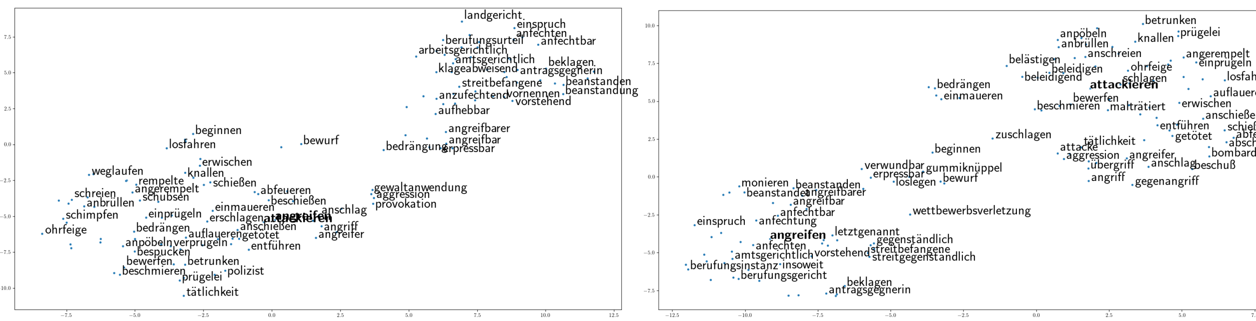

Figure 2: Contrast for the word pair "angreifen"/"attackieren" in the standard (left) and decision (right) model (using t-SNE).

solely relies on the word vectors of both models, embedding them into a common vector space. We have created the translation using the 150.000 most frequent words, as this achieved the best results.

For every lemma w we obtain two word vectors $\underline{w}^s$ and $\underline{w}^d$ in the common vector space – one for the standard and one for the decision model. The decision-to-standard translation $s(w)$ is then the lemma v for which $\sigma(\underline{v}^s, \underline{w}^d)$ is minimal. The standard-to-decision translation $d(w)$ is defined analogously. We can further compute for every word w a list $C^s(w)$ of words for which $\sigma(\underline{v}^s, \underline{w}^d)$ is the smallest, i.e., which are the words closest to the decision-to-standard translation of w. We will consider the 20 most similar words $C^s_{20}(w)$ as a reference for the semantically close words in the standard model. $C^d_{20}(w)$ is defined analogously.

Contrary to the standard-to-decision translation, the decision-to-standard translation does not translate many (10.416) of the very frequent words (first 25.000) to themselves. For 4.415 words we even have $w \notin C^s_{20}(w)$, which points to an actual mismatch in translation. Hereafter, we will consequently focus on the decision-to-standard translation. This inquiry is also of great practical interest, as it highlights problems laypersons might have when reading legal texts. Since our subsequent analysis involves manual work, we had to further reduce our scope to the 5.000 most frequent words in the decision corpus. 1.123 of those words are not translated to themselves. For 660 $w \notin C^s_{20}(w)$ holds, for 430 it does not. We manually classified the 1.123 translations into 10 categories.

Notably, only 185 of the 1123 translations are wrong or misleading. 155 are not fully incorrect, but also not helpful, e.g. "schreiben" (EN: to write) to "Brief" (EN: letter). In 38 cases, words are mapped to similar versions of the same general concept (e.g. "Mann" (EN: man) to "Frau" (EN: woman)), 44 are translations of a name to another name (e.g. "Berlin" to "Rostock", both are cities), 96 spelling/grammar variants (e.g. "bestand" (EN: existed) to "bestehen" (EN: to exist)), and 81 are artefacts from the corpora like enumeration markers (e.g. "aa") or abbreviations translated to other abbreviations (e.g. "isv" (EN: in the sense of) to "isd" (EN: in the sense of the)). 47 translations explain abbreviations (e.g. "BImSchG" as Bundesimmissionsschutzgesetz, EN: federal emission protection law)).

In multiple cases, a legal term was translated to a name (232) or a football term (16). This pattern indicates that some legal terms behave for a layperson as words without an inherent meaning, that are – just like names – used to denote a specific concept. We argue that these words possibly require more elaborated explanations to be understood by laypersons. Thus, the list of the relevant words, which can be extracted from our translations, may serve as a reference point for legal terms that need to be explained in detail.

In 226 cases, the translation is generally helpful. An interesting example of this is the translation of "billig", which usually means cheap. In a legal context, "billig" refers to the reasonable exercise of one's own rights. The decision-to-standard translation of "billig" is "anständig" (EN: decent, proper, ethical). Notably, the standard-to-decision translation of "billig" is "preiswert" (EN: cheap), showing awareness of the difference in meaning. Similarly, the word "beibringen" is translated to "nachreichen" (EN: to hand in later). The word "erklären", which means to explain in the standard language, is translated to "verkündet" (EN: to be declared), which accurately relates to the legal meaning of "erklären" – to declare formally.

5 QUALITATIVE ANALYSIS OF THE DIFFERENCES

Although the results presented so far provide quantitative insight into the differences between the two models, it remains difficult to draw concrete conclusions about specific differences in associated meaning between 'standard' and legal language. Since qualitative research that includes a sufficient number of experts and laypeople is expensive, our study presents graphical representations as an alternative to further investigate those differences. In line with the previous section, we introduce two types of representations, one focusing on a single word w and one on a pair of words $\langle w, v\rangle$ (each for both models). Our tool to generate these representations is available at `galvusdamor.github.io/judicialSemantics` together with additional examples.

