A Bilingual Longitudinal Analysis of Privacy Policies Measuring the Impacts of the GDPR and the CCPA/CPRA

Henry Hosseini
University of Münster
Münster, Germany
henry.hosseini@wi.uni-muenster.de

Christine Utz
CISPA Helmholtz Center for Information Security
Saarbrücken, Germany
christine.utz@cispa.de

Martin Degeling
Stiftung Neue Verantwortung
Berlin, Germany
martin@degeling.com

Thomas Hupperich
University of Münster
Münster, Germany
thomas.hupperich@wi.uni-muenster.de

ABSTRACT

Privacy policies are the main mechanism for websites to describe their practices in collecting and processing visitors' personal data. Their format and content are subject to legal requirements that have changed due to recent new privacy regulations including the General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA), and California Privacy Rights Act (CPRA). Studying how privacy policies are adapted to such regulatory change can help identify shortcomings in implementing the law and inform future legislatory initiatives. Existing work in this area mostly studied effects of the GDPR on privacy policies or the "Do Not Sell My Personal Information" link mandated by the CCPA. Methodologically, insights were mainly drawn from English-language privacy policies using keyword-based analyses or machine learning classifiers.

In this work, we address this research gap and conduct a bilingual study of privacy policies in English and German that investigates the effects of the GDPR and CCPA/CPRA on privacy policy content, using established methods from corpus linguistics that are language-independent and do not rely on keyword lists or classifiers that may date quickly. We find that, unlike for the GDPR, the CCPA's requirements were not yet widely implemented when it first became enforceable but only with its amendment, the CPRA. Before that, websites used more than 60 variants of the "Do Not Sell" link instead of the mandated wording and did not prominently reference individual rights granted by the CCPA/CPRA. While companies outside California and the US did adapt their disclosures to the CCPA/CPRA, this was limited to English-language policies and did not spill over to policies in German. For GDPR enforcement, we find websites to increasingly rely on legitimate interests to justify data collection, raising concerns whether individuals' interests in the privacy of their personal information are still sufficiently considered.

KEYWORDS

privacy, privacy policy, GDPR, CCPA, CPRA, corpus linguistics

Proceedings on Privacy Enhancing Technologies 2024(2), 434–463

https://doi.org/10.56553/popets-2024-0058

1 INTRODUCTION

In the light of pervasive data collection through digital services, including mobile devices, the Internet of Things (IoT), and the modern Web, recent years have seen jurisdictions across the globe update existing or pass new privacy legislation. Three prominent examples are the EU's General Data Protection Regulation (GDPR) [21] and, for the US state of California, the California Consumer Privacy Act (CCPA) [71] and its extension, the California Privacy Rights Act (CPRA) [72]. Albeit quite different in scope and approach, the common goal of these laws is to create higher standards for the protection of personal information in an increasingly interconnected environment. Regulatory instruments for this include the requirement of a legal basis for data collection, transparency mechanisms that require companies to disclose their data processing practices, and providing people with individual rights regarding how companies process and use their personal information.

New privacy legislation coming into force provides researchers with the unique opportunity to study how service providers adapt to such regulatory change, to identify obstacles towards compliance, and to provide regulators with insights for future regulatory efforts. On the Web, the established approach of websites to inform about their privacy-related practices and let visitors acknowledge them are privacy notices, such as privacy policies and consent notices. When new privacy regulations such as the GDPR and CCPA became enforceable, companies updated their privacy policies to comply with new transparency requirements and inform customers about their data rights [15]. Techniques used by privacy researchers to identify updates in online privacy policies include statistical analysis of text features such as changes in sentence, word, and syllable counts [7, 44] or hashing of sentences and measuring their number of changes [15]. Approaches to measuring content change, such as changes in data retention in privacy policies, include machine learning and deep learning classifiers [44] or searching for keywords related to the enforced privacy regulation [4, 15, 76].

While most of this prior work concerns changes in privacy policies around the GDPR enforcement date, there is a notable lack of work regarding the effects of the CCPA/CPRA. Existing research in this area has focused on how websites implement the Act's requirement to allow Californians to opt out of the sale of their personal information [57, 75] and how people perceive these mechanisms [29, 57], while, to the best of our knowledge, a more profound

content analysis of post-CCPA privacy policies is missing. Additionally, the vast majority of existing work in privacy policy content analysis, including work not explicitly focusing on changes due to new legislation, only considered policies in the English language, which leaves privacy disclosure practices in large parts of the world underexplored. Only recently has the analysis of non-English privacy policies started to receive attention, with existing work either exploring privacy policies at one or two points in time [5, 39] or focusing on descriptive statistics and the prevalence of specific key phrases [15] or corpus creation and annotation [2, 14].

We address this research gap by conducting a diachronic bilingual content analysis of privacy policies in English and German, as these are the two most widely spoken languages in Europe as a first or second language [18] and we are familiar with both. Using established natural language processing (NLP) techniques, we revisit the enforcement of both the GDPR and the CCPA/CPRA and investigate how the content of websites' privacy policies has changed as these regulations became effective or enforceable, examining three corpora of privacy policies in English and German to assess how new privacy laws affect privacy disclosures on the Web. In summary, we contribute to privacy policy research as follows:

- We investigate how the content of privacy policies changed under the GDPR and the CCPA/CPRA after they became enforceable (May 25, 2018 for GDPR; July 1, 2020 for CCPA) and effective (January 1, 2023 for CPRA) by examining wording, phrases, association strength, and legal terms. Applying the same methods to multiple privacy regulations coming into effect allows for direct comparison of their effects.
- We provide further insights into CCPA adoption over time by analyzing variants of the "Do Not Sell" link and their evolution. We also provide first insights into how websites adapt to the CPRA.
- To foster multilingual privacy policy analysis, we study the effects of privacy laws on privacy policy content in the two most frequently spoken languages in the EU, English and German, in a longitudinal analysis. While we find that policies in both languages adapt the terminology of new privacy laws, German policies more explicitly refer to concrete provisions, while English policies are more descriptive.

On a methodological level, our results show that established NLP methods such as keyness analysis and topic modeling are well suited to identify prominent topics in both English and German privacy policies, as well as how they evolve over time, especially in reaction to new legislation. Our approach complements prior approaches to privacy policy content analysis, while not relying on lists of key phrases or deep learning models that would need to be updated or retrained if the regulatory environment changes. At the same time, our analysis supports the development of deep learning models by providing them with results for comparison with theirs.

2 BACKGROUND

As background for our work we outline the privacy laws of interest and the provisions expected to impact websites' privacy policies.

Legislative goals and process. In the EU, the General Data Protection Regulation (GDPR) [21] was the first in a proposed series of regulations to update and harmonize privacy laws across member states on a high level. It was passed in 2016 and became effective on May 25, 2018. The US state of California introduced the California Consumer Privacy Act (CCPA; Section 1798.100 of the California Civil Code [CCC] [71]) to strengthen the data privacy rights of consumers within state boundaries. It was passed on June 28, 2018, became effective on January 1, 2020, and became enforceable on July 1, 2020. Its guarantees were later expanded by the California Privacy Rights Act (CPRA) [72], a California ballot proposition approved on November 3, 2020, and took effect on January 1, 2023. Its enforcement was scheduled for July 1, 2023, but was delayed by a Californian superior court until at least March 29, 2024 [51]. Despite their common goal of protecting the personal information of individuals, the CCPA/CPRA and the GDPR differ in several important points [35, 58], including the scope of protected and regulated parties, the definition of personal data and the lawfulness of its processing, individual rights, and the dimension of fines.

Regulated entities. The GDPR and the CCPA/CPRA use similar characteristics to define the entities protected and bound by the respective laws but with differences in terminology and scope. The GDPR distinguishes between "data subjects," "data controllers," and "data processors." *Data subjects* are already identified or identifiable natural individuals regardless of their residency. According to Article 4 GDPR *data controllers* decide on the reasons for which and the methods by which personal data is processed. The *data processor*, usually a third party, processes personal data on behalf of the data controller and according to their instructions. The CCPA/CPRA uses the term "consumer" for the data subject, with the difference of including only Californian residents, while data controllers and data processors are "businesses" and "service providers," respectively.

Applicability. According to Article 3, the GDPR applies to both data controllers and processors if they are established in the EU and process personal data or, if established only outside the EU, they offer services to data subjects in the EU. On the contrary, according to its Section 1798.140(c)(1), the CCPA/CPRA does not bind all businesses and service providers but only those who (1) do business in California (2) with Californian residents and (3) either (i) buy, sell, or share the personal data of at least 100,000 consumers or only collect the personal data of at least 50,000 consumers, or (ii) had a gross annual revenue of at least US $25 million in the preceding year, or (iii) generate at least 50 % of their annual revenue from selling or sharing personal information.

Permissibility of data collection and processing. Both laws fundamentally differ in their approach to under what conditions they allow the processing of individuals' personal information. Under the GDPR, the processing of personal data is only lawful if at least one of the six legal bases in Article 6 GDPR positively applies, two of which are freely given, specific, and unambiguous consent to the data processing and necessity for the data controller's legitimate interests. By contrast, the CCPA follows an opt-out approach in its Section 1798.120 by providing Californians with the right to opt out of the sale of their personal information. Section 1798.135 establishes that consumers must be made aware of this right through *"a clear and conspicuous link on the business's Internet homepage, titled 'Do Not Sell My Personal Information,' to an Internet Web page that enables a consumer [...] to opt out of the sale of*

[their] personal information." Further, consumers' associated rights under this section must be described in an online privacy policy or *"[a]ny California-specific description of consumers' rights"* (Section 1798.135(a)(2) CCC). As "sell" is defined in Section 1798.140(t)(1) as communicating a consumer's personal information *"for monetary or other valuable consideration,"* a business sharing consumers' personal data for any benefit can be understood as a sale. Thus, this provision widely requires commercial websites to provide a "Do Not Sell" link and associated disclosures in their privacy policy. The CPRA amended Sec. 1798.135(a)(1) CCC to require the wording "Do Not Sell or Share My Personal Information." It also introduced a new right to limit the use of *sensitive personal information* (SPI), including racial origin and ethnicity, religious, political, and philosophical beliefs, sexual orientation and activity, financial information, and health status and history. Companies must publish a second link on their home page titled, "Limit the Use of My Sensitive Personal Information," which can be combined with the "Do Not Sell or Share" link into a "single, clearly-labeled link."

Transparency requirements. The GDPR also lays down extensive transparency requirements for processors of personal data. Article 12 poses that data subjects need to be informed about the processing of their personal data *"in a concise, transparent, intelligible, and easily accessible form, using clear and plain language."* Article 13 specifies what information needs to be provided, including contact information, the purposes and legal basis for the processing, and the data subject's rights regarding their personal data. As IP addresses are considered personal data under the GDPR [20] and websites typically store them at least temporarily in web server logs, the requirement for a privacy policy and associated disclosures under the GDPR widely applies to websites.

Individual rights. The GDPR and the CCPA/CPRA grant individuals certain rights regarding knowledge and control of how companies use their personal data. These include the "right to know" what data companies have collected about them and the "right of deletion" of collected and processed data, which grants the "right to be forgotten." The GDPR's "right of rectification" initially did not appear in the CCPA but was introduced by the CPRA. It allows affected individuals to ask data controllers to correct inaccurate information or complement incomplete personal data.

3 RELATED WORK

In this work, we build upon earlier findings from web privacy measurements and privacy policy analysis to create a corpus of website privacy policies and study how their content evolved in response to the GDPR and the CCPA/CPRA.

Privacy policy analysis. Previous work has extensively studied online privacy policies, including their prevalence [55], readability [48, 64], and user perception [44]. Recent research in this area has focused on automated content analysis, extraction, and summarization of data practices using natural language processing (NLP) and machine learning (ML) techniques [5, 31, 45, 73, 78], with some focusing on longitudinal aspects [1, 76]. One particular challenge is the high frequency of changes, which makes it challenging to trace the evolution of privacy policy content over time. Our work contributes to overcoming this challenge by comprising larger, bilingual corpora and extensive topic modeling.

Effects of privacy laws on privacy disclosures. Other work that more specifically focused on the effects of new privacy legislation on privacy policies was conducted when the GDPR came into effect in 2018. Degeling et al. [15] monitored changes in privacy policies on European websites over the course of 2018 and found an average increase in the prevalence of privacy policies of 4.9 % and an increase in the average length of 18.0 %. Content-wise, they identified an increase in the prevalence of GDPR-related terminology, especially that related to user rights and legal bases of processing, but did not conduct a more thorough content analysis. Linden et al. [44] analyzed changes in presentation, textual features, coverage, compliance, and specificity of 6,278 English-language privacy policies between January 2016 and May 2019. They confirmed an increase in average length and also found improvements in user experience, topic coverage, and specificity, though most of these improvements only concerned policies targeted at an EU audience. Wagner [76] conducted a longitudinal analysis of around 50,000 privacy policies from 1996 to 2021, using archival data and methods from ML and NLP to study data practices and the rights granted to users and reserved for companies over time. While she found some types of personal data to be less often collected after the introduction of the GDPR and the CCPA, there was an increase in the collection of location and implicitly collected data, as well as data sharing with unnamed third parties, and website visitors often lack a meaningful choice in how their personal data is used.

Multiple studies measured the prevalence of cookie consent notices as a more recent transparency mechanism for a website's data processing practices. They found that many notices do not offer sufficient choice to deny data collection, do not have a backend that properly implements the visitor's selection, or use dark patterns to nudge visitors into consenting to all data processing [15, 47, 56, 74].

Effects of the CCPA on websites. The CCPA's requirement to provide a "Do Not Sell" link (see Section 2) was among the first effects of this law to be investigated by web privacy research. O'Connor et al. [57] manually and automatically analyzed popular US websites in July 2020 and January 2021 for how implementations of this requirement evolved after the CCPA became enforceable. They already found deviations from the mandated wording of the "Do Not Sell" link, which they partially attributed to deceptive purposes, but unlike this work they did not track the evolution of specific wordings over time. They also noticed widespread use of dark patterns to make the "Do Not Sell" link less visible on websites and, through two user studies, found that these techniques decreased interaction rates and hindered website visitors from exercising their right to opt out of the sale of their personal information. Van Nortwick and Wilson [75] measured the prevalence and implementation of "Do Not Sell My Personal Information" items on 497,870 English-language websites from the Tranco website ranking and a list of domains known to be third-party trackers or advertisers. They found "Do Not Sell" links on 9,838 sites, with a slow increase in adoption between July/August and November/December 2020, and partially attributed the low adoption rates to the CCPA not applying to the majority of websites (see Section 2). After the initially permissible alternative wording "Do Not Sell My Info" had been removed from

the CCPA proposal in December 2020, the study found that only a few websites had updated their "Do Not Sell" links accordingly. The links were found to be often placed in website footers, where they are poorly visible and/or hidden from non-Californian visitors via dynamic link hiding. Our work adds to these findings by also investigating the prevalence and structure of alternative wordings for the "Do Not Sell" link and their evolution over time, including the new wording mandated by the CPRA.

Proposed approaches other than a link to implement the "Do Not Sell" requirement include icons [29] and browser-based mechanisms [26, 79]. For the latter approach, Global Privacy Control (GPC) [26], the California Attorney General has expressed that websites are legally obliged to treat the GPC signal sent by the browser as a "Do Not Sell" request under the CCPA [70].

Studying the effects of the CCPA beyond this requirement, Chen et al. [12] analyzed 95 privacy policies from popular websites across the United States to evaluate the clarity and effectiveness of CCPA disclosures. They concluded that information relevant to the consumer was often obfuscated and unclear.

4 APPROACH

In this work, we address these research gaps by conducting a bilingual diachronic analysis of how the GDPR and the CCPA/CPRA affected privacy policies on the Web. We examine and compare their content regarding textual characteristics and topics. This section provides an overview of our study design. We describe our preliminary CCPA study, which intended to give a first impression of how websites adapt to this law, followed by the methods to perform our main analyses for an in-depth investigation of the GDPR's and CCPA/CPRA's effects on privacy policy content. Figure 1 illustrates the used methods and data corpora.

4.1 Preliminary CCPA Study

In September 2019, we conducted a pre-study to understand if and how websites were preparing to adapt to the CCPA and whether they already contained CCPA-related privacy mechanisms and disclosures. For this pre-study, we investigated a combined set of domains from two different website rankings to get a broad first impression of websites' privacy practices: To account for local developments, we used the Alexa top list of popular websites [3] from September 2019, as Alexa provided website popularity rankings by region.[1] We added to the pre-study domain set the 500 most popular websites for the US states of New York and California, as well as those for Germany, Australia, India, and Israel. These specific regions were selected for their geographic variety and economic strength. To account for global popularity, we added to the pre-study set of websites the top 500 domains on the Tranco ranking [41] from September 27, 2019 (ID: 3QNL). At that time, Tranco had started to evolve into the most popular website ranking for research purposes. To simulate a resident in the location of each country-specific domain top list, we connected to VPN servers in the respective regions. Our setup on a university server automatically established a connection to the VPN server, performed website scraping, and disconnected immediately after that task was finished. We used the Open Web Privacy Measurement (OpenWPM) framework [19] to crawl popular websites for privacy policies, CCPA-related web pages, and homepages. The homepages were searched for policy links with keywords pointing towards privacy policies in English and German. To identify links hinting at California privacy notices, we searched for URLs containing a combination of "California" and "privacy" or "California" and "right" in German and English. We filtered the downloaded web pages for duplicates, leading to a data set of 11,559 web pages belonging to 2,523 domains. The results of this preliminary study are described in Section 5.1.

[1] Amazon retired Alexa Internet's services on May 1, 2022, but past URLs providing the top lists by country can be accessed via the Internet Archive: https://web.archive.org/web/20190916195153/https://www.alexa.com/topsites/countries/.

