A Tale of Two Regulatory Regimes: Creation and Analysis of a Bilingual Privacy Policy Corpus

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[*◇]**

[*]Carnegie Mellon University, Pittsburgh, PA, USA, {siddhana, aravicha}@andrew.cmu.edu, {sadeh, rexc}@cs.cmu.edu, tristan.dhellemmes@laposte.net

[†]University of Münster, Münster, Germany, {henry.hosseini, thomas.hupperich}@wi.uni-muenster.de

[‡]Ruhr University Bochum, Bochum, Germany, {christine.utz, martin.degeling}@rub.de

[‡‡]AWS AI, Santa Clara, CA, USA, vinayshk@amazon.com

[‖]Clark University, Worcester, MA, USA, peter.garth.story@gmail.com

[¶]University of Massachusetts Amherst, Amherst, MA, USA, jmangat@umass.edu

[§]Fordham University School of Law, New York, NY, USA, tnorton1@fordham.edu

[††]Pennsylvania State University, University Park, PA, USA, shomir@psu.edu

Abstract

Over the past decade, researchers have started to explore the use of NLP to develop tools aimed at helping the public, vendors, and regulators analyze disclosures made in privacy policies. With the introduction of new privacy regulations, the language of privacy policies is also evolving, and disclosures made by the same organization are not always the same in different languages, especially when used to communicate with users who fall under different jurisdictions. This work explores the use of language technologies to capture and analyze these differences at scale. We introduce an annotation scheme designed to capture the nuances of two new landmark privacy regulations, namely the EU's GDPR and California's CCPA/CPRA. We then introduce the first bilingual corpus of mobile app privacy policies consisting of 64 privacy policies in English (292K words) and 91 privacy policies in German (478K words), respectively with manual annotations for 8K and 19K fine-grained data practices. The annotations are used to develop computational methods that can automatically extract "disclosures" from privacy policies. Analysis of a subset of 59 "semi-parallel" policies reveals differences that can be attributed to different regulatory regimes, suggesting that systematic analysis of policies using automated language technologies is indeed a worthwhile endeavor.

Keywords: privacy policy, privacy policy corpus, text corpus, bilingual, GDPR, CCPA, CPRA

1. Introduction

Privacy policies are the primary mechanism by which organizations disclose their data practices. These disclosures are intended to inform consumers about how their data will be handled and about their rights in relation to this data, such as the right to review or delete data or to restrict its collection and use (by, e.g., an "opt-in" or "opt-out"). However, privacy policies are often long, vague, ambiguous, and difficult to read (Reidenberg et al., 2015). Legislative bodies have responded by imposing new requirements about the information privacy policies must disclose and by enshrining the privacy rights of consumers from whom data is collected. Because of the sheer number of privacy policies available on the Internet, the impact of these new regulations on data collection and use practices, particularly including specific protections afforded to consumers, has eluded systematic analysis. This work explores the use of language technologies to systematically capture and analyze this impact at scale, with a focus on comparing provisions offered by the same technology (namely, mobile apps) in jurisdictions subject to different privacy regulations. Specifically, we make the following contributions:

Firstly, we extend the OPP-115 annotation scheme introduced in Wilson et al. (2016a) to capture concepts and requirements introduced in new privacy regulations — the EU's General Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA). These include the purpose for and mode of data collection as well as data subject rights. We discuss the challenges of collecting fine-grained annotations in the ambiguous and vague text of privacy policies. We also report on the process that we took to collect annotations in two different languages.

Secondly, we introduce a manually annotated corpus[1] of 64 privacy policies (292K words total) in English and 91 policies in German (478K words total) for smartphone apps sampled from the Google Play Store. It includes a subcorpus of 59 "semi-parallel" policy pairs, i.e., policies in English (aimed at app users in the US) and German (aimed at app users in Germany) for

[**]These co-authors contributed to this work while working as students or visitors at Carnegie Mellon University.

[◇]Arora and Sadeh are the corresponding authors of the present article.

[1]MAPP corpus will be accessible at `https://usableprivacy.org/data` sometime in June 2022.

the same mobile app. Upon observing that English and German privacy policies can significantly differ, we analyze differences in privacy protections offered to users in different jurisdictions (US versus EU) and evaluate computational techniques for such analyses. Our corpus respectively includes manual annotations of 8K and 19K fine-grained data practices (e.g., who collects the data, what type of data, and with whom it is shared) in English and German, intended to facilitate the development of computational methods for automated analysis of privacy disclosures[2] at scale. We emphasize that corpora containing non-English text are essential to ensure such methods for privacy policy analysis are also available to users outside the English-speaking world. To the best of our knowledge, our corpus is the first semi-parallel bilingual resource for privacy policy text and a first step towards analyzing differences in privacy policies between languages and regulatory regimes.

Finally, we report on the initial development and evaluation of classifiers to automatically annotate English and German privacy policy disclosures based on our new scheme. In particular, we present analyses of differences in disclosures between English and German privacy policy text, focusing on our semi-parallel corpus of 59 policies. Our results suggest that the EU's GDPR, as reflected in privacy disclosures in the text of German privacy policies, has had a moderating effect on the sharing of data with third parties and has contributed to more granular disclosures of data collection practices compared to the disclosures found in English privacy policies for the US public. Despite important differences between privacy disclosures in the US and in the EU, our analysis also suggests the existence of spill-over effects: more stringent EU regulations may also benefit US residents as a number of organizations do not differentiate between EU and non-EU residents, possibly due to the overhead involved. After demonstrating the performance of our classifiers and showing interesting differences in privacy disclosures between English and German privacy policies, we use our system to run this analysis at scale on 22,329 English and 1,864 German privacy policies retrieved from websites. We discuss what percentage of German privacy policies satisfy GDPR requirements in terms of including a discussion of the legal basis for the processing of data, and what percentage of websites also have such discussions in their English policies. We believe that such an analysis can help regulators and policy makers evaluate the impact of new privacy disclosure requirements.

2. Related Work

Policy Analysis Privacy policies are known to be long and difficult to read (McDonald and Cranor, 2008; Reidenberg et al., 2015). As a result, there has been an interest in developing automated techniques to analyze the text of privacy policies at scale (Ammar et al., 2012; Liu et al., 2014; Sadeh et al., 2013; Wilson et al., 2016b; Ravichander et al., 2021), with initial work focusing on analyzing their readability (Fabian et al., 2017; Massey et al., 2013; Meiselwitz, 2013; Ermakova et al., 2015). Recently, there has been interest in the automated collection of privacy policies (Zimmeck et al., 2017; Story et al., 2019; Hosseini et al., 2021) and the identification of their described data practices in a structured format (Costante et al., 2012a; Ammar et al., 2012; Costante et al., 2012b; Liu et al., 2014; Ramanath et al., 2014; Wilson et al., 2016a; Kumar et al., 2019; Bui et al., 2021). Such techniques have been used to help users navigate privacy policies[3], to automatically extract opt-out choices (Kumar et al., 2020), to check for compliance with regulatory requirements (Zimmeck et al., 2017; Story et al., 2019; Zimmeck et al., 2019), and to enable downstream applications such as question answering (Harkous et al., 2018; Ravichander et al., 2019; Ahmad et al., 2020; Sathyendra et al., 2017). Our work aims to extend this line of research (see also (Sørensen and Kosta, 2019; Urban et al., 2018; Degeling et al., 2019)) by supporting the systematic analysis of privacy policies in multiple languages, so as to evaluate the impact of new regulatory requirements at scale and analyze differences in the text across jurisdictions (since privacy policies are often localized into different languages).