Each figure shows a set of words W selected with regard to w or $\langle w, v\rangle$. We first compute a set of base words W^0 and enrich this set to provide more context by adding words that provide the 'semantic definition' for each word of W^0. For each $v \in W^0$, we determine for both models separately the N (a parameter) most similar words to v. Words that have a cosine similarity of ≥ 0.8 to a word already in the representation are ignored to avoid cluttering the graphic. This results in $2N$ words for every word v – denoted as $\mathcal{D}(v)$. We add the words $\mathcal{D}_N(v)$ to W^0 yielding W. We then display the word vectors of W using principal component analysis (PCA) or t-distributed stochastic neighbour embedding (t-SNE) [17] in order to reduce them from 300 dimensions to 2 dimensions.

Embeddings. We first focus on a single word w. To provide context, our next step is to select a set of words W^0 that, with sufficient contrastiveness, represent the meaning associated with w for both models. We select W_0 to comprise shared close and exclusive close words. Shared close words are words v that are close to w in both models. We use the five words $w_1, \dots, w_5$ that are most similar to w and that also occur in the list of the 5.000 most similar words in the model. Exclusive close words are words that are close to w in only one model. For each of the models, we take the ten most similar words that do not belong to the 5.000 most similar words in the other model. This results in a set W_0 of (up to) 31 words that we enrich with $N = 4$. As an example we selected the word "beibringen" – a layman's translation would be "to teach". The word "beibringen" has a D-value of 0.0068 and a $S(w)$ value of 0.331. Figure 1 shows both graphics using PCA to reduce the dimensionality. The base word w is surrounded by a box. Exclusive close words are underlined (for the decision model) or highlighted in bold (for the standard model). For "beibringen", there are no shared close words, which would otherwise be bold and underlined. The graphics immediately reveal a difference between the models. In the standard model, "beibringen" is semantically close to words that refer to the concept of learning, while in the decision model, "beibringen" is close to words referring to handing in additional documents or other type of evidence – which is not at all related to the meaning of "to teach". This example illustrates the capabilities of our method for clarifying differences in meaning.

Contrast Pairs. Now focusing on word pairs, our aim is to explain differences between two words v and w that occur in one model only. We determine $N = 7$ words $\mathcal{D}_t(\cdot)$ for both v and w that define them for both models and obtain a set W_0 of 15 words. If the words are close in one model but distant in the other, this definition will contain shared close words from one model and exclusive close words from the other, replicating the structure of the previous graphics for w. We then enrich W_0 using $N = 4$. Figure 2 depicts the result of this method for the word pair with the second highest $\delta(w, v)$, "angreifen"/"attackieren", using t-SNE. Both terms are commonly translated as "to attack". This is reflected in the standard model, where both words have almost the same meaning – that of physical attack. In the decision model, the figure suggests otherwise: the physical attack is linked to "attackieren", while "angreifen" refers to a mental process of objecting to something.

6 CONCLUSION

In a society governed by the rule of law, not only jurists but also the general public should be able to understand court decisions – especially if one is personally affected by them. The language of court decisions is, however, often complicated and hard to comprehend for laypersons. Our goal was to investigate the established hypothesis according to which one major source of this difficulty is ambiguity caused by legal jargon. To this end, we gathered a text corpus comprising over half a million court decisions. Relying on distributional semantics, we were able to present quantitative evidence for the assumed ambiguity by indirectly comparing the subsequently learned model with a 'standard' model. To make the results more tangible, we introduced graphical representations.

REFERENCES

[1] Fatemeh Torabi Asr, Robert Zinkov, and Michael Jones. 2018. Querying word embeddings for similarity and relatedness. In *NAACL-HLT*. 675–684.
[2] Gemma Boleda. 2020. Distributional Semantics and Linguistic Theory. *Annu. Rev. Linguist.* 6 (2020), 213–234.
[3] U. Quasthoff D. Goldhahn, T. Eckart. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *LREC*. 759–765.
[4] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv:1402.3722* (2014).
[5] Hanjo Hamann. 2021. Der blinde Fleck der deutschen Rechtswissenschaft. *JuristenZeitung* 76 (2021), 656–665.
[6] Sandra Hansen, Ralph Dirksen, Martin Küchler, Kerstin Kunz, and Stella Neumann. 2006. Comprehensible legal texts – utopia or a question of wording? On processing rephrased German court decisions. *Hermes – Journal of Language and Communication Studies* 36 (2006), 15–40.
[7] Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. 61–65.
[8] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised Machine Translation Using Monolingual Corpora Only. *arXiv preprint arXiv:1711.00043* (2017).
[9] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. *CoRR* cs.CL/0205028 (2002).
[10] David Mellinkoff. 1963. *The Language of the Law*.
[11] Patrick O'Connor and Andre Kleyner. 2012. *Practical reliability engineering*.
[12] Sebastian Omlor. 2022. Legal Research in Germany between Print and Electronic Media: An Overview. https://www.nyulawglobal.org/globalex/Germany1.html
[13] Lee Petherbridge and David L. Schwartz. 2015. An Empirical Assessment of the Supreme Court's Use of Legal Scholarship. *Northwestern University Law Review* 106 (2015), 995–1032.
[14] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).
[15] Statistisches Bundesamt. 2019. Rechtspflege Zivilgericht. Fachserie 10, Reihe 2.1.
[16] Don Tuggener. 2016. *Incremental Coreference Resolution for German*. Ph.D. Dissertation. University of Zurich, Faculty of Arts.
[17] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).