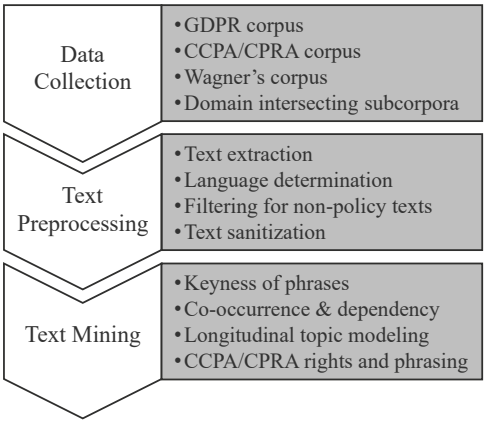


Figure 1: Overview of the used methods.

4.2 Data Collection

Our main analyses use three different corpora, as described below.

4.2.1 Analyzed Corpora. Investigating the impact of privacy regulations requires longitudinal corpora of privacy policies. In each of our analyses, we use distinct privacy policy corpora and/or domain popularity rankings, depending on the focus of the analysis.

GDPR Corpus. This multilingual corpus by Degeling et al. [15] provides a longitudinal snapshot of websites' privacy and cookie policies before and after the enforcement of the GDPR and was created to find evidence for GDPR-related changes on websites. For 28 European countries, the 500 most popular domains according to the Alexa website ranking were visited 15 times in the period between December 2017 and December 2018. Each month, one crawl was conducted, except for May, when three crawls were conducted to capture GDPR enforcement effects in a more fine-grained way. The websites' homepages were searched for links containing specific keywords that typically occur in the links of privacy and cookie policies. The raw data set consists of 127,328 web pages with privacy statements in 24 different languages and provides the basis of our main study regarding GDPR effects.

CCPA/CPRA Corpus. To investigate CCPA-related modifications, we selected the most popular 100K domains of the research-oriented Tranco list [41] as of November 5, 2019 (ID: GVWK). From December 2019 to July 2020, we performed a total of 15 crawls, visiting each of these domains and downloading their website's homepage as well as any identified privacy or cookie policy. In a second set of

crawls in February 2021 we visited the homepages of the top 10K Tranco domains using the list from January 31, 2021 (ID: WQW9) and downloaded their privacy and cookie policies. To capture the privacy policy landscape after the CPRA had taken effect, we conducted a third set of crawls in January and February 2023, revisiting the top 100K domains on the Tranco list from December 23, 2022 (ID: 829V9). For all crawls, we used a server located in California to simulate the geolocation of Californian residents instead of connecting to VPN servers. Overall, the crawls resulted in a raw data set of 1,458,802 privacy and cookie policy pages and 1,309,003 homepages as a basis for our main study of CCPA/CPRA effects.

Wagner's Corpus. The longitudinal corpus of privacy policies by Wagner [76] consists of 645,124 English-language privacy policies from between December 1996 and 2021, collected using the Internet Archive's Wayback Machine. The domains of this corpus were selected by combining the top 1K domains and randomly selected domains ranked between 1K and 10K from the Tranco lists from October 1, 2019 (ID: JL9Y) and March 31, 2021 (ID: ZLZG), as well as the Alexa top 1K domains for 2010–2021, Alexa top 500 between 2003 and 2009, and Alexa top 100 for 2002, resulting in a total of 4,997 domains. We use this corpus to compare and validate the results of our study on English privacy policy texts. For consistency, we extracted the privacy policies from this corpus that covered the same periods of time as each of the GDPR and CCPA/CPRA corpora, which yielded 101,181 privacy policies.

4.2.2 Domain Intersection. We detected that the Tranco lists we used to create the CCPA/CPRA corpus were subject to fluctuations, i. e., domains did not consistently show up in the ranking over time. We found only 48.5 % of the top 100K domains on the used Tranco lists (see Section 4.2.1) of to be persistent (see Appendix F). To conduct a thorough longitudinal analysis, the investigated domain sets must be cleaned of such fluctuations. Hence, we created *intersecting corpora*, subsets for each corpus and language that only contain privacy policies of policy domains present at all points in time comprised by the respective longitudinal corpus. This procedure resulted in 479 and 138 intersecting domains for the English and German privacy policies of the GDPR corpus, respectively. Similarly, 1,946 and 42 intersecting domains were identified for English and German privacy policies in the CCPA/CPRA corpus. We applied the same method to the Wagner corpus, resulting in 542 and 655 intersecting domains for each time frame of collection of the GDPR and CCPA/CPRA corpora. Appendix C lists the top intersecting policy domains based on the Tranco list from December 22, 2022 (ID: 82V9V), which we used for our crawls in early 2023. While this process of cleaning fluctuations reduced the size of the corpora, it ensured the consistency of data over time for longitudinal analyses.

4.3 Text Preprocessing

For all corpora, we applied the best practices for privacy policy preprocessing identified in our earlier work [34]. Following these, we used the Boilerpipe text extractor with the NumWordsRulesExtractor setting [38] to obtain the plain text of privacy policies from web pages, determined the languages of the texts by applying a majority voting scheme on the results of multiple language detection libraries, and identified non-privacy policies by applying trained classifiers [34] that achieved F1 scores of 99.1 % and 99.8 % for English and German. For each data collection time point, we manually inspected a random sample of $\tilde{1}0$ % of the downloaded policies and the final output for correctness and found no issues.

Table 1: Number of unique privacy policies in each corpus.

Corpus	All Domains		Only Intersecting	
	English	German	English	German
GDPR	27,151	7,878	12,016	2,763
CCPA/CPRA	784,561	33,913	78,418	1,242
Wagner's	101,181	0	67,638	0

Previous research has shown that segmentation of legal texts requires more sophisticated approaches, as standard NLP toolkits are challenged by the complex structure of privacy policies [25, 68]. We obtained the best qualitative results for tokenization and part-of-speech tagging for both English and German privacy policies from the SoMaJo library [62] combined with its part-of-speech tagger SoMeWeta [61]. Manual inspection showed that they performed best among the tested NLP toolkits in (1) correctly stripping excessive punctuation, such as bullets of a bulleted list concatenated with the first token of a list item, (2) handling of punctuation symbols in references to laws and regulations, and (3) not splitting tokens containing intra-term hyphens by default.

For further sanitization, we used the Spacy library [52] for lemmatization and replaced email addresses, phone numbers, and URLs in the policy texts with placeholders using Textacy [17] and the token tags of SoMaJo and SoMeWeta. Privacy policies may also contain the names of brands and organizations. To remove bias, we replaced them with a placeholder. We identified these names using the named entity recognition (NER) functionality of Spacy [52], Stanza [63], and Flair [43] cumulatively and manually excluded falsely identified names. Appendix G shows examples for the preprocessing step. Finally, we removed duplicate policies for each data collection time point. Appendix L depicts the composition of the original and final sanitized corpora. The final number of privacy policies in English and German and the number of privacy policies for the intersecting domains in each corpus are shown in Table 1.

4.4 Text Mining

After preparing the corpora, we mined the privacy policy texts for keyphrases, co-occurrences, wordings, and relevant topics.

4.4.1 Keyness Analysis. Keyness analysis aims to identify terms that stand out while comparing corpora. In corpus linguistics, the compared corpora are referred to as reference corpus (R) and target corpus (T). This analysis is performed by measuring whether the frequency of a term in the target corpus stands out statistically compared to its frequency in the reference corpus. In other words, the null hypothesis H_0 is defined as there being no difference between the frequency of a term in the compared corpora. The typical statistical measure for this comparison is the log-likelihood ratio (G^2) value, computed as follows [9]:

$$G^2 = 2 \times \left(\mathrm{O}_{11} \times \ln \frac{\mathrm{O}_{11}}{\mathrm{E}_{11}} + \mathrm{O}_{21} \times \ln \frac{\mathrm{O}_{21}}{\mathrm{E}_{21}} \right)$$

For our keyness analysis, we measure n-grams, with n ranging from 1 to 5. Occurring n-grams are counted only once per policy text, i. e., we consider n-gram types and not n-gram frequencies. O_{11} and O_{12} refer to the observed frequencies of an n-gram in the target and reference corpus, respectively. With w referring to an n-gram, E_{11} and E_{21} are the expected n-gram frequencies in the target and reference corpus and are calculated as follows:

$$\mathrm{E}_{11} = \frac{\text{n-grams in C} \times (\text{ freq. of } w \text{ in C} + \text{ freq. of } w \text{ in R})}{\text{total no. of n-grams in C and R}}$$

$$\mathrm{E}_{21} = \frac{\text{n-grams in R} \times (\text{ freq. of } w \text{ in C} + \text{ freq. of } w \text{ in R})}{\text{total no. of n-grams in C and R}}$$

Since log-likelihood is sensitive to corpora of different sizes [77], we used the Bayesian Information Criterion (BIC), which is calculated as $BIC = G^2 - \ln(N)$, where N stands for the combined number of n-grams in both corpora. A BIC value larger than 2 ($p < 0.0018$) indicates positive evidence against H_0, while a value larger than 10 ($p < 0.0000024$) indicates very strong evidence against H_0. In case of the existence of very strong evidence, an n-gram is considered to be more associated with the target corpus if $O_{11} > E_{11}$ and the normalized frequency in the target corpus is higher. The normalized frequency refers to the frequency per million n-grams to ensure the comparability of corpora of different sizes. To remove noise for this analysis, we filtered n-grams starting and ending with connector words (e. g., then, too, …), containing placeholders for email addresses, phone numbers, or URLs, as well as n-grams with a lower normalized frequency than 10 per million. Unless stated otherwise, we report on n-grams with BIC > 10, indicating very strong evidence against H_0.

While statistical evidence of a difference in frequency is a necessary condition of keyness [77], it is insufficient to indicate prominence. While the BIC value indicates the presence or absence of statistical evidence against H_0, it is not a measure for effect size, i. e., the magnitude of difference between the normalized frequencies of an n-gram across the compared corpora [24]. Hence, to complement this metric, we calculated the log ratio to determine effect size [30]:

$$\text{Log Ratio} = \log_2 \frac{NF_{w,T}}{NF_{w,R}}$$

where $NF_{w,T}$ and $NF_{w,R}$ indicate the normalized frequencies of w in the target and reference corpus per million, respectively. Each additional point of the log ratio score signifies a doubling of the disparity between the two corpora for the considered n-gram. In case of the absence of a term, a tiny value (0.00000000000000000001) was considered instead. To provide the normalized rate difference of an n-gram in case of its absence in one of the compared corpora, we report the difference coefficient (DiffC), calculated as [33, 42]:

$$\text{Difference Coefficient} = \frac{NF_{w,T} - NF_{w,R}}{NF_{w,T} + NF_{w,R}}$$

The difference coefficient ranges between +1 (if an n-gram appears only in the target corpus) and -1 (if an n-gram appears only in the reference corpus). A DiffC of 0 indicates no difference in the normalized frequencies of an n-gram in the compared corpora.

Keyness analysis on corpora from the same field but different points in time can shed light on shifts in language and terminology usage. In our case we investigate such shifts with regard to the most significant terms associated with privacy policies before and after the regulatory regimes of the GDPR and CCPA became enforceable and the CPRA became effective. This required us to split our data into reference and target corpora at a specific point in time presumed to be a turning point in websites' decisions to adapt their privacy policies to regulatory change. Determining the turning point for this type of analysis is not trivial, as privacy policies might have changed several months before and after the enforcement dates of the respective laws. Previous work provides varying evidence of when websites started to adapt to new privacy legislation. For the GDPR, the points in time that saw the most changes in websites' privacy policies were found to be around the GDPR enforcement date in late May 2018 [4], one month before [15], and June 2018 [44]. Wagner's longitudinal study found a peak in the total number of unique privacy policy texts for 2020 and attributed this to updates due to the CCPA [76]. Despite these differences, the identified times of maximum change all hovered around the respective enforcement date, so we decided to use the final enforcement date of each regulation (or effectiveness date, if not enforced at the time of writing – GDPR: May 25, 2018; CCPA: July 1, 2020; CPRA: January 1, 2023) as the turning point to split our corpora into pre- and post-enforcement subcorpora for our analyses. In each comparison, the pre-enforcement subcorpus is the reference corpus, and the post-enforcement subcorpus is the target corpus. We used the CCPA/CPRA subcorpora from February 2021 and early 2023 as additional target corpora to present an updated picture of changes in the privacy policy landscape.

4.4.2 Co-Occurrence & Dependency Analysis. Previous work has identified terms commonly associated with privacy policy texts and terms defined in privacy regulations. These terms were either identified manually, e. g., by reviewing privacy policies and regulations [15], or semi-automatically via guided topic modeling [37] or unsupervised topic modeling followed by consulting domain experts [69]. Our analyses go beyond this and leverage a co-occurrence analysis from the field of corpus linguistics [9], which allows us to discover contiguous phrases with common terms in privacy policies such as "collect," "process" and "share" to identify data practices exclusively associated with these terms in privacy policies. In our study, this analysis is not only limited to identifying the most significant co-occurrences but also changes in statistical collocation strength in privacy policies over time. This requires statistical measures that a) measure the exclusivity of the co-occurrences, b) are independent of corpus size, and c) result in scores that make future analyses comparable with the current state. Therefore, we chose the log Dice score [67], which is calculated as:

$$\text{log Dice } = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

where f_{xy} is the number of co-occurrences of two words x and y in a predefined window of tokens and f_x and f_y are the number of occurrences of x and y in the corpus, respectively. The theoretical values of log Dice range between 0 and 14.

We parsed the privacy policies using Spacy to identify dependency bi-grams that are syntactically connected via a direct head-dependent relationship, i. e., the direct object of a verb. These direct objects consist of noun chunks instead of single nouns to add semantic meaningfulness. Examples of such head-dependent bi-grams are *collect_personal information* and *share_aggregated data*.

4.4.3 Topic Changes. The evolution of content in privacy policies over time is one of the less explored topics in privacy policy research. The traditional method to observe topic change over time is Dynamic Topic Modeling (DTM) as developed by Blei and Lafferty [8]. Drawbacks of DTM include the lack of consideration of the appearance or disappearance of new topics and the requirement to predetermine the number of topics (k) in the corpus. We experimented on each of our corpora to determine k statistically [6, 11, 16, 27] using the ldatuning package [54]. The privacy policies were preprocessed as described in Section 4.3 and segmented into subtopic passages using TextTiling [32]. The resulting number of passages per crawl and language are listed in Table 2. For each corpus and language, we trained 25 models using Latent Dirichlet Allocation (LDA), starting with 20 topics and increasing their number to 500. The resulting extreme values of four statistical tests on these LDA models converged between 440 and 500 topics as the optimal range, as shown in Appendix I. In the end, due to the aforementioned drawback of DTM, this traditional method was not able to infer a concrete number of topics and provide topical insights. Hence, we utilized a more suitable topic modeling method.

BERTopic [28] is an alternative to traditional topic modeling that does not require the number of topics in advance. In this method, each text is converted into vector embeddings using sentence-BERT [65], followed by reducing the dimensionality of these embeddings to cluster semantically similar texts. Dimensionality reduction and clustering are performed using the UMAP [50] and HDBSCAN [49] algorithms, respectively. To obtain the representative terms for each topic, a modified procedure for Term Frequency Inverse Document Frequency (TF-IDF) is applied that calculates the most important words per topic. This procedure, *class-based TF-IDF*, treats the passages inside a cluster as a single text. As a result, the score of each word x within a class (cluster) c is calculated as:

$$W_{x,c} = tf_{x,c} \times \log(1 + \frac{A}{f_x})$$

where tf refers to the frequency of word x within the cluster c, A represents the average number of words per cluster, and f indicates the frequency of word x across all clusters.

BERTopic allows for dynamic topic modeling that considers new topics appearing over time by first generating a general topic model. Then, the class-based TF-IDF representation is recalculated for each cluster c and time t. This way, topic representations at each point in time can be calculated without the need to train t separate models.

4.4.4 Measuring CCPA-related Terminology. The CCPA states in its Section 1798.135 that a link with the exact wording of "Do Not Sell My Personal Information" must be present on homepages and privacy policy pages of websites. In the corpora, we searched for specific items satisfying this unique requirement. As first investigations hinted at the absence of this phrasing, we crafted the regular expressions in Appendix A to capture similar wordings.

Table 2: Corpus stats on the number of privacy policies (PP), passages (Psg), and average (Avg) passages per policy. Only policies of intersecting domains over each corpus are listed.