Multilingual Text Classification There has been a growing effort in the language technologies community to create benchmark resources in non-English languages (Bender, 2011). In recent years, such resources have been constructed for tasks including natural language inference (Hu et al., 2020; Conneau et al., 2018), question answering (Artetxe et al., 2019; Clark et al., 2020), and sentiment analysis (Cieliebak et al., 2017; Amiri et al., 2015; Apidianaki et al., 2016; Vincent and Winterstein, 2013). Our work contributes to this effort by creating the first semi-parallel bilingual corpus of fine-grained data practice annotations to support the comparison of privacy policies across different languages and regulatory regimes.

3. Corpus Creation and Annotation

3.1. Privacy Policy Selection

With the explosion in smartphone ownership and the vast array of sensitive data collected by smartphones, the data practices of mobile apps have become particularly fertile ground for privacy research (e.g., (Lin et al., 2012; Almuhimedi et al., 2015; Kelley et al., 2013; Liu et al., 2016; Zimmeck et al., 2017)). This motivates our focus on the privacy policies of mobile apps. We collected mobile app privacy policies by sampling from a representative subset of app categories in the Google

[2]Privacy disclosures refer to statements made by an entity in its privacy policy about the type of data it collects, the purposes for collection, choices available to data subjects and other relevant data practices.

[3]https://explore.usableprivacy.org/?view=machine

Play Store, namely Family, Finance, Games, Medical, Productivity, Shopping, Social, and Sports.

To more efficiently find apps with German policies, we relied on the "Recommended for you" Google Play Store feature and configured a server with a German IP address and browser locale. In each app category, we used Mozilla Firefox in Private Browsing mode (to avoid personalization) to retrieve the following app metadata: app store listing URL, app category, name, number of downloads, and privacy policy link. The resulting list comprised 1,161 apps. Two German-speaking authors manually accessed all app listings and followed the privacy policy links for each. This usually led to an English-language version on the publisher's website. We manually examined the policy links and ensured that, if multiple policies were available in English, the US policy was downloaded. The website was then searched for a German privacy policy.

Some apps only had privacy policies in English or German, and 261 apps did not have a privacy policy in either language. Because we are interested in comparing privacy protections afforded to US and German users, we focused on 222 apps which had policies in both German and English. Among these apps, 101 had identical policy pairs (e.g., Facebook Messenger and Facebook LITE) to the apps that had been already included in our corpus. We eliminated these apps from our corpus, along with 5 other apps where an auto-translate button or disfluencies hinted at them being the result of automated translation, a situation where linguistic analysis has little to offer. The whole process yielded a total of 116 privacy policy pairs. A graphical depiction of the described process is provided in Figure 3 in the appendix. The privacy policy links of these 116 apps were automatically crawled and downloaded, intentionally keeping any extracted HTML content surrounding the policy text to train our classifiers to distinguish between privacy policy text and irrelevant text. We manually inspected the extracted documents for download errors. The process was carried out in July 2020.

In our analysis, we distinguish between policies with sections that explicitly single out EU residents from policies that do not. This is because the former will typically grant EU residents privacy protections that they may not grant to non-EU residents. Among our 116 policies, 27 policies were identified as explicitly singling out EU residents. Occasionally policies also single out US residents such as Californians or other specific categories of consumers such as children. This will be briefly discussed in Section 6.1.

We also analyzed policies for the presence of markers indicative of GDPR, as proposed by Degeling et al. (2019). Specifically, we use the following keyphrases in English and German as indicators that a policy has been written in response to GDPR requirements: *data protection officer, legal basis, legitimate interest, rectification, erasure, data portability,* and *supervising authority*. We consider a privacy policy to be "GDPR-aware" if it contains at least three of these keyphrases or acknowledges the EU–US Privacy Shield[4]. Out of our policy pairs for 116 apps, 63 English policies and 74 German policies were found to be "GDPR-aware."

3.2. Annotation Scheme and Process

Previous work on English-language privacy policies by some the co-authors and their collaborators produced the OPP-115 annotation scheme and corpus (Wilson et al., 2016a). We updated and refined this scheme to capture important concepts and protections introduced in two new landmark privacy regulations, namely the EU's GDPR and California's CCPA/ CPRA. Below, we review the structure of OPP-115-style annotations, then describe the genesis of the modified annotation scheme used to create our corpus. We refer to it as the *MAPP annotation scheme*.

The OPP-115 annotation scheme focuses on the identification of *data practices*, which are statements about personal data collection, use, sharing, and other related activities. The scheme places each data practice in one of ten themes ("categories") identified in 2015, but we observe that some have diminished in frequency in privacy policies, have become outdated (e.g., the Do Not Track standard), or lack a unifying structure (e.g., the *Other* category). Our work focuses on refining annotations of two categories of data practices from the OPP-115 annotation scheme, namely *First-Party Collection/Use* and *Third-Party Collection/Use*. Together, these two practices account for more than 60% of the annotations in the OPP-115 corpus (Wilson et al., 2016a) and are a central component of all privacy policies. We redefine these categories of data practices as follows:

1. First Party Collection/Use: Privacy practices describing data collection/use by the organization that published (or "controls") the mobile app.

2. Third-Party Collection/Use: Privacy practices describing data sharing with third parties or data collection by third parties. A third party is a organization other than the first-party organization that published (or "controls") the mobile app.

Similar to the OPP-115 annotation scheme, each data practice in our scheme is associated with a set of category-specific *attributes* that detail the practice. Each attribute is instantiated with a choice of one of several predetermined *values*, unless the attribute is defined as optional. To ground the data practice in textual evidence, annotators justify their attribute and value selections by associating them with spans of text (see Figure 1). A sample annotation from our corpus is shown

[4]While the European Court of Justice invalidated the US-EU Privacy Shield in July 2020, the privacy policies that acknowledge Privacy Shield in July 2020, the time we collected our policies, can be assumed to have been GDPR-aware, given that Privacy Shield was designed to help US companies comply with GDPR.

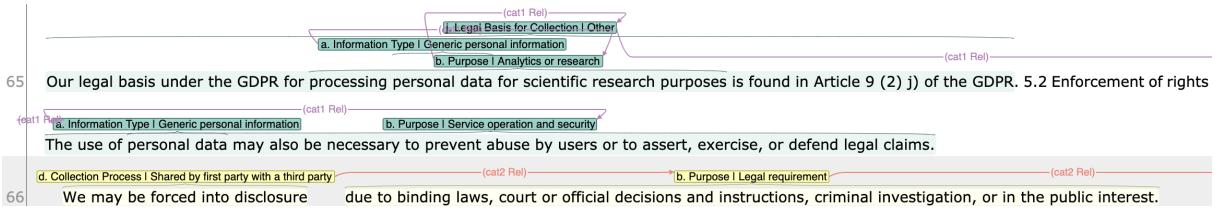

Figure 1: Screenshot of our annotation tool. Green and yellow tags indicate first-party and third-party data practices, respectively. In each tag the pipe character separates the annotated attribute and its value. Red lines link annotations to fully describe a data practice.

in Table 1. We provide more details about the definitions of attributes in our annotation scheme in Table 7 in the appendix. Working with privacy law experts from the US and EU, we refined the original OPP-115 attributes and values for the two selected categories into a scheme suitable for GDPR and CCPA/CPRA. Although OPP-115 broadly captures the principles of GDPR (Poplavska et al., 2020) and CCPA/CPRA, our goal was to model these new regulations' requirements at a finer level of detail. In OPP-115, the two categories we focus on had 14 attributes and 89 values, whereas MAPP has 19 attributes and 124 values.[5]

"Practice":	First Party Collection/Use
"Attribute":	Information Type
"Value":	Location
"selectedText":	"information about the purchase or transaction"
"Practice":	Third-Party Collection/Use
"Attribute":	Collection Process
"Value":	Shared by first party with third party
"selectedText":	"we share information with"

Table 1: Sample span annotations from our MAPP corpus for a segment of Instagram.com's privacy policy.

We configured the INCEpTION annotation platform (de Castilho et al., 2018) for annotators to apply our scheme. Figure 1 shows a screenshot of the annotation tool. Annotators could label text spans for a data practice across sentences and create arbitrarily many data practices per sentence, reflecting the flexible information density of these documents.

To manually annotate the privacy policies, we recruited teams of law students in Germany and the US, who respectively had strong fluency in German and English. We provided our annotation scheme in English and German to annotators working in these respective languages. The two groups of annotators comprised 10 English-speaking and 12 German-speaking law students recruited at universities in the US and Germany. Annotators spent an average of 1 hour and 52 minutes per policy. They were trained by reviewing the annotation scheme in a virtual meeting with the authors, which included training videos about the usage of the annotation tool. Annotators worked independently but conferred with the authors weekly to discuss complex cases. The researchers would also review cases of disagreement between annotators to mitigate confusion. Due to differences in annotators' availability and speed, we ultimately obtained 91 German and 64 English fully annotated privacy policies, i.e., with three annotations each, with an intersection of 59 policies. We refer to the overall corpus of German and English annotations as the *MAPP* (Multilingual Annotation of Privacy Policies) Corpus, and the semi-parallel subcorpus of 59 English-German policy pairs as *MAPP-59*.

4. Corpus Composition

4.1. Annotations

Table 3 shows annotation statistics. Our annotators identified approximately 26K text spans for the 64 English policies and 39K for the 91 German policies, with 8K annotated data practices for English and 19K for German. Compared to the OPP-115 Corpus (Wilson et al., 2016a) (200 data practices per policy), our corpus contains fewer annotated data practices per policy.[6] However, our corpus, with 155 annotated policies, is comparable in size (≈115 policies) to previous corpora (Wilson et al., 2016a; Ravichander et al., 2019; Ahmad et al., 2020). We also compute the percentage of segments where at least 2 out of 3 annotators found at least one annotation span for a given category/attribute, which we denote as the *coverage* of that category or attribute. Our coverage of first-party and third-party data practice categories is also similar to OPP-115 (Wilson et al., 2016a). A detailed comparison with prior corpora is shown in Table 2.

As mentioned in Section 3.2, our corpus includes a semi-parallel subcorpus of English and German privacy policies for 59 mobile apps available in the US and the EU. With MAPP-59, we aimed to identify how the same app differs in privacy policy disclosures across two different languages. This corpus helped to understand differences arising from (1) different regulatory regimes in which the app operates or (2) specific lin-

[5]Note that these counts contain duplicate attributes and values across the two categories. For example, roughly half of the attributes are the same for both data practice categories in OPP-115 and MAPP, and they can take the same values.

[6]The MAPP corpus covers only two data practice categories, compared to the 10 categories covered in the OPP-115 dataset. However, the total number of data practices within these two categories are comparable in the two corpora.

	PrivacyQA (Ravichander et al., 2019)	PolicyQA (Ahmad et al., 2020)	OPP-115 (Wilson et al., 2016a)	MAPP
Documents	35	115	115	155
Task	QA	QA	Text classification	Text classification
Privacy policy source	Mobile applications	Websites	Websites	Mobile applications
Annotator	Domain experts	Mechanical Turkers	Domain experts	Domain experts
Annotation scheme	-	-	OPP-115	OPP-115 refinement for GDPR / CCPA
#Attributes	-	-	14	19
#Values	-	-	89	124
Coverage (first party)	-	-	0.27	0.31 (en) / 0.32 (de)
Coverage (third party)	-	-	0.21	0.14 (en) / 0.12 (de)
Languages	English	English	English	English, German

Table 2: Comparison of the MAPP Corpus with other privacy policy corpora. Our corpus is comparable in size, coverage, and annotation scheme and introduces bilinguality. The number of attributes and values refer to the OPP-115 categories focused on in this work (First Party Collection/Use and Third-Party Collection/Use).

	English	German
Documents	64	91
Words	292,576 (4,571)	478,560 (5,258)
Data Practices	8,475 (132)	19,388 (213)
Attributes	16,300 (254)	29,356 (323)
Text Spans	26,221 (409)	39,809 (437)

Table 3: MAPP Corpus statistics by language, with per-policy mean statistics in brackets.

Category / Attribute	English		German	
	Coverage	FK	Coverage	FK
First Party	0.31	0.61	0.33	0.52
Third Party	0.14	0.52	0.13	0.47
Inform. Type	0.29	0.54	0.28	0.48
Purpose	0.26	0.63	0.23	0.58
Collect. Process	0.20	0.44	0.12	0.33
Legal Basis	0.05	0.37	0.07	0.39
3rd Party Entity	0.11	0.49	0.10	0.36

Table 4: Inter-annotator agreement (Fleiss' Kappa, FK) for data practices and attributes in the MAPP corpus.

guistic differences that may change how readers interpret the language.

4.2. Inter-Annotator Agreement

As in prior work, we often observed ambiguity in privacy policies, with even law experts disagreeing about interpretations of their text (Reidenberg et al., 2015). We automatically divided privacy policies into segments based on each policy's HTML (Zimmeck et al., 2019); each segment can be thought of as roughly equivalent to a paragraph. We analyze the agreement between annotators on the segment level since annotations often span sentence boundaries. We associated a given data practice with a segment if the segment had at least one text span associated with that data practice. We then computed Fleiss' Kappa agreement values (Fleiss, 1971) between annotators for a given language at the segment level, separately for each level of the annotation scheme (i.e., data practice category, attribute, and value). In this work, we focus on building classifiers to automatically detect attributes and values for which we had sufficient agreement between annotators (Fleiss' Kappa > 0.3) and a sufficient number of training instances (coverage ≥ 0.02).