Corpus	Crawl	English			German		
		PP	Psg	Avg	PP	Psg	Avg
GDPR Corpus	2017-12-06	724	21,058	29	168	3,588	21
	2018-01-22	726	21,448	29	166	3,523	21
	2018-02-26	703	21,521	30	172	3,742	21
	2018-03-26	755	24,591	32	175	3,862	22
	2018-04-24	757	25,362	33	173	3,911	22
	2018-05-07	778	26,336	33	173	3,992	23
	2018-05-18	778	26,798	34	171	3,972	23
	2018-05-25	832	31,914	38	183	7,370	40
	2018-06-28	841	32,098	38	190	8,124	42
	2018-07-18	874	33,400	38	198	8,873	44
	2018-08-21	832	32,761	39	196	8,562	43
	2018-09-21	845	32,223	38	197	8,919	45
	2018-10-12	866	33,101	38	198	9,004	45
	2018-11-30	842	31,289	37	201	9,211	45
	2018-12-28	863	33,095	38	202	9,265	45
CCPA/CPRA Corpus	2019-12-09	4,213	108,748	25	71	2,464	34
	2020-01-02	4,169	109,449	26	71	2,464	34
	2020-01-16	4,220	110,971	26	70	2,465	35
	2020-02-03	4,267	110,933	25	71	2,467	34
	2020-02-26	4,257	111,647	26	66	2,294	34
	2020-03-09	4,302	112,431	26	70	2,396	34
	2020-03-23	4,485	116,835	26	70	2,390	34
	2020-04-09	4,119	107,853	26	69	2,374	34
	2020-04-27	4,544	118,012	25	68	2,282	33
	2020-05-13	4,506	117,872	26	68	2,279	33
	2020-05-22	4,513	117,376	26	69	2,405	34
	2020-06-09	4,572	120,252	26	70	2,399	34
	2020-06-23	4,612	121,110	26	71	2,438	34
	2020-07-09	4,662	123,694	26	72	2,455	34
	2020-07-30	4,705	124,177	26	73	2,473	33
	2021-02-18	5,014	143,792	28	86	2,765	32
	2023-01-14	3,670	114,425	31	55	1,822	33
	2023-02-13	3,588	110,663	30	52	1,804	34
Wagner's Corpus	2017-12	2,627	64,689	25	n/a		
	2018-01	2,723	69,176	25			
	2018-02	2,721	69,528	26			
	2018-03	2,793	70,346	25			
	2018-04	2,676	66,754	25			
	2018-05	3,202	83,499	26			
	2018-06	3,079	86,057	28			
	2018-07	3,096	88,063	28			
	2018-08	3,162	89,949	28			
	2018-09	3,192	90,786	28			
	2018-10	3,165	89,430	28			
	2018-11	3,080	89,779	29			
	2018-12	3,123	90,585	29			
	2019-12	3,240	100,681	31			
	2020-01	3,066	100,331	33			
	2020-02	3,098	101,912	33			
	2020-03	2,876	95,281	33			
	2020-04	2,817	92,056	33			
	2020-05	3,516	117,598	33			
	2020-06	3,413	115,029	34			
	2020-07	3,432	116,255	34			
	2021-02	3,541	121,198	34			

5 RESULTS

In the following, we report on the results of the preliminary study, followed by the main study. We present the significant shifts in terminology, legal references, and user rights, as well as topic trends in privacy policies after the enforcement of the GDPR and CCPA and the evolution of CCPA/CPRA-related wordings on homepages. We also show the first effects of the CPRA in early 2023.

5.1 Preliminary CCPA Analysis

As described in Section 4.1, we accessed 2,523 domains from six specific locations to collect their privacy policies and search their homepages for "Do Not Sell" links. Overall, 8,305 privacy policies were retrieved, of which 488 – less than 6 % – contained CCPA-related disclosures, as determined by the presence of the string "CCPA". Table 3 provides an overview of the number of collected privacy policies by VPN server location and how many of them included CCPA disclosures. We cannot observe any strong trend that website visitors from California would see CCPA-related content in privacy policies more often than non-Californians.

Table 3: Privacy policies collected from 2,523 domains in the preliminary CCPA analysis, categorized by access location.

	Germany	California	New York	Australia	Israel	India	Total
Privacy Policy	1,452	1,399	1,388	1,246	1,431	1,389	8,305
CCPA Content	83	79	79	76	88	83	488

We also analyzed the presence of CCPA-related terminology in the privacy policies by websites' top-level domains (TLD). The highest prevalence was observed in privacy policies from `.com` domains, where 424 out of 5,111 policies (8.3 %) contained CCPA-related disclosures. 21 out of 442 (4.8 %) privacy policies with a `.org` TLD and 4 out of 224 privacy policies (1.8 %) with the `.co.il` TLD featured CCPA content. The remaining TLDs included `.de`, `.com.au`, `.us`, and `.ca.us`, whose privacy policies did not contain disclosures related to the CCPA.

Finally, we searched the domains' homepages for links containing the phrase "Do Not Sell My Personal Information." At the time of this pre-study in October 2019, none of the inspected homepages had a link with this exact wording, hinting at websites not having taken preparations for the CCPA back then.

> In October 2019, websites were not yet prepared for the CCPA: Although 5.8 % of the inspected privacy policies included CCPA-related disclosures, no homepage contained a link with the exact wording "Do Not Sell My Personal Information."

5.2 Keyness Analysis

As discussed in Section 4.4.1, the GDPR, CCPA/CPRA, and Wagner's corpora were split based on the enforcement / effectiveness dates of the GDPR, CCPA, and CPRA to find terms that occurred more often after the enforcement dates based on statistical evidence.

5.2.1 GDPR Enforcement. The analysis of the English GDPR corpus demonstrated an increase in the usage of phrases that refer to the individual rights of data subjects under the GDPR, which aligns with the findings of previous work [76]. Examples of such phrases are *restrict_processing*, *object_to_processing_of_personal*, *right_to_withdraw_consent*, and *rectification*. Furthermore, the right to data portability (Article 20 GDPR) is reflected in the increased frequency of *readable_format* and *machine_readable* after the GDPR enforcement date. The increased prevalence of phrases such as *compliance_with_legal* and *comply_with_legal_obligation* hints at compliance with the law becoming increasingly important for data controllers. In addition, the log ratio of phrases such as *perform_contract* (3.06), *legitimate_interest* (2.71), *base_on_consent* (2.18), *contractual_obligation* (0.94), and *legal_obligation* (0.67) provides evidence for how often data controllers ground their data processing on each of the legal bases for data collection or processing in Article 6 after the GDPR went into effect. For comparison against existing corpora, Table 4 lists further statistically significant phrases and their log ratio in the GDPR and Wagner corpora.

Comparing these insights with the most statistically significant increased phrases in the German privacy policies after GDPR enforcement paints a different picture. The most prominent phrases are references to individual GDPR. Examples include *article_16*, *article_17*, as well as many phrases containing *article_6* and its paragraphs and enumerated subcases; detailed statistics are included in Appendix B. German privacy policies directly referencing GDPR provisions and those in English using a more descriptive approach could be rooted in different legal traditions: The common law system prevalent in the English-speaking world has legal precedent as its main source of law, while the civil law system that governs, among other jurisdictions, Germany and much of Europe, focuses on legal codes. Thus, legal texts in German are more likely to directly reference a law's individual provisions. In addition, more than 70 statistically significant phrases in the German corpus included the term "process", e. g., *object to processing* (log ratio 6.98), *process restriction* (3.30), *legal basis for processing* (3.04), which reflects the increased importance of transparency and accountability of data controllers about data processing after GDPR enforcement in German privacy policies. In comparison, we found around 20 such statistically significant phrases in English privacy policies.

> After GDPR enforcement, German privacy policies more prominently referenced concrete GDPR provisions to provide a legal basis for data processing, while English privacy policies favored a more descriptive approach.

5.2.2 CCPA Enforcement & CPRA Taking Effect. The n-grams listed in Table 5 indicate the frequency changes between the pre-CCPA and the July 2020 subcorpora, the latter of which was collected

Table 4: Log ratio values of English phrases with statistically significant occurrence increase after GDPR enforcement. An extended version is included in Appendix K.

Phrase	GDPR	Wagner's
data_portability	3.60	3.08
restrict_processing	3.45	2.71
readable_format	3.38	3.14
supervisory_authority	3.29	2.42
general_data_protection_regulation	3.23	2.92
right_to_withdraw_consent	3.14	2.39
machine_readable	3.11	1.81
legal_basis	3.11	2.77
object_to_processing_of_personal	3.09	2.63
perform_contract	3.06	2.40
lodge_complaint	2.97	3.13
right_to_object	2.86	2.56
enter_into_contract	2.85	1.85
legitimate_interest	2.71	2.80
consent_to_process	2.64	2.42
right_to_receive	2.64	1.14
erasure	2.58	2.75

after the CCPA enforcement date of July 1, 2020. The log ratio values of the phrase *opt_out_of_sale* and *right_to_opt* indicate an increase in 26.1 and 20.4 percentage points, respectively, and are relatively small compared to the observed effect sizes after GDPR enforcement. The occurrence of *californian_resident* has increased by 14 percentage points, which is comparable to Wagner reporting a 20 percentage point increase in mentioning Californians [76]. In the German policies, we could not observe any terms with statistically significant changes after the CCPA enforcement date.

Table 5: Log ratio values of English phrases with statistically significant occurrence increase after CCPA enforcement.

Phrase	CCPA/CPRA	Wagner's
discriminate	0.41	0.25
commercial_information	0.40	0.19
opt_out_of_sale	0.34	0.19
california_consumer_privacy_act	0.33	0.27
do_not_sell_personal_information	0.30	0.16
right_to_opt	0.27	0.15
california_resident	0.19	0.09

Comparing the pre-CCPA and February 2021 subcorpora yields a similar picture, which also holds for the Wagner-2021 corpus: Changes in the occurrence of CCPA-related terms are either not statistically significant or their log ratios do not differ from those of the July 2020 comparison. However, the comparison of the phrases between the pre-CCPA policies and those collected in Jan. and Feb. 2023 shows a substantial statistically significant increase in the occurrence of the four CCPA consumer rights, as well as for the two consumer rights newly added by the CPRA, the right to correction of personal information and the right to limit the use of sensitive personal information. Table 6 compares the metrics for these rights between the pre-CCPA reference subcorpus and the July 2020, Feb. 2021, and Jan. & Feb. 2023 target subcorpora, showing a statistically significant increase in the occurrence of these phrases in Jan. and Feb. 2023. Values with a $BIC > 10$ lead to rejecting H_0 (no difference in frequency). The higher normalized frequencies in the target corpora (NFT) compared to the normalized frequencies in the reference corpus (NFR) and the positive difference coefficient values (DiffC) indicate a higher association with the three target corpora. This phenomenon could be due to the CPRA taking effect on January 1, 2023 and websites preparing for its enforcement, originally planned for July 1, 2023. These early adjustments to privacy policies to include consumer rights under the CPRA before its enforcement date indicate the willingness of businesses to comply with the CPRA.

CCPA consumer rights appeared significantly more often in English privacy policies in early 2023, especially the two consumer rights newly added by the CPRA. Such increases were not observable for CCPA consumer rights in July 2020.

5.3 Co-Occurrence & Dependency Analysis

Comparison of the co-occurrence strength in English privacy policies before and after the enforcement dates or, if not enforced at the time of writing, effectiveness dates of the GDPR and CCPA/CPRA as described in Section 4.4.2 revealed intriguing insights. Similarly to Wagner's work [76], we identified an increased co-occurrence of *collect_precise location data* in the January & February 2023 privacy policies. We additionally observed an increased exclusivity in co-occurrence strength for, e. g., *collect_voice data*, *use_algorithm-based technology*, and *personalize_child*, which would require more in-depth investigation. Due to space constraints we list the observed occurrences of common verbs in privacy policies in Appendix H.

Table 6: Keyness statistics for the phrases related to CCPA/CPRA consumer privacy rights in the English corpora.

Right	Corpus	Time Frame	Metrics				
			BIC	LR	DiffC	NFT	NFR
Opt-out	CCPA/CPRA	Jul. 2020	12.00	0.27	0.09	14.15	11.71
		Feb. 2021	-20.03	-0.02	0.00	11.59	11.71
		Jan. & Feb. 2023	190.14	0.64	0.22	18.22	11.71
	Wagner	Jul. 2020	-13.12	0.16	0.05	16.07	14.43
		Feb. 2021	0.24	0.26	0.09	17.31	14.43
Know	CCPA/CPRA	Jul. 2020	11.31	0.33	0.11	9.50	7.55
		Feb. 2021	-11.46	-0.24	-0.08	6.38	7.55
		Jan. & Feb. 2023	91.83	0.59	0.20	11.33	7.55
	Wagner	Jul. 2020	-10.85	0.23	0.08	9.97	8.49
		Feb. 2021	-0.11	0.33	0.12	10.70	8.49
Non-discrimination	CCPA/CPRA	Jul. 2020	-7.24	0.39	0.14	2.83	2.16
		Feb. 2021	-19.96	-0.05	-0.02	2.08	2.16
		Jan. & Feb. 2023	95.22	1.01	0.34	4.35	2.16
	Wagner	Jul. 2020	-11.33	0.46	0.16	2.52	1.83
		Feb. 2021	1.53	0.69	0.24	2.96	1.83
Delete	CCPA/CPRA	Jul. 2020	-13.88	0.19	0.07	5.42	4.74
		Feb. 2021	-12.10	0.26	0.09	5.70	4.74
		Jan. & Feb. 2023	9.64	0.40	0.14	6.25	4.74
	Wagner	Jul. 2020	-16.20	0.21	0.07	4.70	4.06
		Feb. 2021	-13.16	0.28	0.10	4.93	4.06
Correct	CCPA/CPRA	Jul. 2020	-15.07	0.29	0.10	1.98	1.62
		Feb. 2021	-11.13	0.46	0.16	2.22	1.62
		Jan. & Feb. 2023	322.22	1.72	0.54	5.34	1.62
	Wagner	Jul. 2020	-15.09	0.41	0.14	1.69	1.27
		Feb. 2021	-13.30	0.47	0.16	1.77	1.27
Limit	CCPA/CPRA	Jul. 2020	-18.64	0.22	0.08	1.00	0.86
		Feb. 2021	37.97	1.32	0.43	2.14	0.86
		Jan. & Feb. 2023	376.58	2.26	0.66	4.12	0.86
	Wagner	Jul. 2020	-19.47	-0.05	-0.02	1.14	1.18
		Feb. 2021	-19.33	0.10	0.03	1.26	1.18

BIC = Bayesian information criterion, LR = log ratio, DiffC = difference coefficient, NF = normalized frequency per million in T (target corpus) and R (reference corpus)

5.4 Topic Changes over Time

In Section 4.4.3, we described how we applied BERTopic to identify topic changes in privacy policies over time. Table 7 presents the number of topics identified by BERTopic in each corpus. The most likely reason for the difference in the number of topics between the GDPR and CCPA/CPRA English corpora and Wagner's corpus is that the latter was compiled using a combination of top-ranked domains from the Alexa and Tranco lists, while the GDPR and CCPA/CPRA corpora used only Alexa and Tranco, respectively. In addition, our CCPA/CPRA English corpus extends to early 2023 and includes more intersecting domains and privacy policies.

For each corpus and language, we first summarize the most frequently emerging topics, independent of the time aspect. Then

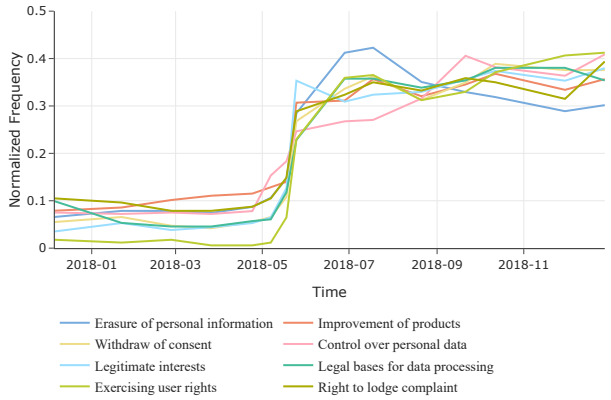


Figure 2: Topic trends in the English GDPR corpus.

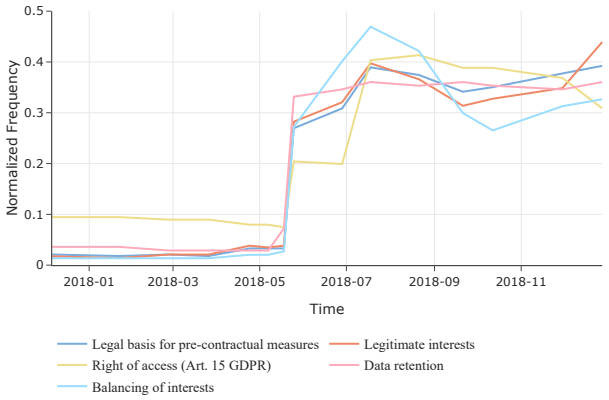


Figure 3: Topic trends in the German GDPR corpus.

we look at the topic distribution over time to highlight past and current topical trends in privacy policies. Each topic distribution is L2-normalized to allow for easier comparison of the magnitude of change over time between the topics. The trends in topics identified for Wagner's corpus are included in Appendix J for comparison.

Table 7: Number of topics identified by BERTopic in each corpus. The number of topics was set to *auto* and the number of minimum documents per topic to 20.

Corpus	English	German	Wagner's
GDPR	6,032	1,511	17,058
CCPA/CPRA	34,041	1,534	17,215

5.4.1 GDPR Enforcement. As previously described, the GDPR corpus was compiled using data from December 2017 to December 2018. Examining the top 10 topics in the English GDPR corpus by frequency of occurrence, these topics cover common broad privacy practices such as cookies and using, sharing, and protecting personal information, as well as more specific topics and data processing purposes such as interest-based advertising, improvement of products and services, and promotional emails and text messages. Looking at the effects before and after the GDPR enforcement date, Figure 2 depicts topics that follow a common trend. All of them had low occurrence prior to the GDPR enforcement date, which increased afterwards. One of these topics is "legitimate interests," which is one of the six legal bases of data collection and processing in Article 6 GDPR. What constitutes "legitimate interest" under the GDPR still requires interpretation by European courts and, consequently, still is the subject of recent research about deceptive design and potentially unfaithful data practices [36] five years after the enforcement of the GDPR [40].

In comparison, the overall top 10 topics in the German GDPR corpus differ and include cookies and their definition, opt-out cookies and links, the legal basis for pre-contractual data processing, and pseudonyms in user profiles. Companies' legitimate interests while protecting rights and freedoms of the affected individual are also among the trending topics after the GDPR enforcement date, as shown in Figure 3, along with the legal basis for processing personal data prior to entering a contract (Article 6(1)(b) GDPR). Reviewing the corresponding policy text clarified that companies use this legal basis to be able to communicate with customers and process their personal data prior to establishing a contractual relationship or to conduct credit investigations before providing financial services.