Table 4 displays the resulting agreement values at the data practice category and attribute levels; Table 8 in the Appendix shows agreements at the value level. We observe that the English annotators displayed a higher level of agreement than the German annotators. Annotators also tended to exhibit lower agreement for more granular annotations (e.g., specific attribute values) and categories and attributes with lower coverage (e.g., *Collection Process*, *Legal Basis for Processing*, and *Third Party Entity*). The same observation also applies to attribute values. We now briefly discuss frequent sources of annotator disagreement.

(1) "*[...]* ***this data will stay on your device unless you enable the functionality of sharing within the app.***
If you opt-in the sharing functionality, we can also ask you to create a ***publicly*** *visible [COMPANY] account [...]*"

The above statement illustrates an example of a disclosure with which annotators struggled, in part due to limitations of our annotation scheme. This disclosure differentiates between collection and storage of information only on the user's device versus on first-party servers. Our annotation scheme does not distinguish between these practices. Instead, annotators were instructed to annotate both practices as *First Party Collection/Use*. However, this particular disclosure states that a user has to opt-in before their data can be stored on first-party servers. It further states that users who opt-in can also agree to have a publicly visible account,

which corresponds to a *Third Party* sharing practice. A more accurate annotation for this segment includes two separate annotations, the first identifying the *First Party Collection/Use* practice with attribute *Choice Type* and value *Opt In* (among others), and the second identifying a *Third Party Sharing* practice. Annotators sometimes struggled with such cases of two practices being disclosed in conjunction.

(2) *"[...] collects 2 basic type of information [...] (1)* ***personally identifiable information*** *and (2) non-personally identifiable information [...]"*

(2) illustrates another source of confusion: a disclosure that explicitly uses the term *personally identifiable information*. Annotators would often mark such a disclosure as having the attribute *Information Type* with the value *Personal Identifier*. However, since this term does not refer to unique identifiers, a more correct value for the attribute would be *Generic Personal Information*. While we instructed our annotators to make such annotations, they did not consistently do so.

5. Prediction of Policy Structure

To automate the annotation of privacy policies at scale, we built classifiers by fine-tuning state-of-the-art language models on our corpora. In what follows, we show our classifiers' ability to extract useful information from privacy policy text.

5.1. Classifier for Information Extraction

Our MAPP corpus consists of 3,976 segments from 64 privacy policies in English and 6,871 segments from 91 privacy policies in German. Our classification task has three sub-tasks for identifying (1) the disclosure of data practice categories (i.e., *First Party Collection/Use* and *Third Party Collection/Use*) in a segment, (2) the disclosure of data practice attributes (e.g., *Information Type*), and (3) the disclosure of values associated with these data practice attributes in the segment.

We trained classifiers to identify *First* and *Third Party* data practice categories with the following five attributes: *Information Type*, *Purpose*, *Collection Process*, *Legal Basis for Processing*, and *Third-Party Entity*. We created gold standard annotations by labeling a segment with a category/attribute if two or more annotators agreed on that category/attribute. We further trained value-level classifiers for nine values of *Information Type*, five values of *Purpose*, two values of *Collection Process*, and one value of *Legal Basis for Processing* (Table 6). All of our classifiers were trained to predict the binary presence or absence of a single category or attribute/value. To build these classifiers, we fine-tuned pre-trained large-scale language models (Devlin et al., 2019): BERT for English policies and Multilingual BERT (M-BERT) for German policies. For the value-level classifiers, we also experimented with using the predictions made by data practice and attribute classifiers to create a binary feature vector. This vector was then concatenated with the BERT encoding as input to the final classifier layer. We describe parameter settings for our classifiers in Section 10.1 in the appendix.

Category / Attribute	English			German		
	P	R	F1	P	R	F1
First Party	0.84	0.69	0.76	0.69	0.78	0.73
Third Party	0.75	0.65	0.70	0.60	0.70	0.64
Inform. Type	0.71	0.72	0.71	0.67	0.77	0.71
Purpose	0.76	0.81	0.79	0.64	0.89	0.74
Collect. Process	0.61	0.58	0.60	0.55	0.77	0.64
Legal Basis for Processing	0.85	0.85	0.85	0.50	0.58	0.54
3rd Party Entity	0.68	0.54	0.60	0.43	0.72	0.54

Table 5: (Performance of BERT and M-BERT for predicting data practice *category* and *attributes* in English and German privacy policy segments. Precision, recall, and F1 values are for the positive class.

Attribute	Value	English			German		
		P	R	F1	P	R	F1
Information Type	Financial	0.77	0.67	0.71	0.49	0.95	0.65
	Contact inform.	0.78	0.56	0.65	0.73	0.84	0.78
	Location	0.47	0.60	0.53	0.47	0.83	0.60
	Demographic data	0.75	0.63	0.69	0.50	1.00	0.67
	User online activities	0.66	0.37	0.47	0.48	0.58	0.52
	IP address and device IDs	0.81	0.65	0.72	0.64	0.85	0.73
	Cookies and tracking elements	0.71	0.79	0.75	0.44	0.59	0.51
	Computer inform.	0.71	0.71	0.71	0.70	0.72	0.71
	Generic personal information	0.57	0.65	0.61	0.61	0.51	0.55
Purpose	Essential service or feature	0.61	0.40	0.48	0.56	0.28	0.38
	Advertising or marketing	0.48	0.75	0.59	0.52	0.78	0.62
	Analytics or research	0.74	0.71	0.73	0.57	0.79	0.66
	Service operation and security	0.58	0.45	0.51	0.67	0.57	0.61
	Legal requirement	0.69	0.52	0.59	0.45	0.65	0.53
Collection Process	Shared by 1st party w/ 3rd party	0.53	0.40	0.45	0.71	0.25	0.37
	Collected on 1st party website/app	0.49	0.58	0.53	0.60	0.40	0.48
Legal Basis for Process.	Legitimate interests of first/third party	0.82	0.82	0.82	0.53	0.68	0.60

Table 6: Performance of BERT and M-BERT for predicting *values* associated with attributes in English and German privacy policy segments. We show precision, recall, and F1 values for the positive class.

5.2. Results

We created a random train-test split for both languages, respectively with 52/12 and 75/16 policies for training/testing in English and German. Table 5 shows our classifiers' performance in predicting data practices and attributes. We report precision, recall, and F1 for the presence of a data practice/attribute in a segment. Our classifiers for predicting the *First Party* and *Third Party* categories on English annotations had a macro F1 score of 0.83, similar to figures in prior work (Kumar

et al., 2019). Note that in Table 5, we only report F1 values for the positive class. However, we found that our classifiers were less accurate at tagging segments with data practice categories for German annotations than for English annotations. This may be attributed to lower agreement between German annotators. Our classifiers are fairly accurate in categorizing the *Information Type* and *Purpose* attributes in both languages. However, for the *Legal Basis for Processing* attribute, the English classifier performs substantially better.

Table 6 shows our classifiers' performance for values of the *Information Type* attribute. We found that attribute values like *Financial*, *Computer Information* and *IP Addresses & Device IDs*, which are associated with annotation spans containing more distinctive language (i.e., higher KL divergence with the vocabulary of text in privacy policies), yielded better classification performance. We observed that *Generic Personal Information* and *User Online Activities* had lower classification performance (Table 6), which may result from lower agreement between annotators than for other values. Furthermore, these values corresponded to longer annotation spans, requiring the more challenging identification of patterns spread across longer text spans. Performance for values of the *Purpose* attribute is shown in Table 6. Again, values with distinct vocabulary in their annotation spans (e.g., *Legal Requirement*) tended to yield better performance. Overall, performance for *Information Type* and *Purpose* values was better than values of other, more nuanced attributes (i.e., *Collection Process* and *Legal Basis for Processing*).

6. Comparing Data Practice Disclosures in English and German Policies

As mentioned in Section 3.1, the research reported here focuses on a set of 116 mobile apps with both English and German privacy policies. We aim to compare disclosures made in English privacy policies with the German privacy policies found for the same app.

Among the 116 mobile apps, our analysis shows that only 54% of apps had English privacy policies that were "GDPR-aware" (see Section 3.1). We further found that 43% of these apps specifically singled out EU residents, suggesting that they may possibly grant GDPR privacy protections only to EU residents. The other 57% likely extend their GDPR protections to non-EU residents. This suggests a spillover effect wherein US residents effectively receive the benefits of more stringent GDPR protections originally intended for EU residents. Such spillover likely makes sense for many organizations, considering the additional complexity they would otherwise have to manage if they were to differentiate between EU and non-EU residents. A similar practice is commonly seen in cookie options offered to non-EU residents in response to EU cookie regulations (the "ePrivacy Directive") originally intended for EU residents only (van Eijk et al., 2019).

Interestingly, nearly 36% of German policies do not acknowledge GDPR requirements. Following manual analysis, these policies were found to be similar to the corresponding English policies and belonged to apps originally introduced in the US. Even among the German policies that explicitly mention GDPR or its German equivalent, DSGVO (Datenschutz-Grundverordnung), about 33% do not seem to have language addressing the legal basis for the processing of data or text enumerating data subject rights, both required disclosures under GDPR. We suspect that, as of July 2020, many US app publishers had not yet realized they were subject to GDPR or made efforts to become fully GDPR compliant. This is consistent with earlier mobile app studies that compared the code of mobile apps with their privacy policy disclosures and found numerous indicators of non-compliance (Zimmeck et al., 2017; Zimmeck et al., 2019).

Similarly, for MAPP-59, we observe that 40 apps and 27 apps are "GDPR-aware" respectively in their English and German privacy policies. We further find that most apps which acknowledge GDPR rights in their German policy also mention these rights in the English version. Most of these privacy policies are for apps from European publishers. A total of 14 apps in MAPP-59 acknowledge GDPR in their German privacy policy but not in the corresponding English policy, which is consistent with the German and English policies respectively being intended for EU users and non-EU users. Beyond such high-level differences, we want to explore the use of our classifiers for a richer comparison of disclosures in the English and German privacy policies of mobile apps. Below, we report our analysis of differences in English and German privacy policy disclosures in our semi-parallel sub-corpus.

6.1. Comparative Analysis of Disclosures in English and German Privacy Policies

A manual analysis of our MAPP-59 policies was performed by two authors who are privacy scholars and fluent in both German and English. German policies were auto-translated into English with DeepL (DeepL GmbH, nd) to limit manual comparison to non-identical text. For each mobile app privacy policy pair, we also compute a BiLingual Evaluation Understudy (BLEU) score to measure the degree of difference between the German and English policy texts.[7] Scores range from 0 to 100, with lower scores indicating significant differences between the German and English texts for a given app and higher scores indicating more similar texts. Figure 2 in Appendix compares the numbers of both first- and third-party disclosures in policy pairs, where each policy pair is represented as two vertically aligned colored dots for that pair's BLEU

[7] We computed the average of the SacreBLEU scores (Post, 2018) (1) between the original English and the German policy auto-translated into English and (2) between the original German and the English policy auto-translated into German.

score. As expected, pairs with low BLEU scores tend to have greater differences in disclosure counts since these policy pairs have greater textual differences.

Our manual inspection of the policy pairs in MAPP-59 found that 13 pairs have substantially different policies, whereas 9 have nearly identical policy texts. Of the 9 nearly identical policies, 7 feature a separate section for EU users that provides more extensive protections to EU residents. Manual inspection of the policies that do not explicitly single out EU residents confirmed that US users of these apps often benefit from some of the more stringent privacy protections afforded by GDPR to EU residents. For instance, we observe that data subject rights like the "right to access" information collected by an app, provided by GDPR to EU residents, are also offered to US residents by 9 apps in our corpus. Among the 34 pairs that are not substantially different or nearly identical, we found 10 cases of additional sections specific to the English policies that primarily (6 out of 10 apps) contained CCPA-specific disclosures, in addition to single instances addressing COPPA or provisions related to policy changes. Two German policies contained additional sections compared to the English version, with one policy explicitly listing third-party entities with whom data is shared, and the other discussing the EU-US Privacy Shield. The presence of additional disclosures in English or German policy texts is consistent with their respective intended target audiences of US and EU-based/German app users.

At the sentence and paragraph level, our analysis identified additional differences in a total of 24 policy pairs. Most frequent (21 instances) were differences in wording that went beyond what one would reasonably expect from a mere translation. These included additional disclosures about types of data collected, third parties with whom data may be shared, different privacy rights, or ways to limit data collection. Again, these differences were generally consistent with disclosure requirements associated with GDPR (German policy text) and US regulations such as CCPA or COPPA (English policy text). These results suggest that it is feasible to automatically analyze the impact of privacy laws on differences in protection offered to users in different jurisdictions, as well as spillover effects. We also noticed grammatical differences such as the use of double negatives in German as opposed to positive statements in English policies (*Does/Does Not* attribute).

6.2. What can we learn from our classifiers?

One goal of our work was to automate the analysis and comparison of privacy policies in different languages. In the following, we describe an instance of our classifiers' ability to automatically identify the differences we discussed in Section 6.1 to answer policy questions. GDPR Article 6 prohibits collecting and processing personal data without a proper legal basis. Therefore, every category of personal data requires the legal basis to be clear and specific. To demonstrate what percentage of websites in Germany fulfill this requirement, we identified German privacy policies for 1,864 websites from the top 10,000 domains in Tranco (Le Pochat et al., 2019) using an established toolchain (Hosseini et al., 2021). Applying our trained classifiers to predict segments discussing *Legal Basis for Processing*, we found that 76% (1,414) of the websites satisfy this requirement. To determine if US residents also benefit from GDPR protections through spillover effects, we used the Princeton-Leuven Longitudinal Corpus of Privacy Policies (Amos et al., 2021) to obtain 22,359 English privacy policies for US-based websites from the time at which we downloaded the German ones (i.e., the second half of 2019). Using our classifiers, we observed that only 19% (4,245) provide this protection.