5.4.2 CCPA Enforcement & CPRA Taking Effect. The top 10 topics in the English CCPA/CPRA corpus, which covers the time between December 2019 and February 2023, reveal a different pattern. These topics include updates to privacy policies, security measures, the purpose of using information, functional cookies, third-party hyperlinks, and the potential requirement to disclose information to law enforcement authorities. Regarding the effects of CCPA/CPRA enforcement, we found that more topics in relation to these regulations have been included since December 2019. This supports our previous observation in Section 5.2.2, a significant increase of phrases in privacy policies referring to CCPA/CPRA-related rights. Figure 4 shows this upward trend for CCPA/CPRA-related topics that include core principles of these laws, such as the option to opt out of the sale of personal information, response time and format of verifiable consumer requests, and individual rights of Californians. A concerning trend is the continuous rise of referrals to "legitimate interests" also found in prior work [76].

The top 10 topics for the German privacy policies in the CCPA/CPRA corpus include the usage and definition of cookies; the rights to access, rectification, and limitation of data processing; the double opt-in process for newsletter registration via confirmation emails; and encryption of personal data sent through contact forms. Another top 10 topic regards opt-out cookies, whose occurrence has been decreasing since February 2021, as shown in Figure 5. The likely cause is the new German Telecommunications-Telemedia Data Protection Act (German abbr.: TTDSG) [10], which came into force in December 2021. Implementing the EU's ePrivacy Directive, Section 25 TTDSG only allows storing (or accessing already stored) information on an end user's device if the user has provided consent based on clear and comprehensive information, unless storing or accessing the information is, from a technical perspective, strictly necessary to provide a service explicitly requested by the user. The

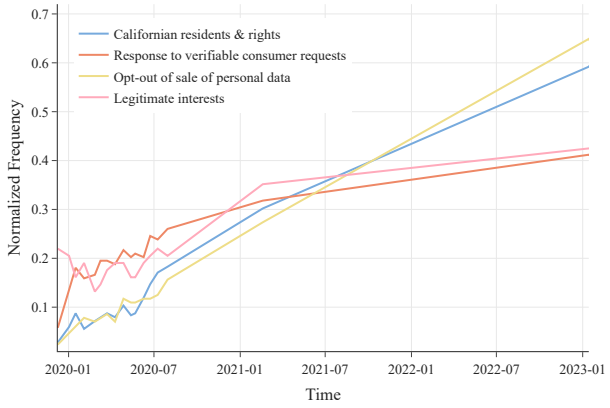


Figure 4: Topic trends in the English CCPA/CPRA corpus.

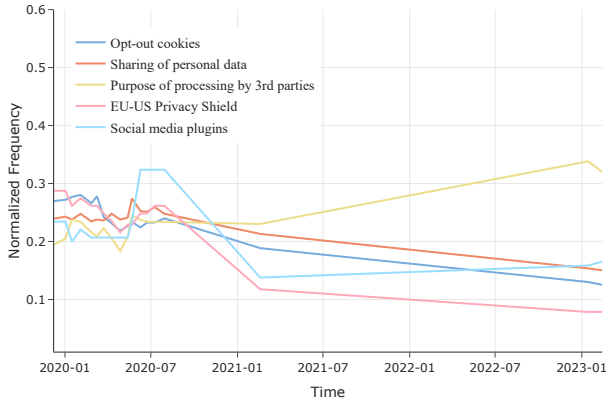


Figure 5: Topic trends in the German CCPA/CPRA corpus.

requirement for active, informed consent makes opt-out mechanisms not TTDSG-compliant [66], which explains privacy policies mentioning opt-out cookies less often.

> Passages referring to legitimate interests of data subjects have been trending in English privacy policies since 2018. German privacy policies have been mentioning opt-out cookies less frequently since February 2021, possibly due to the TTDSG, which became effective in Germany in December 2021.

5.5 "Do Not Sell" Link Over Time

In our main study, the search for "Do Not Sell" links on the homepages of the domains analyzed for the CCPA/CPRA corpus paints a different picture compared to the pre-study. Table 8 shows an increase in the appearance of "Do Not Sell My Personal Information" over time, particularly a monotonous increase of 1.85 percentage points from December 2019 to the second half of January 2020, and a total of 3.57 percentage points to July 2020. The most likely reason could be the initially declared CCPA enforcement date of January 1, 2020 and its postponement to July 1, 2020. Between July 2020 and January 2023, another 4.61 percent of the homepages included the wording required by the CCPA.

We also inspected the homepages for the link wording mandated by the CPRA, which added the aspect of *sharing*: "Do Not Sell or Share My Personal Information." Before 2023, no homepage in our CCPA/CPRA corpus contained this required link, while 1.85 % of homepages had added it in January 2023 and 3.12 % in February 2023. This indicates that website owners were gradually adapting to the CPRA's requirements and preparing for its enforcement.

Although many websites used to use different wordings for their "Do Not Sell" link in 2020, the prevalence of the exact wording stipulated by the CCPA/CPRA increased with each crawl, while nonstandard wordings gradually disappeared from homepages. We observed 60 wordings, listed in Appendix D, which differed in 1) use of acronyms or abbreviations ("PI" for "personal information" or "info" for "information"), 2) substitution of the term "information" with "data," 3) appending words referring to the legislation mandating this link or its applicability, such as "CA," "California," or "CCPA," and / or 4) capitalization of all characters.

> In early 2023, websites had started to adopt the CPRA wording for the "Do Not Sell" link, while back in 2020 they used 60 different wordings instead of the one mandated by the CCPA.

5.6 Global Effects of the CCPA/CPRA

As outlined in Section 2, the CCPA/CPRA could affect any company that does business in California or deals with Californians, even if it is based in another US state or country. Consequently, we were interested in how many companies in our CCPA/CPRA corpus resided in California or offered services to Californians from elsewhere and had prepared themselves for the CCPA/CPRA coming into effect. As prior work has identified "spillover effects" of the GDPR on privacy practices in other jurisdictions [5], this raised the question whether the CCPA/CPRA have also led to changes in privacy policies in languages other than English, in our case German. As a metric for whether a company had adapted its privacy disclosures to California regulations, we used the presence of "Do Not Sell" mechanisms. To determine where companies were based, we used the Free Company Dataset from People Data Labs [59], which contains metadata on 12 million companies in the world, including company name, website, country, and region. 3,138 of the 4,674 domains (67.1 %) in our English-language CCPA/CPRA subcorpus were listed in the Free Company Dataset. This does not indicate a shortcoming of this data set, as not all domains necessarily belong to companies. 509 of these 3,138 domains (16.02 %) were linked to companies located in California. None of the 56 domains in the German CCPA/CPRA subcorpus belonged to companies from California. The higher number of domains compared to intersecting privacy policy domains is due to companies owning multiple TLDs, such as Google, Blogspot, or ESPN. Redirects to privacy policies with different domains are also common [15].

Our analysis shows that companies with German privacy policies did not contain "Do Not Sell" links on their homepages or in their privacy policies at any time. However, many companies with English privacy policies inside and outside the US in early 2023 included "Do Not Sell" links on their homepages or declared not to sell or share personal information in their privacy policies. Outside the US, we identified 153 such companies from 35 countries around

Table 8: Prevalence and evolution of the most common wordings for the "Do Not Sell" link on websites' homepages in the CCPA/CPRA corpus over time. The full table listing all discovered wordings can be found in Appendix D.

	2019	2020														2023	
	Dec	Jan		Feb		Mar		Apr		May		Jun		Jul		Jan	Feb
Wording	09	02	16	03	26	09	23	09	27	13	22	09	23	09	30	14	13
Do Not Sell My Personal Information	0.57	1.25	2.42	2.51	2.37	2.60	3.08	2.28	3.35	2.24	2.15	3.19	3.32	4.14	4.14	8.75	7.08
Do Not Sell or Share My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.85	3.12
Do Not Sell My Info	1.09	1.39	1.64	1.52	1.83	1.81	1.81	1.87	1.78	1.75	1.86	1.73	1.84	1.99	2.10	1.14	1.00
Do not sell my personal information	0.02	0.14	0.33	0.33	0.35	0.23	0.31	0.29	0.57	0.29	0.27	0.48	0.48	0.45	0.47	0.79	0.95
Do not sell my info	0.21	0.22	0.28	0.40	0.42	0.46	0.45	0.46	0.48	1.04	1.13	1.25	0.52	0.54	0.68	0.54	0.42
Do Not Sell or Share My Personal Data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.39

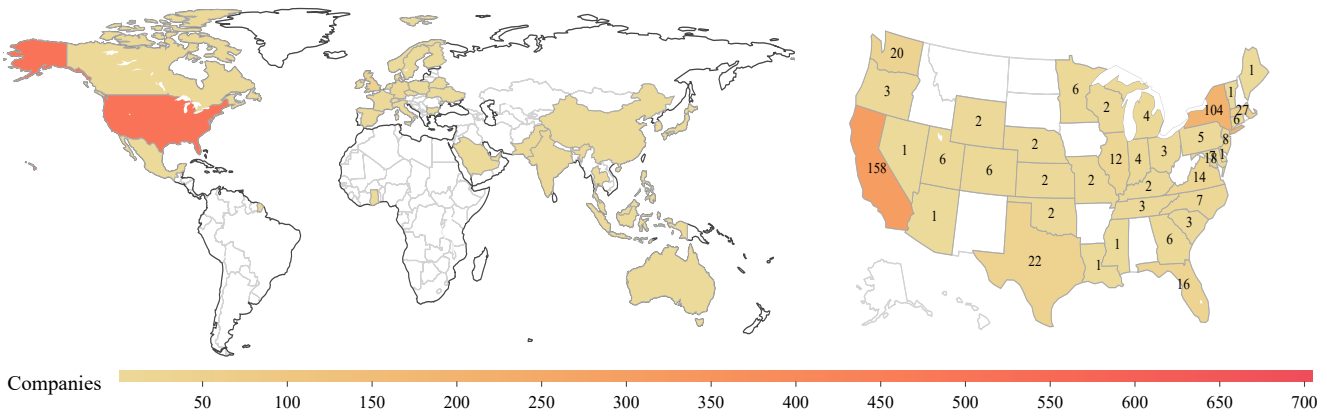


Figure 6: The distribution of companies by country in the world (left, N=153) and US state (right, N=489) whose websites' homepages or privacy policies contained "Do Not Sell" links or statements in early 2023.

the world, including 48 in the UK, followed by 17 companies in Germany, 14 in Canada, 11 in India, and 6 in France and the Netherlands, as illustrated in Figure 6. Most of them are in industries such as the Internet, computer software, IT and services, publishing, online media, and marketing and advertising. As previous work has suggested [75], these types of companies residing outside the US are more likely to target an international audience and, thus, to collect personal data from Californian residents, as opposed to, e. g., the real estate industry. We made similar observations for companies within the United States, where we identified 489 companies with attributable states and 20 with non-attributable states in early 2023 with "Do Not Sell" links on their homepages or declarations not to sell or share personal information in their privacy policies. As expected, most of these companies (158) resided in California, followed by companies in New York (104), Massachusetts (27), Texas (22), and Washington (20). The higher number of "Do Not Sell" links and declarations on New York companies' websites compared to other US states could be explained by New York being the country's finance and investment center, while California is a tech hub, both attractive environments for company headquarters.

With regard to the definition of companies required to comply with the CCPA/CPRA (see Section 2), a relevant limitation of the Free Company Dataset is the lack of annual revenue and the unique number of individuals affected by data collection. Therefore, we cannot investigate the compliance of companies with CCPA/CPRA.

The CCPA/CPRA had an impact not only on transparency regarding the data practices of American companies but also on those of companies in 35 other countries around the world, as evidenced by the presence of "Do Not Sell" links on their websites in early 2023.

6 DISCUSSION & LIMITATIONS

In this section we discuss our findings, their implications for future privacy policy analyses, and limitations of our approach.

Diachronic bilingual privacy policy analysis. By applying statistical and modern analysis methods and dividing the English and German GDPR and CCPA/CPRA corpora and the English Wagner corpus by the dates of three important privacy regulations, we strengthened the extensive findings of previous research and discovered new trends and topics in English privacy policies by comparing them with our findings in German privacy policies in fixed sets of domains. The bilingual comparison of the two most commonly used languages in the European Union sheds light on language-specific developments, such as English privacy policies increasingly referring to legitimate interests, while German privacy policies abandoned the concept of opt-out cookies. To the best of our knowledge, our work is the first to provide longitudinal insights into the changes in German privacy policies over a data collection period of almost two years, though our available German data is

still too limited for a thorough comparison of CCPA/CPRA effects. We also provided first insights into the CPRA taking effect.

Benefits of our methods. Measuring whether predefined terms from multilingual word lists occurred in privacy policies [15], as well as applying trained deep learning classifiers based on the annotated OPP-115 corpus from 2016 [1, 13, 31, 44, 53, 76] are well-established methods in privacy policy analysis. The keyness analysis employed in this paper is language-independent and allows for closer investigation of changes in privacy policies independent of (incomplete) word lists or machine-learning models which might output false positives and negatives or require discarding trained models due to low precision [76]. Therefore, this method withstands the test of time and can be employed in future research. Our settings for keyness analysis are the strictest in the field of corpus linguistics [46, 77]. These settings include the defined value of 10 for the BIC and a normalized minimum of 10 occurrences per million n-grams. A unique setting of our keyness analysis is to consider the occurrence of each n-gram at most once per policy, which enables us to map the number of occurrences of each n-gram to the number of inspected policies. While LDA topic modeling has been used to identify new topics in privacy policies after GDPR enforcement [69], we used a modern topic modeling technique over time based on sentence-BERT to discover new trends in privacy policies, thus providing privacy policy researchers with new tools.

Our recommendations. Addressing regulators and enforcement agencies, we strongly suggest that future regulations incorporate more concrete requirements for implementing transparency and control mechanisms. New regulations should be accompanied by additional guidelines in non-legal language. Concrete examples or even sample code could provide further guidance. As evidenced by our initial observation of 60 wordings for the "Do Not Sell" link, this still does not guarantee that affected businesses are initially legally compliant, but would head towards providing them with practical and non-ambiguous guidelines. In this light, CPRA guidelines now allowing for unspecified combinations of the two links could make compliance more difficult for websites. Another example of how more concrete legal requirements could boost compliance is the possible effect of the TTDSG on German privacy policies regarding opt-out cookies, as the GDPR had deliberately not clarified the usage of cookies, but deferred them to a future ePrivacy Regulation. We recommend that affected businesses ask regulators to provide clear and practical guidance on how to implement legal requirements concretely, as these businesses are directly bound by data protection legislation.

Limitations of our analysis. Although our focus on intersecting privacy policy domains (see Section 4.2) restricts the size of our data set, this allows for a thorough comparison of changes over the enforcement of three crucial privacy regulations. The small set of domains for the CCPA/CPRA corpus is, besides the aforementioned fluctuations in the used Tranco lists, caused by our February 2021 data collection, in which only the top 10K domains of the Tranco list were visited, in contrast to the top 100K domains for all other website crawls. Moreover, we are aware of the data gap in our CCPA/CPRA corpus for the second half of 2021 and 2022. Filling data gaps from archives like the Internet Archive's Wayback Machine was not feasible, as we collected the homepages and privacy policies using servers in California to simulate the location of Californian residents. However, the analyses have shown clear longitudinal results and trends for the privacy policies of consistently present domains, and we do not expect the privacy policies of domains only occasionally appearing on the Tranco lists during this data gap to have a significant influence on our results.

Code availability. During the longitudinal analysis of our corpora, we developed customized code and expanded existing libraries to perform our analysis. To enable the privacy policy research community to perform similar analyses for the enforcement of future privacy regulations, we make our code available on GitHub[2].

7 CONCLUSION & FUTURE WORK

In this work, we analyzed how the enforcement of the GDPR and CCPA and the CPRA taking effect influenced the language of privacy policies. We conducted text and topic modeling analyses based on modern linguistic standards, providing more details into their effect sizes in privacy policies while confirming previous findings that the enforcement of the GDPR was reflected in the privacy policies around the time it came into effect.

Our findings indicate that for the CCPA significant changes in the texts and topics of privacy policies mainly occurred after the CPRA had become effective on January 1, 2023. Earlier, we had observed widely differing wordings for the "Do Not Sell (or Share) My Personal Information" link, while over time the wording mandated by the CCPA/CPRA had become more widespread. This illustrates that, even when laws clearly state very specific requirements, companies can find it difficult to implement them in practice, which illustrates a need for regulators to provide further guidance. The topic modeling over time showed a gradually rising and concerning trend in the usage of "legitimate interests" since 2021 as the legal basis of data processing in privacy policies.

At the time of writing, the CPRA (California Privacy Rights Act) is due to be enforced in 2024. Other US states that have started to follow suit and passed their own state privacy laws that will soon become effective include Utah, Colorado, Connecticut, Virginia, Iowa, Indiana, Tennessee, Montana, and Texas [60]. We encourage privacy researchers to draw inspiration from our work and collect longitudinal data to observe the effect of these legislations on companies' privacy practices, including their websites and privacy policies, and observe how they affect privacy policy language.

ACKNOWLEDGMENTS

The authors express their gratitude to Lukas Ottenjann for implementing the preliminary CCPA study. We also thank our anonymous reviewers for their feedback. This project received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 462287308 (HU 3005/2-1).

[2]https://github.com/ITSec-Uni-Munster/Bilingual-Longitudinal-Analysis-of-Privacy-Policies.

REFERENCES

[1] Andrick Adhikari, Sanchari Das, and Rinku Dewri. 2023. Evolution of Composition, Readability, and Structure of Privacy Policies over Two Decades. *Proceedings on Privacy Enhancing Technologies* 2023, 3 (May 2023), 138–153. https://doi.org/10.56553/popets-2023-0074

[2] Hend Al-Khalifa, Malak Mashaabi, Ghadi Al-Yahya, and Raghad Alnashwan. 2023. The Saudi Privacy Policy Dataset. *arXiv preprint arXiv:2304.02757* (2023), 8 pages. https://doi.org/10.48550/arXiv.2304.02757

[3] Alexa Internet, Inc. 2019. The top 500 sites on the Web. Retrieved April 30, 2022 from https://www.alexa.com/topsites

[4] Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *The Web Conference 2021 – Proceedings of the World Wide Web Conference (WWW '21)*. ACM, New York, NY, USA, 2165–2176. https://doi.org/10.1145/3442381.3450048

[5] Siddhant Arora, Henry Hosseini, Christine Utz, Vinayshekhar Bannihatti Kumar, Tristan Dhellemmes, Abhilasha Ravichander, Peter Story, Jasmine Mangat, Rex Chen, Martin Degeling, Tom Norton, Thomas Hupperich, Shomir Wilson, and Norman Sadeh. 2022. A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. ELRA, Paris, France, 5460–5472. http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.585.pdf

[6] Rajkumar Arun, Venkatasubramaniyan Suresh, C. E. Veni Madhavan, and M. Narasimha Murthy. 2010. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Advances in Knowledge Discovery and Data Mining – Proceedings of the 14th Pacific-Asia Conference (PAKDD 2010)*. Springer, Berlin, Heidelberg, Germany, 391–402. https://doi.org/10.1007/978-3-642-13657-3_43

[7] Nastaran Bateni, Jasmin Kaur, Rozita Dara, and Fei Song. 2022. Content Analysis of Privacy Policies Before and After GDPR. In *Proceedings of the 2022 19th Annual International Conference on Privacy, Security & Trust (PST 2022)*. IEEE, Washington, DC, USA. https://doi.org/10.1109/PST55820.2022.9851983

[8] David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. ACM, New York, NY, USA, 113–120. https://doi.org/10.1145/1143844.1143859

[9] Vaclav Brezina. 2018. *Statistics in Corpus Linguistics – A Practical Guide*. Cambridge University Press, Cambridge, United Kingdom. https://doi.org/10.1017/9781316410899

[10] Bundesministerium der Justiz. 2021. Gesetz über den Datenschutz und den Schutz der Privatsphäre in der Telekommunikation und bei Telemedien (Telekommunikation-Telemedien-Datenschutz-Gesetz – TTDSG). Retrieved December 15, 2023 from https://www.gesetze-im-internet.de/ttdsg/

[11] Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. 2009. A density-based method for adaptive LDA model selection. *Neurocomputing* 72, 7-9 (March 2009), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

[12] Rex Chen, Fei Fang, Thomas Norton, Aleecia M. McDonald, and Norman Sadeh. 2021. Fighting the Fog: Evaluating the Clarity of Privacy Disclosures in the Age of CCPA. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society (WPES '21)*. ACM, New York, NY, USA, 73–102. https://doi.org/10.1145/3463676.3485601

[13] Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. 2023. PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (ACL 2023)*. ACL, Stroudsburg, PA, USA, 352–365. https://doi.org/10.18653/v1/2023.acl-short.31

[14] Francesco Ciclosi, Silvia Vidor, and Fabio Massacci. 2023. Building cross-language corpora for human understanding of privacy policies. *arXiv preprint arXiv:2302.05355* (2023), 20 pages. https://doi.org/10.48550/arXiv.2302.05355

[15] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. Internet Society, Reston, VA, USA. https://doi.org/10.14722/ndss.2019.23378

[16] Romain Deveaud, Eric Sanjuan, and Patrice Bellot. 2014. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval. *Revue des sciences et technologies de l'information* 17, 1 (2014), 61–84. https://doi.org/10.3166/DN.17.1.61-84

[17] Burton DeWilde. 2016. textacy: NLP, before and after spaCy. Retrieved December 16, 2023 from https://github.com/chartbeat-labs/textacy

[18] Directorate-General for Communication. 2014. Special Eurobarometer 386: Europeans and their Languages. Retrieved December 16, 2023 from https://data.europa.eu/data/datasets/s1049_77_1_ebs386?locale=en

[19] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-million-site Measurement and Analysis. In *Proceedings of the 2016 ACM Conference on Computer and Communications Security (CCS '16)*. ACM, New York, NY, USA, 1388–1401. https://doi.org/10.1145/2976749.2978313

[20] The European Commission. 2023. What is personal data? Retrieved December 16, 2023 from https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

[21] The European Parliament and the Council of the European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* L 119/1 (4 May 2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[22] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. SAGE Publications Ltd, London, United Kingdom. https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-r/book236067

[23] Ronald Aylmer Fisher. 1992. *Statistical Methods for Research Workers*. Springer, New York, NY, USA. https://doi.org/10.1007/978-1-4612-4380-9_6

[24] Costas Gabrielatos. 2018. Keyness analysis: Nature, metrics and techniques. In *Corpus Approaches to Discourse*, Charlotte Taylor and Anna Marchi (Eds.). Routledge, London, United Kingdom, 225–258. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315179346-11/keyness-analysis-costas-gabrielatos

[25] Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021. Sentence Boundary Detection in German Legal Documents. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021)*. SCITEPRESS, Setúbal, Portugal, 812–821. https://doi.org/10.5220/0010246308120821

[26] The Global Privacy Control group. 2020. Global Privacy Control. Retrieved December 16, 2023 from https://globalprivacycontrol.org/

[27] Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101, suppl 1 (April 2004), 5228–5235. https://doi.org/10.1073/pnas.0307752101

[28] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022), 10 pages. https://doi.org/10.48550/arXiv.2203.05794

[29] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. 2021. Toggles, Dollar Signs, and Triangles: How to (In)Effectively Convey Privacy Choices with Icons and Link Texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Article 63, 25 pages. https://doi.org/10.1145/3411764.3445387

[30] Andrew Hardie. 2014. Statistical identification of keywords, lockwords and collocations as a two-step procedure. In *ICAME 35 – Corpus Linguistics, Context and Culture*. University of Nottingham, Nottingham, United Kingdom. https://www.nottingham.ac.uk/conference/fac-arts/english/icame-35/documents/icame35-book-of-abstracts.pdf

[31] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security '18)*. USENIX Association, Berkeley, CA, USA, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[32] Marti A. Hearst. 1997. Text Tiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23, 1 (March 1997), 33–64. https://aclanthology.org/J97-1003

[33] Knut Hofland and Stig Johansson. 1982. *Word Frequencies in British and American English* (2 ed.). Norwegian Computing Centre for the Humanities, Bergen, Norway.

[34] Henry Hosseini, Martin Degeling, Christine Utz, and Thomas Hupperich. 2021. Unifying Privacy Policy Detection. *Proceedings on Privacy Enhancing Technologies* 2021, 4 (July 2021), 480–499. https://doi.org/10.2478/popets-2021-0081

[35] Laura Jehl and Alan Freil. 2018. CCPA and GDPR Comparison Chart. Retrieved December 16, 2023 from https://web.archive.org/web/20230609115637/https://www.bakerlaw.com/webfiles/Privacy/2018/Articles/CCPA-GDPR-Chart.pdf

[36] Irene Kamara and Paul De Hert. 2018. Understanding the Balancing Act Behind the Legitimate Interest of the Controller Ground: A Pragmatic Approach. *Brussels Privacy Hub Working Paper* 4, 12 (Aug. 2018), 35 pages. https://brusselsprivacyhub.eu/BPH-Working-Paper-VOL4-N12.pdf

[37] Jasmin Kaur, Rozita A. Dara, Charlie Obimbo, Fei Song, and Karen Menard. 2018. A comprehensive keyword analysis of online privacy policies. *Information Security Journal: A Global Perspective* 27, 5-6 (May 2018), 260–275. https://doi.org/10.1080/19393555.2019.1606368

[38] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate Detection using Shallow Text Features. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WDSM '10)*. ACM, New York, NY, USA, 441–450. https://doi.org/10.1145/1718487.1718542

[39] Konrad Kollnig, Lu Zhang, Jun Zhao, and Nigel Shadbolt. 2023. Before and after China's new Data Laws: Privacy in Apps. In *Workshop on Technology and Consumer Protection (ConPro '23)*. IEEE Computer Society, Los Alamitos, CA, USA. https://conpro23.ieee-security.org/papers/kollnig-conpro23.pdf

[40] Lin Kyi, Sushil Ammanaghatta Shivakumar, Cristiana Teixeira Santos, Franziska Roesner, Frederike Zufall, and Asia J. Biega. 2023. Investigating Deceptive Design in GDPR's Legitimate Interest. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 2023)*. ACM, New York, NY, USA,

Article 583, 16 pages. https://doi.org/10.1145/3544548.3580637
[41] Victor Le Pochat, Tom Van Goethem, Samaneh Talajizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. TRANCO: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS '19)*. Internet Society, Reston, VA, USA. https://doi.org/10.14722/ndss.2019.23386
[42] Geoffrey Leech and Roger Fallon. 1992. Computer corpora – What do they tell us about culture? *ICAME Journal – Computers in English Linguistics* 16 (April 1992), 29–50. http://icame.uib.no/archives/No_16_ICAME_Journal_index.pdf
[43] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained Named Entity Recognition in Legal Documents. In *Semantic Systems. The Power of AI and Knowledge Graphs. Proceedings of the 15th International Conference (SEMANTiCS 2019)*. Springer Nature Switzerland AG, Cham, Switzerland, 272–287. https://doi.org/10.1007/978-3-030-33220-4_20
[44] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (Jan. 2020), 47–64. https://doi.org/10.2478/popets-2020-0004
[45] Fei Liu, Rohan Ramanath, Norman Saden, and Noah A. Smith. 2014. A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)*. ACL, Stroudsburg, PA, USA, 884–894. https://aclanthology.org/C14-1084.pdf
[46] Stefania M. Maci and Michele Sala. 2022. *Corpus Linguistics and Translation Tools for Digital Humanities: Research Methods and Applications* (1 ed.). Bloomsbury Publishing, New York, NY, USA. https://www.bloomsbury.com/us/corpus-linguistics-and-translation-tools-for-digital-humanities-9781350275225/
[47] Célestin Matte, Nataliia Bielova, and Cristiana Santos. 2020. Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework. In *Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP '20)*. IEEE Computer Society, Los Alamitos, CA, USA, 791–809. https://doi.org/10.1109/SP40000.2020.00076
[48] Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 543–568. http://hdl.handle.net/1811/72839
[49] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software* 2, 11 (March 2017), 205. https://doi.org/10.21105/joss.00205
[50] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426* (2018), 63 pages. https://doi.org/10.48550/arXiv.1802.03426
[51] Steven M. Millendorf. 2023. CPRA Enforcement Delayed Until at Least March 29, 2024. Retrieved December 16, 2023 from https://www.foley.com/en/insights/publications/2023/07/cpra-enforcement-delayed-march-29-2024
[52] Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2020. spaCy: Industrial-strength Natural Language Processing in Python. Retrieved December 16, 2023 from https://doi.org/10.5281/zenodo.1212303
[53] Najmeh Mousavi Nejad, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, and Damien Graux. 2020. Establishing a Strong Baseline for Privacy Policy Classification. In *Systems Security and Privacy Protection. Proceedings of the 35th IFIP TC 11 International Conference (SEC 2020)*. Springer Nature Switzerland AG, Cham, Switzerland, 370–383. https://doi.org/10.1007/978-3-030-58201-2_25
[54] Nikita Murzintcev and Nathan Chaney. 2016. ldatuning. Retrieved December 16, 2023 from https://github.com/nikita-moor/ldatuning
[55] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2017. A study of web privacy policies across industries. *Journal of Information Privacy and Security* 13, 4 (2017), 169–185. https://doi.org/10.1080/15536548.2017.1394064
[56] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376321
[57] Sean O'Connor, Ryan Nurwono, Aden Siebel, and Eleanor Birrell. 2021. (Un)clear and (In)conspicuous: The Right to Opt-out of Sale under CCPA. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society (WPES '21)*. ACM, New York, NY, USA, 59–72. https://doi.org/10.1145/3463676.3485598
[58] OneTrust Data Guidance and the Future of Privacy Forum. 2019. Comparing Privacy Laws: GDPR v. CCPA. Retrieved December 16, 2023 from https://fpf.org/wp-content/uploads/2019/12/ComparingPrivacyLaws_GDPR_CCPA.pdf
[59] People Data Labs. 2023. Free Company Dataset. Retrieved December 16, 2023 from https://www.peopledatalabs.com/company-dataset
[60] F. Paul Pittman. 2023. Texas Passes Comprehensive Data Privacy Law. Retrieved December 16, 2023 from https://www.whitecase.com/insight-alert/texas-passes-comprehensive-data-privacy-law
[61] Thomas Proisl. 2018. SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. ELRA, Paris, France, 665–670. https://aclanthology.org/L18-1106/
[62] Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC 2016)*. ACL, Stroudsburg, PA, USA, 57–62. http://aclweb.org/anthology/W16-2607
[63] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL 2020)*. ACL, Stroudsburg, PA, USA, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14
[64] Joel R. Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James Graves, Fei Liu, Aleecia McDonald, Thomas Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2015. Disagreeable Privacy Policies: Mismatches between Meaning and Users' Understanding. *Berkeley Technology Law Journal* 30, 1 (2015), 39–88. https://doi.org/10.2139/ssrn.2418297
[65] Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. *arXiv preprint arXiv:2004.09813* (2020), 14 pages. https://doi.org/10.48550/arXiv.2004.09813
[66] Anne Riechert and Thomas Wilmer. 2022. *TTDSG – Telekommunikation-Telemedien-Datenschutz-Gesetz* (1 ed.). Erich Schmidt Verlag, Berlin, Germany. https://www.esv.info/978-3-503-20979-8
[67] Pavel Rychlý. 2008. A Lexicographer-Friendly Association Score. In *Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Language Processing (RASLAN 2008)*. Masaryk University, Brno, Czech Republic, 6–9. https://nlp.fi.muni.cz/raslan/2008/papers/13.pdf
[68] George Sanchez. 2019. Sentence Boundary Detection in Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2019 (NAACL 2019)*. ACL, Stroudsburg, PA, USA, 31–38. https://doi.org/10.18653/v1/W19-2204
[69] David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19 Companion)*. IW3C2, Geneva, Switzerland, 563–568. https://doi.org/10.1145/3308560.3317585
[70] State of California Department of Justice, Office of the Attorney General. 2023. California Consumer Privacy Act (CCPA) – Frequently Asked Questions. Retrieved December 16, 2023 from https://oag.ca.gov/privacy/ccpa
[71] State of California Legislative Counsel. 2018. Assembly Bill No. 375 – Chapter 55. Retrieved December 16, 2023 from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375
[72] State of California Legislative Counsel. 2020. Proposition 24. Retrieved December 16, 2023 from https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5
[73] Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: Towards an Implementation of the EU GDPR on Internet Privacy Policy Evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics (IWSPA '18)*. ACM, New York, NY, USA, 15–21. https://doi.org/10.1145/3180445.3180447
[74] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. 2019. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19)*. ACM, New York, NY, USA, 973–990. https://doi.org/10.1145/3319535.3354212
[75] Maggie Van Nortwick and Christo Wilson. 2022. Setting the Bar Low: Are Websites Complying With the Minimum Requirements of the CCPA? *Proceedings on Privacy Enhancing Technologies* 2022, 1 (Jan. 2022), 608–628. https://doi.org/10.2478/popets-2022-0030
[76] Isabel Wagner. 2023. Privacy Policies across the Ages: Content of Privacy Policies 1996-2021. *ACM Transactions on Privacy and Security* 26, 3, Article 32 (May 2023), 32 pages. https://doi.org/10.1145/3590152
[77] Andrew Wilson. 2013. Embracing Bayes factors for key item analysis in corpus linguistics. In *New Approaches to the Study of Linguistic Variability*, Markus Bieswanger and Amei Koll-Stobbe (Eds.). Peter Lang, Lausanne, Switzerland, 3–11. https://beckassets.blob.core.windows.net/product/preamble/13062959/9783631615041_intro_005.pdf
[78] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL '16)*. ACL, Stroudsburg, PA, USA, 1330–1340. https://doi.org/10.18653/v1/P16-1126
[79] Sebastian Zimmeck and Kuba Alicki. 2019. Standardizing and Implementing Do Not Sell. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society (WPES '20)*. ACM, New York, NY, USA, 15–20. https://doi.org/10.1145/3411497.3420224

A REGULAR EXPRESSIONS TO DETECT "DO NOT SELL" LINK VARIANTS

The following listing shows the regular expressions that we used to search for "Do Not Sell or Share" links on homepages and in privacy statements.

Listing 1: "Do Not Sell or Share" on homepages

```
^(?!^do not sell$)(ccpa|cpra|ca|california
    )?\s?[-:]?\s?(do not|don't)\ssell\s?(
    or)?\s?(share)?\s?(my)?\s?(personal)?\
    s?(info|information|data|PI)?\s?(\(ca
    \)|\(california\)|\(ccpa\)|\(cpra\))?$
```

Listing 2: "Do Not Sell or Share" in privacy statements

```
(ccpa|cpra|ca(lifornia)?)?\s?[-:]?\s?((do
    not)|(don't)|(does␣not)|(doesn't))\
    ssell\s?(or)?\s?(share)?\s?(my|your)?\
    s?(personal)?\s?(info(rmation)?|data|
    PI)\s?(\(ca(lifornia)?\)|\(ccpa\)|\(
    cpra\))?
```

B GDPR ARTICLES IN GERMAN PRIVACY POLICIES

The following table shows the keyness statistics for references to GDPR articles in the German GDPR corpus after the GDPR enforcement date compared to privacy policies before that date. BIC = Bayesian information criterion, LR = log ratio, PercDiff = percentage points difference, DiffC = difference coefficient, NF = normalized frequency per million in T (target corpus) and R (reference corpus).