Further analyzing the 172 websites with both US and German privacy policies, we observed that 62 websites provide the legal basis for processing in both policies. 45 websites provide a legal basis only in the German policy, whereas 26 websites do so only in the US policy. This analysis shows that our corpus can be used to build automatic systems that can enable regulators to ask such policy questions and to systematically analyze the impact of jurisdiction-specific privacy regulations on the privacy disclosures made by apps in a given jurisdiction. We hope that such analyses can inform future public policy debates to understand the impact of new privacy requirements and in particular the way in which they are reflected in the text of privacy policies.

7. Conclusion

In this work, we introduced MAPP, the first bilingual corpus of privacy policies, to facilitate the development of classifiers that can automatically identify data practice disclosures in privacy policies of two different languages. We identified how privacy disclosures differ in policies published in English and German, and presented initial evidence of the effectiveness of our classifiers at automatically identifying these differences. We discussed how privacy regulations can account for some of these differences. For future work, we aim to build classifiers that can automatically identify more nuanced data practice attributes and values. We believe that this type of analysis can contribute to a more systematic understanding of the protections afforded in practice to consumers under different regulatory regimes and could ultimately help inform the development of more effective privacy regulations.

8. Acknowledgment

This research was supported in part by the National Science Foundation Secure and Trustworthy Computing program (CNS-1330596, CNS-1330214, CNS-1914486, CNS-1914444). For additional details on the "Usable Privacy Policy Project" and the "Automatically Answering People's Privacy Questions" under which this work was performed, see: `https://usableprivacy.org`.

9. Bibliographical References

Ahmad, W., Chi, J., Tian, Y., and Chang, K.-W. (2020). PolicyQA: A Reading Comprehension Dataset for Privacy Policies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP '20, pages 743–749. ACL.

Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., Cranor, L. F., and Agarwal, Y. (2015). Your Location Has Been Shared 5,398 Times! A Field Study on Mobile App Privacy Nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 787–796. ACM.

Amiri, F., Scerri, S., and Khodashahi, M. (2015). Lexicon-based Sentiment Analysis for Persian Text. In *Proceedings of Recent Advances in Natural Language Processing*, pages 9–16. ACL.

Ammar, W., Wilson, S., Sadeh, N., and Smith, N. A. (2012). Automatic categorization of privacy policies: A pilot study. Technical Report CMU-ISR-12-114, Carnegie Mellon University.

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and Mayer, J. (2021). Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of The Web Conference 2021*, WWW '21, page 22. ACM.

Apidianaki, M., Tannier, X., and Richart, C. (2016). Datasets for Aspect-Based Sentiment Analysis in French. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC '16, pages 1122–1126. European Language Resources Association.

Artetxe, M., Ruder, S., and Yogatama, D. (2019). On the cross-lingual transferability of monolingual representations. arXiv ePrint 1910.11856.

Bender, E. M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3):1–26.

Bui, D., Shin, K. G., Choi, J.-M., and Shin, J. (2021). Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110.

Cieliebak, M., Deriu, J. M., Egger, D., and Uzdilli, F. (2017). A Twitter Corpus and Benchmark Resources for German Sentiment Analysis. In *Proceedings of the 5th International Workshop on Natural Language Processing for Social Media*, SocialNLP '17, pages 45–51. ACL.

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. (2018). XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 2475–2485. ACL.

Costante, E., den Hartog, J., and Petković, M. (2012a). What websites know about you. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 146–159. Springer.

Costante, E., Sun, Y., Petković, M., and den Hartog, J. (2012b). A Machine Learning Solution to Assess Privacy Policy Completeness. In *Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, WPES '12, pages 91–96. ACM.

de Castilho, R. E., Klie, J.-C., Kumar, N., Boullosa, B., and Gurevych, I. (2018). Linking Text and Knowledge using the INCEpTION annotation platform. In *Proceedings of the 14th eScience IEEE International Conference*, eScience '18, pages 327–328. IEEE Press.

DeepL GmbH. (n.d.). DeepL Translate.

Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F., and Holz, T. (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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers) of *NAACL-HLT '19*, pages 4171–4186. ACL.

Ermakova, T., Fabian, B., and Babina, E. (2015). Readability of Privacy Policies of Healthcare Websites. In *Wirtschaftsinformatik Proceedings 2015*, WI '15, page 73.

Fabian, B., Ermakova, T., and Lentz, T. (2017). Large-Scale Readability Analysis of Privacy Policies. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 18–25. ACM.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. (2018). Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *Proceedings of the 27th USENIX Security Symposium*, USENIX Security '18, pages 531–548. USENIX Association.

Hosseini, H., Degeling, M., Utz, C., and Hupperich, T. (2021). Unifying Privacy Policy Detection. *Proceedings on Privacy Enhancing Technologies*, 2021:480–499.

Hu, H., Richardson, K., Xu, L., Li, L., Kübler, S., and Moss, L. (2020). OCNLI: Original Chinese Natural Language Inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Lan-*

guage Processing, EMNLP '20, pages 3512–3526. ACL.

Kelley, P. G., Cranor, L. F., and Sadeh, N. (2013). Privacy as Part of the App Decision-Making Process. In *Proceedings of the 31st Annual ACM Conference on Human Factors in Computing Systems*, CHI '13, pages 3393–3402. ACM.

Kumar, V. B., Ravichander, A., Story, P., and Sadeh, N. (2019). Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *Proceedings of the 2019 AAAI Spring Symposium on Privacy Enhancing AI and Language Technologies*, PAL '19. AAAI Press.

Kumar, V. B., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., Schaub, F., and Sadeh, N. (2020). Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of the 29th International World Wide Web Conference*, WWW '20, pages 1943–1954. ACM.

Le Pochat, V., Van Goethem, T., Talajizadehkhoob, S., Korczyński, M., and Joosen, W. (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.

Lin, J., Amini, S., Hong, J. I., Sadeh, N., Lindqvist, J., and Zhang, J. (2012). Expectation and Purpose: Understanding Users' Mental Models of Mobile App Privacy through Crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 501–510. ACM.

Liu, F., Ramanath, R., Sadeh, N., and Smith, N. A. (2014). A Step Towards Usable Privacy Policy: Automatic Alignment of Privacy Statements. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, COLING '14, pages 884–894. ACL.

Liu, B., Andersen, M. S., Schaub, F., Almuhimedi, H., Zhang, S., Sadeh, N., Acquisti, A., and Agarwal, Y. (2016). Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. In *Proceedings of the 12th USENIX Conference on Usable Privacy and Security*, SOUPS '16, pages 27–41. USENIX Association.