Table 9: Occurrence of legal references in German privacy policies after the GDPR enforcement date.

Article	BIC	LR	PercDiff	DiffC	NFT	NFR
6	412.23	2.73	562.61	0.74	44.22	6.67
6(1)	404.27	2.85	621.08	0.76	41.531	5.76
6(1)(a)	140.24	4.15	1,680.05	0.89	11.39	0.64
6(1)(c)	97.39	2.75	574.02	0.74	11.71	1.74
6(1)(f)	92.19	2.58	495.88	0.71	11.99	2.01
15	275.28	3.23	835.97	0.81	25.67	2.74
16	307.87	5.04	3,194.32	0.94	21.08	0.64
17	348.64	4.62	2,366.11	0.92	24.80	1.01
18	208.01	4.58	2,298.12	0.92	15.35	0.64
20	265.78	3.47	1,009.84	0.84	23.34	2.10
21	227.23	3.08	744.41	0.79	22.39	2.65
28	139.14	3.86	1,346.98	0.87	11.91	0.82
46	201.80	4.39	1,992.95	0.91	15.31	0.73
77	221.46	6.28	7,687.71	0.98	14.24	0.18

C TOP INTERSECTING DOMAINS

The following table presents the top domains in our CCPA/CPRA corpus for both English and German that we found to be intersecting over all data collection time points, ranked and based on the Tranco list.

Table 10: Top-ranked intersecting privacy policy domains of the German and English CCPA/CPRA corpora based on the Tranco list from December 22, 2022 (ID: 82V9V).

English		German	
Domain	Rank	Domain	Rank
google.com	1	mpg.de	2045
facebook.com	4	ccc.de	2423
microsoft.com	5	fraunhofer.de	3155
twitter.com	9	mobile.de	3626
apple.com	12	adition.com	3693
linkedin.com	13	bahn.de	3742
yahoo.com	16	tum.de	4986
amazon.com	17	idealo.de	5219
cloudflare.com	19	derstandard.at	5284
wordpress.org	30	bayern.de	5387
github.com	32	ndr.de	5633
pinterest.com	33	uni-hamburg.de	5772
zoom.us	39	kit.edu	6533
reddit.com	40	bundesregierung.de	7238
adobe.com	44	wetteronline.de	7593
intuit.com	62	post.ch	7677
tiktok.com	65	vodafone.de	7962
mozilla.org	71	auswaertiges-amt.de	8040
paypal.com	83	deutsche-bank.de	8153
spotify.com	84	uni-tuebingen.de	8212
opera.com	89	gutefrage.net	8599
nih.gov	90	mdr.de	8715
nytimes.com	96	uni-kiel.de	8911
dropbox.com	100	uni-goettingen.de	9228
flickr.com	101	uni-stuttgart.de	9285
digicert.com	104	swr.de	9524
salesforce.com	107	uni-bremen.de	10019
medium.com	109	presseportal.de	10949
imdb.com	113	deutschepost.de	11097
soundcloud.com	118	tagesanzeiger.ch	11260
apache.org	119	united-domains.de	11264
fandom.com	122	lidl.de	11515
theguardian.com	142	daad.de	12361
stackoverflow.com	144	giga.de	12381
sciencedirect.com	151	wko.at	13206
xhamster.com	153	fau.de	14711
etsy.com	154	abendblatt.de	15695
twitch.tv	155	webgains.com	16528
epicgames.com	156	etracker.com	18311
amazon.co.uk	159	winfuture.de	18325
amazon.in	161	comdirect.de	23585
creativecommons.org	169	homepage-baukasten.de	399833

D "DO NOT SELL" VARIANTS

Table 11: "Do Not Sell" variants on homepages and their prevalence over time in the CCPA/CPRA corpus.

	2019	2020														2023	
	Dec	Jan		Feb		Mar		Apr		May		Jun		Jul		Jan	Feb
Wording	09	02	16	03	26	09	23	09	27	13	22	09	23	09	30	14	13
Do Not Sell My Personal Information	0.57	1.25	2.42	2.51	2.37	2.60	3.08	2.28	3.35	2.24	2.15	3.19	3.32	4.14	4.14	8.75	7.08
Do Not Sell or Share My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.85	3.12
Do Not Sell My Info	1.09	1.39	1.64	1.52	1.83	1.81	1.81	1.87	1.78	1.75	1.86	1.73	1.84	1.99	2.10	1.14	1.00
Do not sell my personal information	0.02	0.14	0.33	0.33	0.35	0.23	0.31	0.29	0.57	0.29	0.27	0.48	0.48	0.45	0.47	0.79	0.95
Do not sell my info	0.21	0.22	0.28	0.40	0.42	0.46	0.45	0.46	0.48	1.04	1.13	1.25	0.52	0.54	0.68	0.54	0.42
Do Not Sell or Share My Personal Data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.39
CCPA: Do not sell my personal info	0.00	0.05	0.17	0.19	0.19	0.16	0.16	0.17	0.15	0.16	0.13	0.15	0.15	0.15	0.13	0.25	0.36
Do Not Sell My Personal Info	0.07	0.38	0.43	0.42	0.40	0.37	0.36	0.39	0.35	0.36	0.35	0.20	0.20	0.28	0.28	0.46	0.33
Do Not Sell My Information	0.00	0.14	0.24	0.23	0.28	0.23	0.27	0.17	0.26	0.13	0.13	0.26	0.28	0.41	0.40	0.41	0.33
Do not sell or share my personal information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.28
Do Not Sell Or Share My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.25
Do not sell my data	0.05	0.05	0.05	0.05	0.05	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04	0.04	0.14	0.14
Do Not Sell My Data	0.02	0.05	0.07	0.05	0.07	0.07	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.06	0.04	0.14	0.11
Do Not Sell or Share My Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.11
Do not Sell or Share My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.11
Do not sell or share my info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.11
Do Not Sell My Personal Data	0.00	0.00	0.05	0.09	0.09	0.09	0.09	0.10	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.00	0.11
CA - Do Not Sell My Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.08	0.08
Don't sell or share my personal information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08
Do Not Sell or Share My Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.08
DO NOT SELL PERSONAL INFORMATION	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
Do not sell my information	0.02	0.05	0.02	0.07	0.09	0.09	0.09	0.10	0.09	0.09	0.07	0.07	0.07	0.11	0.11	0.05	0.06
California Do Not Sell	0.00	0.07	0.09	0.05	0.02	0.05	0.09	0.07	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.05	0.06
Do Not Sell My PI	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.02	0.02	0.02	0.05	0.06
CA: Do Not Sell My Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.05	0.06
Do not Sell My Personal Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.06
DO NOT SELL OR SHARE MY INFORMATION	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.06
DO NOT SELL OR SHARE MY PERSONAL INFORMATION	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.06
Do Not Sell my Data	0.00	0.05	0.05	0.05	0.05	0.05	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03
CA: Do Not Sell My Personal Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.00	0.03
Do not sell or share my personal data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.03
Do Not Sell Or Share My Personal Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03
CA Do Not Sell or Share My Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03
California - Do Not Sell My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.03
California Do Not Sell My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
CA Do Not Sell My Info	0.26	0.31	0.36	0.28	0.31	0.30	0.31	0.32	0.29	0.29	0.27	0.26	0.26	0.26	0.28	0.00	0.00
Do Not Sell My Info (California)	0.07	0.10	0.14	0.12	0.14	0.12	0.13	0.19	0.15	0.16	0.16	0.07	0.09	0.11	0.09	0.00	0.00
CCPA Do Not Sell My Info	0.05	0.05	0.05	0.05	0.07	0.07	0.07	0.07	0.00	0.07	0.07	0.07	0.07	0.06	0.06	0.00	0.00
CA Do Not Sell My Personal Information	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00
California Do Not Sell My Info	0.00	0.10	0.07	0.09	0.09	0.07	0.07	0.10	0.07	0.07	0.07	0.09	0.09	0.09	0.06	0.00	0.00
Do not Sell My Info	0.00	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.06	0.06	0.00	0.00
California: Do Not Sell My Info	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00
Do Not Sell my Personal Info	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00
Do Not Sell My Info (CA)	0.00	0.02	0.02	0.02	0.05	0.05	0.04	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.00	0.00
DO NOT SELL MY PERSONAL INFO	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00
CCPA-Do Not Sell My Personal Information	0.00	0.00	0.07	0.02	0.07	0.07	0.07	0.00	0.07	0.00	0.00	0.07	0.07	0.06	0.06	0.00	0.00
DO NOT SELL MY PERSONAL INFORMATION	0.00	0.00	0.05	0.07	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00
Do not sell my info (California)	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.11	0.11	0.09	0.09	0.08	0.00
Don't Sell My Personal Info	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.04	0.04	0.02	0.04	0.02	0.02	0.00	0.00
Do Not Sell My Personal Information (CCPA)	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.00	0.02	0.02	0.02	0.02	0.00	0.00
Don't sell my info	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.00	0.00
CCPA Do Not Sell My Personal Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.02	0.00	0.00	0.11	0.00	0.00
Do not sell my Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00
Don't sell my data	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.00
Don't Sell My Info	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00
CA: Do not sell my personal information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00
DO NOT SELL MY DATA	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00
Do not sell my info (CA)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00
California Do Not Sell My Information	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00

E COMPANIES WITH "DO NOT SELL" LINKS OR STATEMENTS

Table 12: Number of companies by country in the world (left) and US state (right) whose websites' homepages or privacy policies contained "Do Not Sell" links or statements in early 2023, as visualized in Figure 6. The states of 20 US companies were not included in the Free Company Dataset; we marked these as N/A.

World		USA	
Country	**Company**	**US State**	**Company**
United States	509	California	158
United Kingdom	48	New York	104
Germany	17	Massachusetts	27
Canada	14	Texas	22
India	11	N/A	20
France	6	Washington	20
Netherlands	6	District of Columbia	18
Denmark	5	Florida	16
Philippines	4	Virginia	14
Switzerland	4	Illinois	12
China	3	New Jersey	8
Finland	3	Maryland	7
Singapore	3	North Carolina	7
Australia	2	Colorado	6
Indonesia	2	Connecticut	6
Italy	2	Georgia	6
Japan	2	Minnesota	6
Norway	2	Utah	6
Belarus	1	Pennsylvania	5
Bermuda	1	Indiana	4
Brunei	1	Michigan	4
Bulgaria	1	Ohio	3
Czechia	1	Oregon	3
Ghana	1	South Carolina	3
Hong Kong	1	Tennessee	3
Ireland	1	Kansas	2
Malaysia	1	Kentucky	2
Mexico	1	Missouri	2
Pakistan	1	Nebraska	2
Poland	1	Oklahoma	2
Saudi Arabia	1	Wisconsin	2
South Korea	1	Wyoming	2
Spain	1	Arizona	1
Sweden	1	Delaware	1
Thailand	1	Louisiana	1
Ukraine	1	Maine	1
United Arab Emirates	1	Mississippi	1
		Nevada	1
		Vermont	1

F FLUCTUATIONS IN TRANCO DOMAINS OVER TIME

Table 13: Ranking fluctuations in the top 115 domains of the Tranco lists used in this paper.

Domain	GVWK	WQW9	82V9V	Domain	GVWK	WQW9	82V9V
google.com	1	1	1	.	.	.	.
facebook.com	2	2	4	⋮	⋮	⋮	⋮
netflix.com	3	16	8	imgur.com	58	109	188
youtube.com	4	3	3	dropbox.com	59	54	100
twitter.com	5	5	9	xinhuanet.com	60	42	693
microsoft.com	6	4	5	nytimes.com	61	50	96
amazon.com	7	19	17	alipay.com	62	57	215
tmall.com	8	6	178	csdn.net	63	69	76
linkedin.com	9	10	13	microsoftonline.com	64	38	49
instagram.com	10	7	10	flickr.com	65	52	101
baidu.com	11	11	11	stackoverflow.com	66	94	144
wikipedia.org	12	13	15	yahoo.co.jp	67	73	120
apple.com	13	12	12	soundcloud.com	68	62	118
qq.com	14	9	18	gravatar.com	69	55	112
yahoo.com	15	18	16	t.co	70	67	68
sohu.com	16	15	116	urbandictionary.com	71	830	1459
live.com	17	14	22	imdb.com	72	80	113
taobao.com	18	21	66	nih.gov	73	64	90
adobe.com	19	23	44	realtor.com	74	644	1588
wikipedia.com	20	3421	7883	icloud.com	75	112	93
doubleclick.net	21	17	31	cnn.com	76	56	117
googletagmanager.com	22	20	25	aliexpress.com	77	83	129
yelp.com	23	242	376	office365.com	78	75	78
windowsupdate.com	24	8	42	whatsapp.com	79	51	38
pinterest.com	25	25	33	akamaiedge.net	80	139	7
macromedia.com	26	88	86	medium.com	81	63	109
blogspot.com	27	44	70	medicalnewstoday.com	82	631	940
msn.com	28	49	74	nflxso.net	83	382	35
reddit.com	29	28	40	apache.org	84	68	119
bing.com	30	27	24	theguardian.com	85	77	142
360.cn	31	24	266	bbc.co.uk	86	82	148
jd.com	32	30	106	w3.org	87	71	141
weibo.com	33	34	75	stackexchange.com	88	247	480
sina.com.cn	34	37	87	paypal.com	89	79	83
giphy.com	35	293	504	europa.eu	90	53	94
wordpress.com	36	32	58	twitch.tv	91	72	155
vimeo.com	37	26	45	sourceforge.net	92	81	158
amazonaws.com	38	39	6	google.com.hk	93	78	111
quizlet.com	39	525	596	livejasmin.com	94	179	723
youtu.be	40	22	37	forbes.com	95	76	131
vk.com	41	45	64	naver.com	96	84	126
ebay.com	42	70	110	walmart.com	97	195	249
fandom.com	43	169	122	babytree.com	98	197	49441
buzzfeed.com	44	362	539	spotify.com	99	74	84
goo.gl	45	35	55	slideshare.net	100	129	275
espn.com	46	186	263	indeed.com	102	166	166
github.com	47	36	32	amazon.co.jp	103	98	205
office.com	48	29	34	bbc.com	104	90	143
tumblr.com	49	46	85	chaturbate.com	105	120	311
bit.ly	50	40	60	thehill.com	106	745	1346
googleusercontent.com	51	33	52	weebly.com	107	97	186
godaddy.com	52	141	284	yandex.ru	108	96	57
google-analytics.com	53	41	80	pornhub.com	109	159	102
mozilla.org	54	47	71	google.co.in	110	116	232
okezone.com	55	58	2928	dailymail.co.uk	111	155	238
skype.com	56	66	98	glassdoor.com	112	645	1254
wordpress.org	57	43	30	tribunnews.com	113	127	6558
.	.	.	.	booking.com	114	200	198
⋮	⋮	⋮	⋮	amazon.in	115	100	161

G EXAMPLES OF PRIVACY POLICY PREPROCESSING

Tables 14 and 15 show examples of how we preprocessed the text of the privacy policies as described in Section 4.3; more concretely, for the topic modeling analysis. In the English example below, company names and an email address were replaced with the placeholders "COMPANYNAME" and "REPLACEDMAIL," respectively. In the German example on the next page, additionally a URL was replaced with "REPLACEDURL." The white spaces before and after the placeholders are put intentionally to prevent accidental concatenation of words with punctuation or placeholders. They do not cause any problems with text processing. The line breaks in the preprocessed text indicate the result of the TextTiling algorithm, i. e., the point where the text segment was split into two tiles. Lemmatization was not applied to the input texts of BERTopic, as this might have resulted in imprecise sentence-BERT embeddings. For the keyness analysis, the texts were afterwards lemmatized with the Spacy library and tokenized with the SoMaJo and SoMeWeta libraries as described in Section 4.3.

Table 14: Original and preprocessed text fragments of the English privacy policy of yahoo.com in February 2023.

Raw Text as Extracted by Boilerpipe	Preprocessed Text
Authorized Agent\nYou may use an authorized agent to submit a request to opt-out of sale, request to know, request to correct, or request to delete on your behalf. If you choose to use an authorized agent to exercise any such rights under the CCPA, you will need to provide the authorized agent written signed permission to act on your behalf. Please direct your authorized agent to email us at california_privacy@yahooinc.com where they will receive instructions on how to submit a request on your behalf.\nAuthorized Agent Request to Opt-Out of Sale\nUsers with Registered Accounts\nAuthorized agents submitting a request on a user\'s behalf to request opt-out of sale of the user\'s personal information must provide Yahoo evidence of the authorized agent\'s power of attorney or an authorization signed by the consumer showing the agent is authorized by the consumer to act on the consumer\'s behalf.\nUsers without Registered Accounts\nRequests to opt out of sale for non-registered users must come from the device on which the user wishes to opt out of sale. As a result, an authorized agent will need to submit the request to opt out of sale from the applicable device. Due to the nature of Yahoo\'s services, we are only able to opt a non-registered user out of sale if such user takes such action on the device on which the user wishes to opt out.\nAuthorized Agent Request to Know or Request to Delete\nIf you choose to use an authorized agent to exercise your request to know or request to delete on your behalf, Yahoo will require you to verify your identity directly with Yahoo and confirm directly with Yahoo that you provided the authorized agent permission to submit the request on your behalf.\nIf you have provided the authorized agent with power of attorney pursuant to California Probate Code sections 4121 to 4130, the above instructions do not apply. Prior to releasing any personal information to the authorized agent or honoring a deletion request, Yahoo will require verification from the authorized agent of power of attorney to act on your behalf.\nWe may deny a request from an authorized agent that does not submit proof that they have been authorized by you to act on your behalf.	Authorized Agent You may use an authorized agent to submit a request to opt-out of sale, request to know, request to correct, or request to delete on your behalf. If you choose to use an authorized agent to exercise any such rights under the CCPA, you will need to provide the authorized agent written signed permission to act on your behalf. Please direct your authorized agent to email us at **REPLACEDEMAIL** where they will receive instructions on how to submit a request on your behalf. Authorized Agent Request to Opt-Out of Sale Users with Registered Accounts Authorized agents submitting a request on a user's behalf to request opt-out of sale of the user's personal information must provide **COMPANYNAME** evidence of the authorized agent's power of attorney or an authorization signed by the consumer showing the agent is authorized by the consumer to act on the consumer's behalf. Users without Registered Accounts Requests to opt out of sale for non-registered users must come from the device on which the user wishes to opt out of sale. As a result, an authorized agent will need to submit the request to opt out of sale from the applicable device. Due to the nature of **COMPANYNAME**'s services, we are only able to opt a non-registered user out of sale if such user takes such action on the device on which the user wishes to opt out. Authorized Agent Request to Know or Request to Delete If you choose to use an authorized agent to exercise your request to know or request to delete on your behalf, **COMPANYNAME** will require you to verify your identity directly with **COMPANYNAME** and confirm directly with **COMPANYNAME** that you provided the authorized agent permission to submit the request on your behalf. If you have provided the authorized agent with power of attorney pursuant to California Probate Code sections 4121 to 4130, the above instructions do not apply. Prior to releasing any personal information to the authorized agent or honoring a deletion request, **COMPANYNAME** will require verification from the authorized agent of power of attorney to act on your behalf. We may deny a request from an authorized agent that does not submit proof that they have been authorized by you to act on your behalf.

Table 15: Original and preprocessed text fragments of the German privacy policy of swr.de in February 2023.

Raw Text as Extracted by Boilerpipe	Preprocessed Text
Für die Ermittlung der anonymen statistischen Kennwerte wird eine Technik der Firma AT Internet (https://www.atinternet.com/de/) genutzt. Die durch diese Technik gesammelten Daten werden ausschließlich anonymisiert auf Servern in Deutschland gespeichert.\nSie haben in unseren Datenschutz-Einstellungen die Möglichkeit, der anonymen Erfassung Ihrer Nutzungsvorgänge zu widersprechen.\nNielsen-Messverfahren\nNielsen, als beauftragtes Marktforschungsunternehmen, setzt zum Zweck der Webanalyse auf dieser Webseite Cookies ein. Die Webanalyse dient dazu, statistische Analysen über die Nutzung dieser Webseite und deren Angebot zu erstellen. Diese Informationen helfen dabei, die Webseite und die damit verbundenen Services im Hinblick auf Effektivität und Effizienz zu verstehen und zu verbessern. Im Rahmen der Webanalyse und mit dem damit verbundenen Cookie werden nur anonyme Nutzerinformationen erfasst.\nSie haben in unseren Datenschutz-Einstellungen die Möglichkeit, der anonymen Erfassung Ihrer Nutzungsvorgänge zu widersprechen.\nBei Fragen bezüglich des Nielsen-Webanalyse-Auftrags wenden Sie sich bitte per E-Mail an info.deutschland@nielsen.com .\nPodigee (Podcast-Auswertung)\nWir nutzen für die Auswertung der Podcast-Nutzung den Podcast-Hosting-Dienst Podigee des Anbieters Podigee GmbH, Schlesische Straße 20, 10997 Berlin, Deutschland. Die Podcasts werden dabei von Podigee ausgewertet und die Daten aufbereitet.\nDie Nutzung erfolgt auf Grundlage unserer berechtigten Interessen, d.h. Interesse an einer sicheren und effizienten Bereitstellung, Analyse sowie Optimierung unseres Podcastangebotes gem. Art. 6 Abs. 1 lit. f. DSGVO.\nPodigee verarbeitet IP-Adressen und Geräteinformationen, um statistische Daten, wie z.B. Abrufzahlen zu ermitteln. Diese Daten werden vor der Speicherung in der Datenbank von Podigee anonymisiert oder pseudonymisiert, sofern sie für die Bereitstellung der Podcasts nicht erforderlich sind	Für die Ermittlung der anonymen statistischen Kennwerte wird eine Technik der Firma **COMPANYNAME** (**REPLACEDURL**) genutzt. Die durch diese Technik gesammelten Daten werden ausschließlich anonymisiert auf Servern in Deutschland gespeichert. Sie haben in unseren Datenschutz-Einstellungen die Möglichkeit, der anonymen Erfassung Ihrer Nutzungsvorgänge zu widersprechen. **COMPANYNAME**-Messverfahren **COMPANYNAME**, als beauftragtes Marktforschungsunternehmen, setzt zum Zweck der Webanalyse auf dieser Webseite Cookies ein. Die Webanalyse dient dazu, statistische Analysen über die Nutzung dieser Webseite und deren Angebot zu erstellen. Diese Informationen helfen dabei, die Webseite und die damit verbundenen Services im Hinblick auf Effektivität und Effizienz zu verstehen und zu verbessern. Im Rahmen der Webanalyse und mit dem damit verbundenen Cookie werden nur anonyme Nutzerinformationen erfasst. Sie haben in unseren Datenschutz-Einstellungen die Möglichkeit, der anonymen Erfassung Ihrer Nutzungsvorgänge zu widersprechen. Bei Fragen bezüglich des **COMPANYNAME**-Webanalyse-Auftrags wenden Sie sich bitte per E-Mail an **REPLACEDEMAIL** . Podigee (Podcast-Auswertung) Wir nutzen für die Auswertung der Podcast-Nutzung den Podcast-Hosting-Dienst Podigee des Anbieters **COMPANYNAME**, Schlesische Straße 20, 10997 Berlin, Deutschland. Die Podcasts werden dabei von Podigee ausgewertet und die Daten aufbereitet. Die Nutzung erfolgt auf Grundlage unserer berechtigten Interessen, d.h. Interesse an einer sicheren und effizienten Bereitstellung, Analyse sowie Optimierung unseres Podcastangebotes gem. Art. 6 Abs. 1 lit. f. DSGVO. Podigee verarbeitet IP-Adressen und Geräteinformationen, um statistische Daten, wie z.B. Abrufzahlen zu ermitteln. Diese Daten werden vor der Speicherung in der Datenbank von Podigee anonymisiert oder pseudonymisiert, sofern sie für die Bereitstellung der Podcasts nicht erforderlich sind.

H NOUN CHUNK BIGRAM DEPENDENCIES

In the following we report the dependency bi-grams that a) occur in the compared sub-corpora, b) are statistically significant in the target corpus (Fischer's exact test [23], $p < 0.05$ with Benjamini-Hochberg correction [22]), c) appear at least ten times per million in the target corpus and d) have a minimum positive change in log Dice value of 1, equaling to a doubling of co-occurrence exclusivity strength [67].

Table 16: Sanitized lemmatized form of the noun chunk dependents with the highest increment in relative frequency per million in the post-GDPR corpus (after GDPR enforcement in May 2018).

Head	Dependents
collect	location data, cookie, image, sensitive data, search, image, voice data, financial transaction data, unique identifier, diagnostic data, account data, name and contact data, technical data, personal data, user data, any special category, your personal data, information and report website usage statistic, functionality cookie, web traffic data, billing information, uri address, certain data, traffic data, specific location, any sensitive information / special category, email address, limited business contact information, usage data, your address book and calendar meeting information, "email header" information
disable	connect service, personalized feature, exist cookie, certain location-base service, third-party cookie
enable	personalization, other feature, your organization, data sharing, consistent experience, company, location-base service, cross-device experience, connection, sharing, functionality, website owner, advertiser, analysis, access, delivery, archiving, session replay, publicly available information
personalize	our product, experience, advertisement, feature, our content
reject	any content, use
send	promotional communication, your activity history, specific instruction, sms, short snippet, your entire document, customer data, content, customer communication, marketing material, marketing email, your personal information, direct marketing communication, query, letter, service message, your personal data, third-party direct marketing communication, our newsletter, assessment, your contact detail, special signal, service communication, data subject access request, important service-relate message, renewal notice, important notice, your user information, some notification, account deletion request, singular email
share	your data, limited account, limited aggregated information, some de-identify data, your confidential information, snippet, all, non-public additional information, your personal data, certain data, personal data section, your feedback, video, aggregated data, certain amount, your email, user location data, aggregated insight, aggregated data, your contact detail, their name, member, your thought, payment partner, my data, agent, services, cookie
use	your personal data, your work, automate process, various tool, app, third-party service, automate system, contact information, browser-base cookie control, particular word, de-identify device, tracking protection, connected service, browser, query, your keyboard, your voice data, control, personalization, global navigation satellite system, purchase, your calendar, third-party app, product, performance, account, web beacon, similar technology, tool, device, our products, our data request feature, global navigation satellite systems, content

Table 17: Sanitized lemmatized form of the noun chunk dependents with the highest increment in relative frequency per million in privacy policies after CCPA enforcement in July 2020.

Head	Dependents
collect	diagnostic data, information, identifier, unidentifiable data
disable	local device storage
enable	motion, voice command, virtual reality experience, mixed reality experience, delivery
personalize	–
reject	–
send	personal data
share	your location, website content, audience segment
use	usage data, control, resource, toggle switch, social media plug-in, pixel tag, smarturl, service offerings, script, exercise my right link, all intellectual property, any intellectual property, representation, commercially reasonable discretion, certain important feature

Table 18: Sanitized lemmatized form of the noun chunk dependents with the highest increment in relative frequency per million in privacy policies in February 2021 compared to the pre-CCPA corpus (enforcement in July 2020).

Head	Dependents
collect	age, voice data, your direct input, access information, user information, follow required diagnostic data, follow additional information, california consumer privacy act, persistent identifier, diagnostic data, customer information, text, sample, contact and payment data, require diagnostic data, financial transaction data, installation date, performance, usage data, device and usage data, image, public information, search, certain category, personal data, child's online contact information, de-identify location data, "email header" information, motion activity, hardware capability, unique cookie id, usage and system operation data, any special category, additional website usage data
disable	other analytic tool, access, your account, ability, app's use, connect service, personalized feature, certain type, camera app's access, local device storage, interest-base ad, any user identification code, option, syncing, your access, feature, notification, display, automatic content recognition, see ad
enable	archiving, web application, optional diagnostic data, functionality, your web experience, share usage data, inclusion, virtual reality experience, certain account feature, tool, bulk consent feature, motion, fast loading, restore, other people, market research, voice command, mixed reality experience
personalize	advertising, publisher site, your feed or job recommendation, child
reject	cookies, change
send	your activity history, service announcement, your voice data, periodic promotional or informational email, diagnostic data, invitation, e-mail address, unique browser id, informational message, specific instruction, unsolicited commercial email, link, notice, direct message, informational communication, feedback, information request, connection request, spam, service notification, text message, report, location data, unsolicited email, notification, log, your search query, copy, error report, information, relate communication, instruction
share	result, account data, your profile data, your location, anything, your account, some de-identify data, all category, your e-mail address, your own list, non-personally identifiable information, common behavior, optional diagnostic data, all, you control, relevant data, video, common attribute, insight, publisher content, those, offer, relevant third party, limited, aggregated information, link, your phone screen, identification, common component, service provider, your phone number, your private personal data, non-personal information, our service, any sensitive personal information, someone else's creative content
use	your contact, automate process, multiple account, payment instrument, facial recognition, laptop, clear browse history, appropriate safeguard, your payment card information, local device storage, payment information, device-base recognition, your postal mailing contact information, web form, publishers' site, your tweet, subscription services, voice and text data, professional or employment relate personal information, device's setting app, automatic scan technology, opt-out tool, facial recognition technology, member' data, your keyboard, automate system, first name, digital service, error report, unique identifier, account, device-base speech recognition, log data

Table 19: Sanitized lemmatized form of the noun chunk dependents with the highest increment in relative frequency per million in privacy policies in January and February 2023 after the CPRA taking effect compared to the pre-CCPA corpus (enforcement in July 2020).

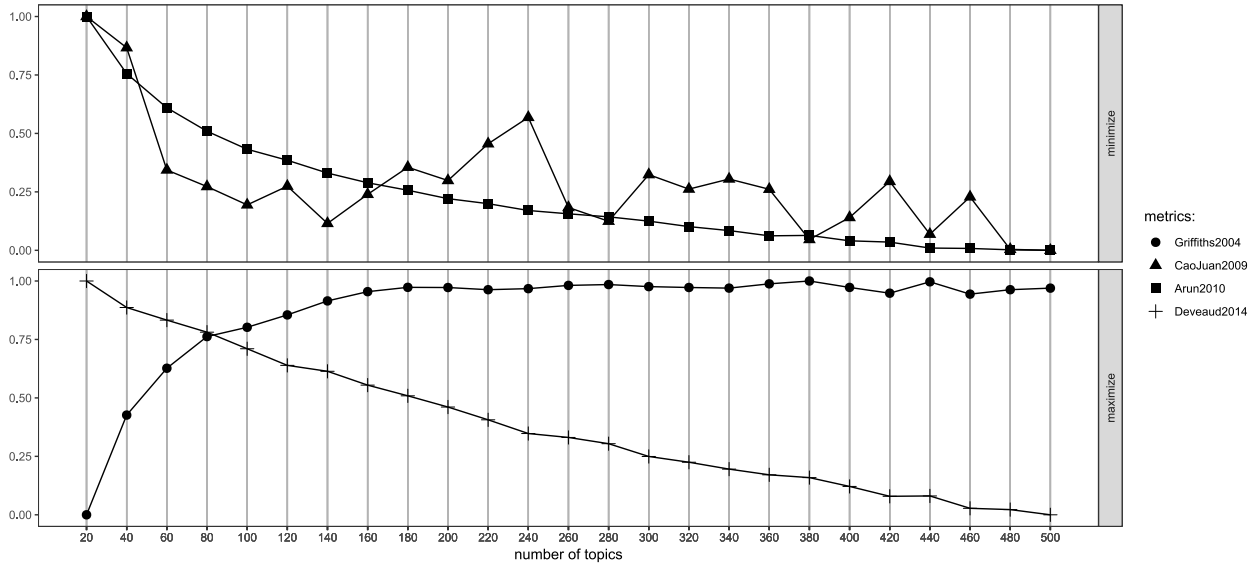
Head	Dependents
collect	additional personal data, age, your direct input, process, certain personal data, geolocation data, california consumer privacy act, precise location data, persistent identifier, diagnostic data, identifier, payment and billing information, contact data, anonymous usage data, sensitive personal information, device identifier, purpose, website, commercial information, child's online contact information, metadata, unique cookie id, limited information, traffic, device information, business or commercial purpose, billing, category
disable	feature, option, app's use, connect service, personalized feature, your account, local device storage, syncing, notification, camera app's access, some cookie, automatic content recognition, functional cookie
enable	web application, secure login, gps feature, our advanced security setting, functionality, purpose, website, virtual reality experience, basic feature, collection, your continue use, employment and education data, fast loading, our user, marketing use, interest-base content, contact, feature, other people, specific functionality, service, voice chat, market research, gps location-base service, voice command, mixed reality experience
personalize	our communication, information, your use, child
reject	cookies
send	personalized offer, your voice data, any personal data, newsletter, invitation, advertising, electronic communication, informational message, link, service-relate message, email marketing, informational communication, transactional and administrative email, important update, marketing and promotional communication, service-relate communication, event invitation, spam, text message, periodic email, administrative email, transactional message, unsolicited email, notification, transactional email, compliance, commercial email
share	your personal information, user-generate content, your location, de-identify information, any personal data, non-personally identifiable information, video, personal information, any personal data, your usage activity, your photo, your data, your phone number, my personal information, non-personal information, any sensitive personal information, your name and mailing address, visitor' personal information, photo, certain device identifier, subscriber record, their contact list, hashed version
use	remarketing service, facial recognition, their own tracking technology, your payment card information, location-base service, local device storage, payment information, different technology, personal information, your postal mailing contact information, fully automate algorithm-base technology, web chat service, optional service, location information, professional or employment relate personal information, device's setting app, facial recognition technology, automate system, tracking technologies, digital service, unique identifier, location-base service, your content, research, usage data, http cookie, publicly available information, publisher network websites, authorized agent, certain information, visit information, sensitive personal information, functionality, previously collect information

I NUMBER OF TOPICS PER CORPUS

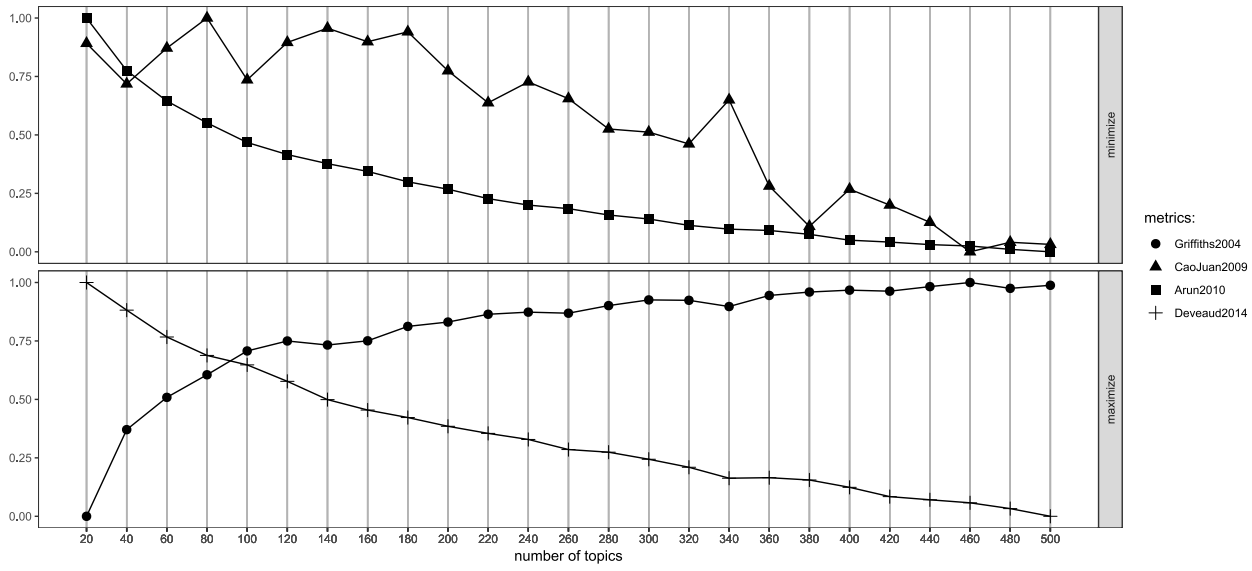
The following figures depict the output of the ldatuning package [54], which performs four statistical tests to determine the number of topics. These tests are based on the work of Arun et al. [6], Cao et al. [11], Deveaud et al. [16], and Griffith and Steyvers [27].