Massey, A. K., Eisenstein, J., Antón, A. I., and Swire, P. P. (2013). Automated text mining for requirements analysis of policy documents. In *21st IEEE International Requirements Engineering Conference*, RE '13, pages 4–13. IEEE.

McDonald, A. M. and Cranor, L. F. (2008). The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3):543–568.

Meiselwitz, G. (2013). Readability assessment of policies and procedures of social networking sites. In *International Conference on Online Communities and Social Computing*, OCSC '13, pages 67–75. Springer.

Poplavska, E., Norton, T. B., Wilson, S., and Sadeh, N. (2020). From prescription to description: Mapping the GDPR to a privacy policy corpus annotation scheme. In *33rd International Conference on Legal Knowledge and Information Systems*, JURIX '20, pages 243–246. IOS Press BV.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the 3rd Conference on Machine Translation*, WMT '18, pages 186–191. ACL.

Ramanath, R., Schaub, F., Wilson, S., Liu, F., Sadeh, N., and Smith, N. (2014). Identifying relevant text fragments to help crowdsource privacy policy annotations. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2 of *HCOMP '14*, pages 54–55. AAAI Press.

Ravichander, A., Black, A. W., Wilson, S., Norton, T., and Sadeh, N. (2019). Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 4947–4958. ACL.

Ravichander, A., Black, A. W., Norton, T., Wilson, S., and Sadeh, N. (2021). Breaking Down Walls of Text: How Can NLP Benefit Consumer Privacy? In *Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, ACL-IJCNLP '21, pages 4125–4140. ACL.

Reidenberg, J. R., Breaux, T., Cranor, L. F., French, B., Grannis, A., Graves, J. T., Liu, F., McDonald, A., Norton, T. B., Ramanath, R., , Russell, N. C., Sadeh, N., and Schaub, F. (2015). Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Technology Law Journal*, 30:39–88.

Sadeh, N., Acquisti, A., Breaux, T. D., Cranor, L. F., Smith, N. A., Liu, F., Schaub, F., Russell, N. C., Schaub, F., and Wilson, S. (2013). The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. Technical Report CMU-ISR-13-119, Carnegie Mellon University.

Sathyendra, K. M., Ravichander, A., Story, P. G., Black, A. W., and Sadeh, N. (2017). Helping users understand privacy notices with automated query answering functionality: An exploratory study. Technical report, CMU.

Sørensen, J. and Kosta, S. (2019). Before and After GDPR: The Changes in Third Party Presence at Public and Private European Websites. In *Proceedings of the 28th International World Wide Web Confer-*

ence, WWW '19, pages 1590–1600. ACM.

Story, P., Zimmeck, S., Ravichander, A., Smullen, D., Wang, Z., Reidenberg, J., Russell, N. C., and Sadeh, N. (2019). Natural language processing for mobile app privacy compliance. In *Proceedings of the 2019 AAAI Spring Symposium on Privacy Enhancing AI and Language Technologies*, PAL '19. AAAI Press.

Urban, T., Tatang, D., Degeling, M., Holz, T., and Pohlmann, N. (2018). The Unwanted Sharing Economy: An Analysis of Cookie Syncing and User Transparency under GDPR. arXiv preprint 1811.08660.

van Eijk, R., Asghari, H., Winter, P., and Narayanan, A. (2019). The Impact of User Location on Cookie Notices (Inside and Outside of the European Union). In *Proceedings of the 2019 Workshop on Technology and Consumer Protection*, ConPro '19. IEEE.

Vincent, M. and Winterstein, G. (2013). Building and exploiting a French corpus for sentiment analysis. In *Proceedings of TALN 2013*, volume 2: Short Papers of *TALN '13*, pages 764–771. ATALA.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., Norton, T. B., Hovy, E., Reidenberg, J., and Sadeh, N. (2016a). The Creation and Analysis of a Website Privacy Policy Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers of *ACL '16*, pages 1330–1340. ACL.

Wilson, S., Schaub, F., Ramanath, R., Sadeh, N., Liu, F., Smith, N. A., and Liu, F. (2016b). Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International World Wide Web Conference*, WWW '16, pages 133–143. ACM.

Zimmeck, S., Wang, Z., Zou, L., Iyengar, R., Liu, B., Schaub, F., Wilson, S., Sadeh, N. M., Bellovin, S. M., and Reidenberg, J. R. (2017). Automated Analysis of Privacy Requirements for Mobile Apps. In *Proceedings of the 24th Annual Network and Distributed System Security Symposium*, NDSS '17. Internet Society.

Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Russell, N. C., and Sadeh, N. (2019). MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86.

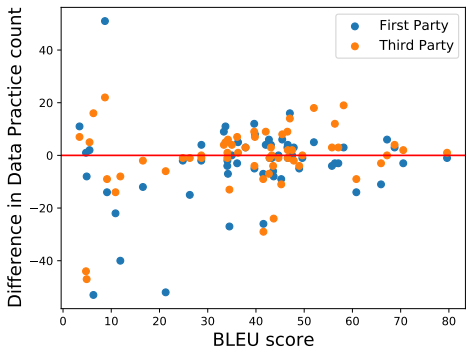

Figure 2: Differences between MAPP-59 pairs of English and German policies in the number of sentences with first-party and third-party disclosures against each pair's BLEU score. Positive values indicate more disclosures in English and negative values indicate more disclosures in German.

10. Appendix

10.1. Reproducibility

We trained all models using a maximum sequence length of 512 tokens and a batch size of 32 for 3 epochs. For segments that contain more than 512 tokens, we used only the first 512 tokens for classification. We trained our classifiers with a batch size of 32, and tuned the learning rate over the following range of values: [5e-6, 1e-5, 2e-5, 3e-5, 5e-5]. For each of these learning rates, we also experimented with another model setting where the final classifier layer is trained with a higher learning rate of 1e-4. To find the best hyperparameters and model settings for training the classifier, we further split the training set into a 80:20 training/-validation split. We ran a grid search over all possible hyperparameters and chose the values with the best F1 score for the positive class on the validation set. We used the best hyperparameters to fine-tune the language model on the entire training set and to evaluate the classifiers' generalization performance on the testing set.

10.2. Annotation Scheme

Table 7: Data practices and attributes in our annotation scheme. (opt) indicates that a data practice attribute is optional.