We tested the range from 20 to 500 with steps of 20 for each corpus and language separately. Not all four statistical tests necessarily pointed to the same optimal number of topics. Therefore, the near-optimal number of topics must be inferred from the resulting figures. For example, in both Figures 10a and 8a, the tests of Arun, Cao, and Griffith converge on 500 topics, while the test of Deveaud is not informative here.

Figure 7: Sample outputs of determining the number of topics for the German corpora.

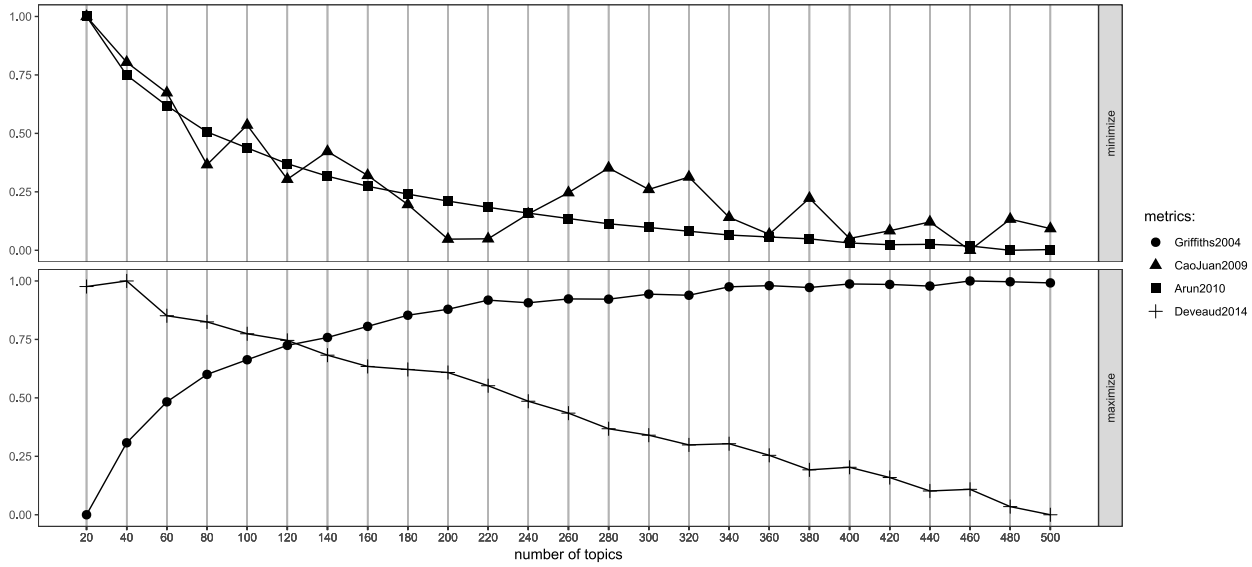


(a) GDPR corpus.

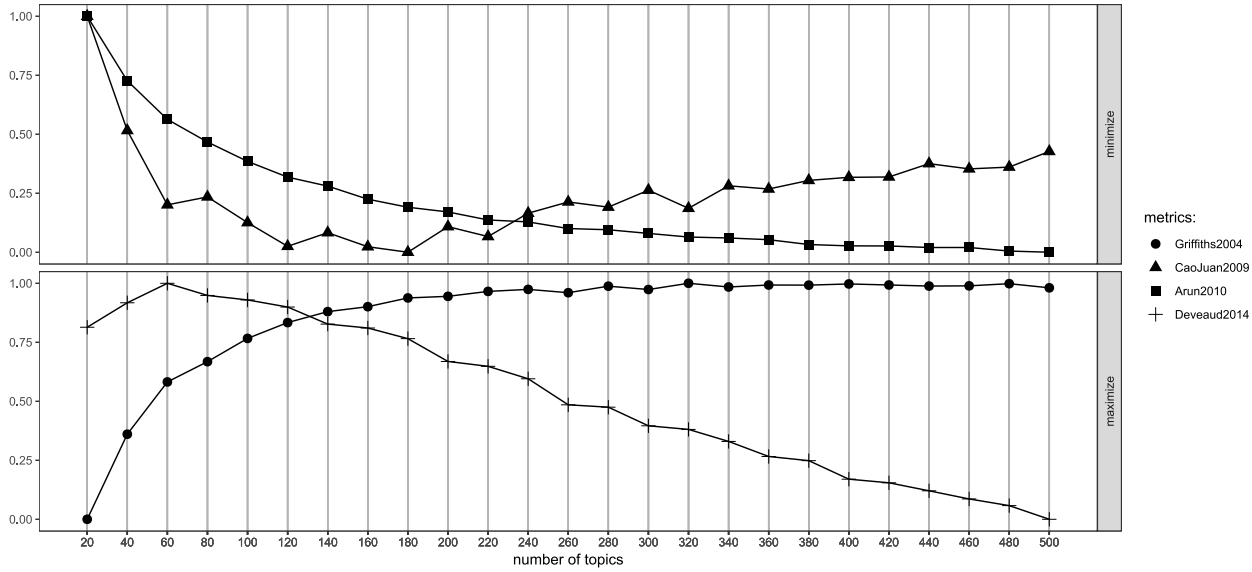


(b) CCPA/CPRA corpus

Figure 9: Sample outputs of determining the number of topics for the English corpora.



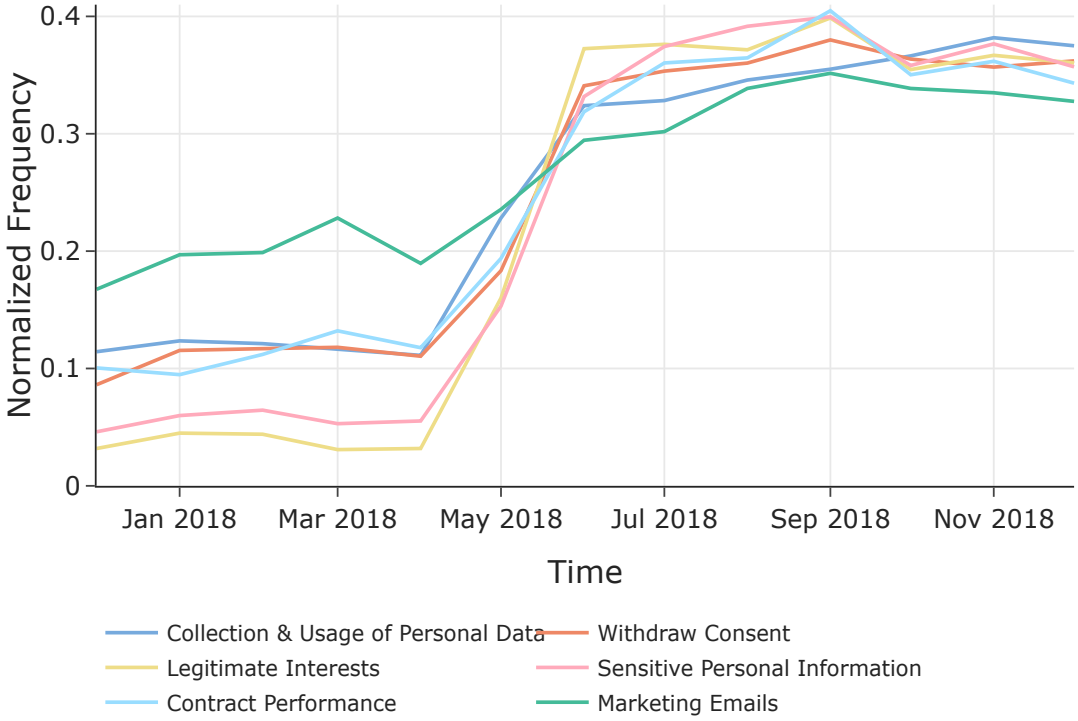
(a) GDPR corpus



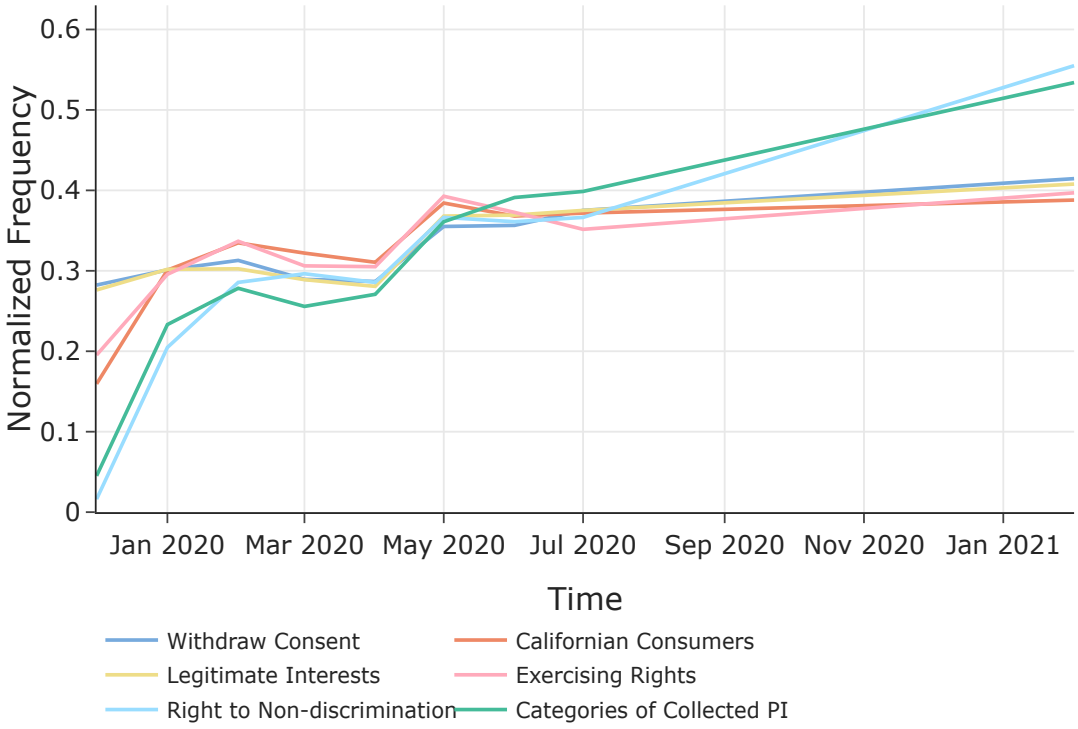
(b) Wagner's GDPR subset corpus

J TOPIC CHANGES IN WAGNER'S CORPUS

Figure 11: Topic trends in Wagner's subset corpora.



(a) GDPR subset corpus



(b) CCPA/CPRA subset corpus

K EXTENDED RESULTS OF THE KEYNESS ANALYSIS

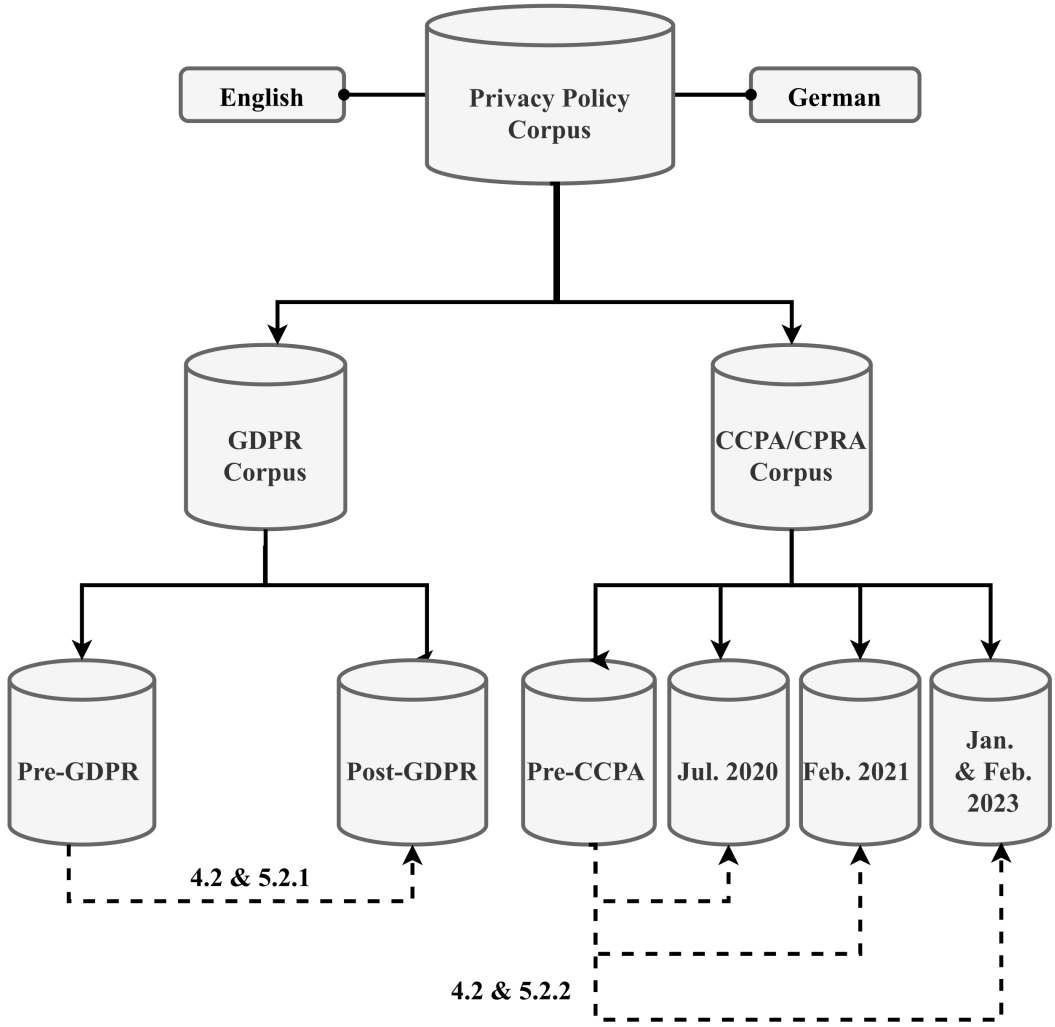
For completeness and due to space constraints, we present more results of the keyness analysis (Section 5.2) in the following.

Table 20: Log ratio (LR) and percentage difference (PercDiff) values of English phrases with a statistically significant increase in occurrence after GDPR enforcement.

	GDPR		Wagner's	
Phrase	**LR**	**PercDiff**	**LR**	**PercDiff**
data_portability	3.60	1,109.03	3.08	742.78
restrict_processing	3.45	991.74	2.71	552.81
readable_format	3.38	943.78	3.14	779.83
supervisory_authority	3.29	879.60	2.42	436.32
general_data_protection_regulation	3.23	839.19	2.92	656.12
right_to_withdraw_consent	3.14	781.77	2.39	424.06
machine_readable	3.11	762.63	1.81	249.55
legal_basis	3.11	761.47	2.77	581.03
object_to_processing_of_personal	3.09	750.24	2.63	518.85
perform_contract	3.06	735.84	2.40	426.21
lodge_complaint	2.97	685.68	3.13	774.94
right_to_object	2.86	623.68	2.56	490.99
enter_into_contract	2.85	618.94	1.85	261.47
legitimate_interest	2.71	555.95	2.80	595.88
consent_to_process	2.64	523.17	2.42	433.76
right_to_receive	2.64	522.81	1.14	120.53
erasure	2.58	496.33	2.75	570.49
profiling	2.31	394.77	2.29	388.12
consent_at_time	2.29	389.36	2.26	378.89
applicable_datum_protection_law	2.25	375.79	2.11	332.80
base_on_consent	2.18	354.09	1.89	270.94
compliance_with_legal	2.10	328.02	1.87	264.14
performance_of_contract	1.91	276.37	1.71	226.82
data_subject	1.89	270.63	3.01	707.77
request_access	1.89	270.06	0.60	51.74
rectification	1.80	248.32	2.93	659.36
exercise_of_right	1.63	209.76	1.50	181.82
contact_data	1.62	207.31	1.43	170.08
withdraw_consent	1.60	203.64	1.99	298.47
data_protection_officer	1.58	199.07	2.09	324.31
legally_require	1.56	195.31	0.61	53.05
make_complaint	1.37	158.87	1.01	100.80
necessary_for_purpose	1.37	158.15	1.37	158.41
right_to_request	1.30	146.87	1.20	129.47
retention_period	1.27	141.23	1.58	198.65
data_protection_authority	1.22	133.27	2.13	336.08
transfer_personal_data	1.11	115.38	2.21	363.48
process_personal_information	1.10	114.21	0.89	84.96

L DATA SETS AND COMPARISONS

Figure 13: Overview of and relationship between the privacy policy corpora used in our analyses. The dotted arrows indicate the compared subset corpora, as well as the respective part of the Results section for easier reference.



Corpus	Time frame
Pre-GDPR	2017-06-12--2018-05-18
Post-GDPR	2018-05-25--2018-12-28
Pre-CCPA	2019-12-09--2020-06-23