Categories	Category Description	Attributes	Attribute Description
First-Party Collection/Use	Privacy practices describing data collection or data use by the company/organization owning the website or mobile app.	Information Type	What category of information is collected or tracked by the company/organization?
		Purpose	What is the purpose of collecting or using user information?
		Collection Process	How does the first party collect, track, or obtain user information?
		Does/Does Not (opt)	Use this optional attribute to denote if the policy explicitly states that something is NOT done. Defaults to "Does."
		Collection Mode (opt)	Use this optional attribute to denote if the data collection performed by the first party is implicit (e.g., the company/organization collects the information without the user's explicit awareness) or explicit (e.g., the user provides the information). Defaults to "not selected."
		Anonymization (opt)	Use this optional attribute if it is explicitly stated whether the information or data practice is linked to the user's identity or if it is anonymous. Defaults to "not selected."
		User Type (opt)	Use this optional attribute if a practice applies specifically to users with an account or users without an account. Defaults to "not selected."
		Choice Type (opt)	Use this optional attribute if user choices are explicitly offered for this practice. Defaults to "not selected."
		Choice Scope (opt)	Use this optional attribute to indicate the scope of user choices. In some cases, even if user choices are not clear or specific, this attribute can be selected. Defaults to "not selected."
		Legal Basis for Processing	The GDPR prohibits the collection and processing of personal data without a proper legal basis. Therefore, every category of personal data requires the legal basis to be clear and specific.
Third-Party Collection/Use	Privacy practices describing data sharing with third parties or data collection by third parties. A third party is a company/organization other than the first-party company/organization that owns the website or mobile app.	Information Type	What category of information is shared with, collected by, or otherwise obtained by the third party?
		Purpose	What is the purpose of a third party receiving or collecting user information?
		Entity	The third parties involved in the data practice.
		Collection Process	How does the third party receive, collect, track, or see user information?
		Does/Does Not (opt)	Use this optional attribute to denote if the policy explicitly states that something is NOT done. Defaults to "Does."
		Anonymization (opt)	Use this optional attribute if it is explicitly stated whether the information or data practice is linked to the user's identity or if it is anonymous. Defaults to "not selected."
		User Type (opt)	Use this optional attribute if this practice applies specifically to users with an account or users without an account.
		Choice Type (opt)	Use this optional attribute if user choices are explicitly offered for this practice. Defaults to "not selected."
		Choice Scope (opt)	Use this optional attribute to indicate the scope of user choices. In some cases, even if user choices are not clear or specific, this attribute can be selected. Defaults to "not selected."

Table 8: Value-level inter-annotator agreement.

Attribute	Value	English		German	
		Coverage	Fleiss' Kappa	Coverage	Fleiss' Kappa
Information Type	Financial	0.02	0.63	0.02	0.59
	Contact information	0.05	0.61	0.05	0.65
	Location	0.03	0.65	0.02	0.61
	Demographic data	0.02	0.66	0.02	0.66
	User online activities	0.08	0.51	0.06	0.51
	IP address and device IDs	0.05	0.68	0.03	0.71
	Cookies and tracking elements	0.05	0.61	0.04	0.47
	Computer information	0.03	0.55	0.04	0.63
	Generic personal information	0.13	0.39	0.12	0.37
Purpose	Essential service or feature	0.1	0.5	0.07	0.43
	Advertising or marketing	0.07	0.55	0.06	0.53
	Analytics or research	0.08	0.54	0.07	0.55
	Service operation and security	0.06	0.49	0.05	0.52
	Legal requirement	0.04	0.55	0.03	0.41
Collection Process	Shared by first party with a third party	0.07	0.42	0.05	0.33
	Collected on first-party website/app	0.1	0.38	0.06	0.26
Legal Basis for Processing	Legitimate interests of first or third party	0.02	0.37	0.02	0.47

10.3. MAPP Corpus Creation Workflow

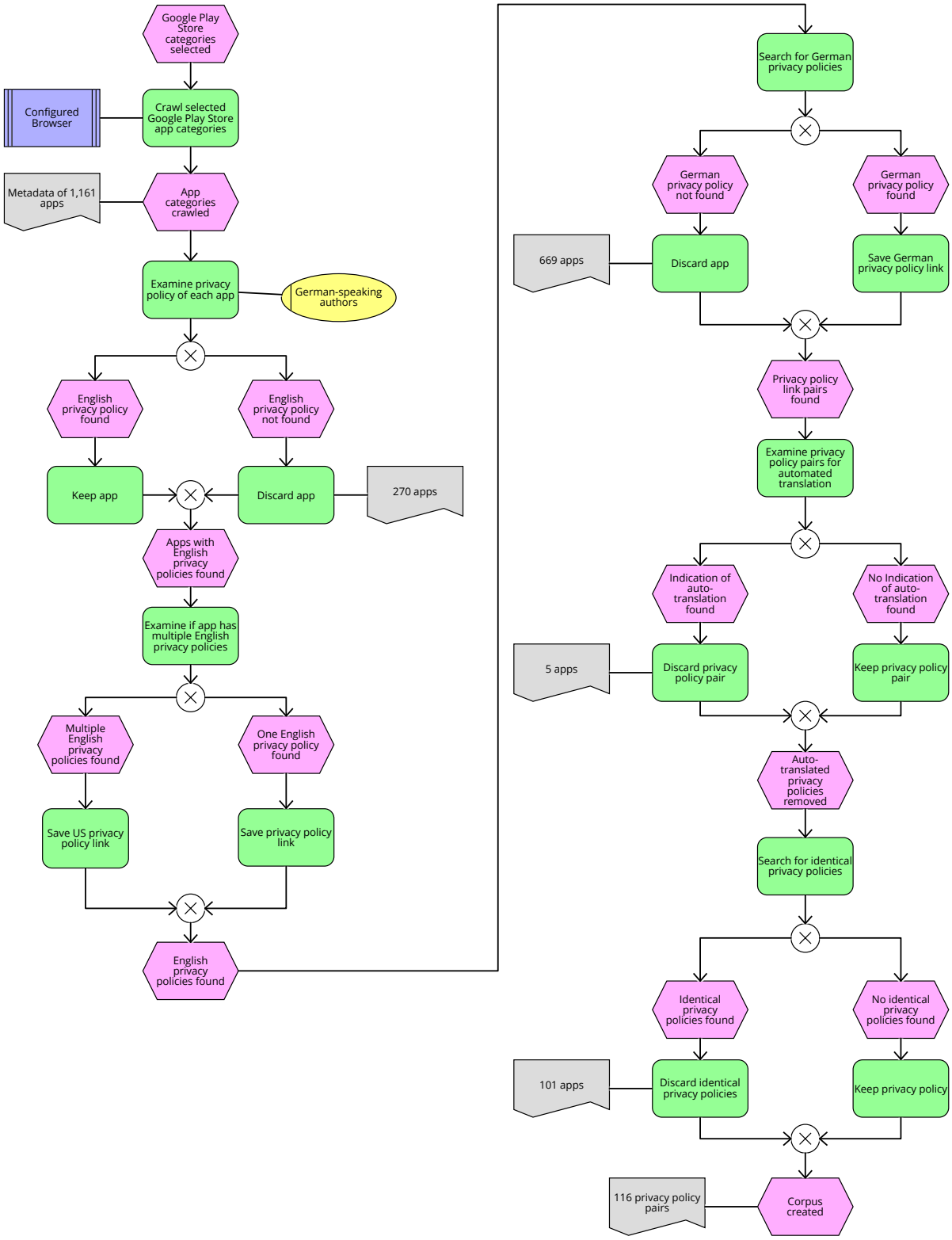

Figure 3: Event-driven process chain describing the workflow to create the MAPP bilingual privacy policy corpus.