Building Cross-language Corpora for Human Understanding of Privacy Policies

Ciclosi, Francesco; Vidor, Silvia; Massacci, Fabio

published in
Digital Sovereignty in Cyber Security
2023

DOI (link to publisher)
10.1007/978-3-031-36096-1_8

document version
Publisher's PDF, also known as Version of record

Building Cross-language Corpora for Human Understanding of Privacy Policies

Francesco Ciclosi[1(✉)], Silvia Vidor[1], and Fabio Massacci[1,2]

[1] University of Trento, Trento, Italy
{francesco.ciclosi,silvia.vidor}@unitn.it, f.massacci@vu.nl
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Abstract. Making sure that users understand privacy policies that impact them is a key challenge for a real GDPR deployment. Research studies are mostly carried in English, but in Europe and elsewhere, users speak a language that is not English. Replicating studies in different languages requires the availability of comparable cross-language privacy policies corpora. This work provides a methodology for building comparable cross-language in a national language and a reference study language. We provide an application example of our methodology comparing English and Italian extending the corpus of one of the first studies about users understanding of technical terms in privacy policies. We also investigate other open issues that can make replication harder.

Keywords: Privacy Policies · Comparable corpora · Methodology · Evaluation · Cross-language corpora

1 Introduction

A well-developed literature exists in relation to the analysis of privacy policies (e.g., [27,28,31]), particularly concerning the development of tools to improve the writing [9] and clarity of the policies themselves [8].

Specific studies on users' understanding of privacy policies are instead less frequent (e.g. [13,19,25]). While the assumption that existing privacy policies are excessively long and complex is at the foundation of the majority of works on the topic [1,20,28,30], the details of users' misunderstanding of policies are less studied (e.g. [24,26]). For example, Tang et al. [24] are the first to focus specifically on misconceptions related to technical terms used in privacy policies.

Yet, the ability of citizens to understand what they accept (e.g., 'are they waving their rights without even understanding it?') is a critical issue to gauge the actual success of privacy legislation initiatives such as the European Union General Data Protection Regulation (GDPR) [18]. The introduction of the GDPR in the EU has increased organizations' focus on complying with data protection principles. This privacy-driven approach progressively has grown the complexity of cybersecurity implementation and consequently their costs [12],

A. Skarmeta et al. (Eds.): CyberSec4Europe 2022, CCIS 1807, pp. 113–131, 2023.
https://doi.org/10.1007/978-3-031-36096-1_8

encouraging the development of GDPR's compliance tools [3], and introducing security workers such as the Data Protection Officer [5] (already existing in the past, but now well-defined and mandatory in all the EU member states). Because an organization must write and comply with privacy policies, it is crucial to assess whether the alleged beneficiaries of the protection granted by these policies (i.e., citizens) can actually understand what they are protected from (and what they are *not* protected from).

A critical issue in this respect is that most studies use English and are based on the U.S. reality (e.g. [21,24,27,29]), which has a different culture as well as differently constructed legislation.

European Union policies are typically written in the national language of a country, and any study that is based on the English language would not accurately reflect the proper understanding of the users: we would be probing at the same time their understanding of English *and* their understanding of the privacy issues. Still, using English as the reference language has advantages, particularly in ensuring that researchers from different countries can access and compare the work.

Given these premises, our main goal is to provide a methodology on how to generate comparable privacy corpora when dealing, on the one hand, with the English language and, on the other hand, with a National language (in this case, Italian).

In this work, we describe how we created such corpora and how to quantify the diversity within the corpora as well as a number of other open issues that makes replication studies way harder than one can initially think of.

2 Related Work

Privacy policies have long been the subject of detailed scientific studies; with the advent of the Internet, online privacy policies have proliferated [2,29] - often in the form of long, complex and, easily misunderstandable statements - in tandem with the necessity to explain users' data treatment for each online service. This, situation has led, to the necessity (that under some legislation, as in the EU, also became an obligation) to identify and develop methods to write privacy policies more easily, but also to make them more accessible to users of all backgrounds. Within the literature on privacy policies, two major issues are at the forefront of the discussion: the construction of adequate corpora for future analyses, and the development of automated tools for the analysis of existing policies and the drafting of future policies on the basis of a continuously evolving legal environment.

Privacy Policies' Corpora Selection. One of the main issues identified by the literature in proceeding with the (automated) analysis of existing privacy policies for their overall improvement is the lack of appropriate datasets from which to start such analysis [1,21]. Specifically, the problem for practitioners lies in the selection of policies that represent adequately the great variety in length, complexity and service coverage present among online privacy policies. The situation is further complicated by the presence of different legal backgrounds concerning

privacy, for example between the European Union (where the content of privacy policies is mainly determined by the GDPR [18]) and China (where the Personal Information Protection Law (PIPL) sets similar requirements [16]) or the United States (where applicable laws vary between different States or circumstances). These differences can result in different policy structures and contents. As a consequence, the privacy policy of the same company can vary significantly depending on the country from which it is read. An additional factor of variety is the dimension of the company which the privacy policy is referred to, an element which generally impacts both the length of the policy and the frequency with which it is updated [27].

The methods used for policy selection vary between different works. Two main approaches emerge from the literature: the identification of criteria for the manual selection of representative policies (e.g. [27]); and the development of web crawlers to extract the highest possible number of policies available online (e.g. [1]). The two approaches seem to reflect the distinct priorities of different studies conducted on privacy policies, which tend to focus either on the characteristics of the selected policies (a preference for "quality") or on the sheer amount of considered policies (a preference for "quantity"). The second approach seems to be the most common in the literature, though a combined method – establishing a set of quality criteria, often starting from a manual selection and analysis of a few policies generally selected on the basis of popularity, and using such criteria as a basis of action for a crawler – is also employed often.

Tools for Analysis. A thriving strand of literature is then dedicated to automated tools aimed at, among others, analysing, creating, synthesizing, verifying the compliance or extracting specific elements of interest from privacy policies. The automated analysis of privacy policies has become a necessity both for the organizations writing the policies and for supervising authorities, but also for users requiring new manners to identify and understand the highlights of such policies [6]. Situated at the intersection between legal and technical domains, such analysis has recently turned to machine learning and text mining in order to automate and potentially improve a process that still relies heavily on human contribution [21].

Automatic privacy policy analysis is generally performed through natural language processing (NLP) techniques, which remain the dominant approach in the area. As reported by Del Alamo et al. [6], NLP techniques can be divided into symbolic (or classic) and statistical (or empirical). The first starts from human-developed rules to process the policy's text and model natural language; the second, instead, applies mathematical techniques to previously-created corpora of policies to develop generalized linguistic models. The attention of practitioners is mostly directed to the statistical approach and to the creation of tools enabling its execution in an automatic manner [30,31].

Readability and Users' Misconceptions. The academic interest over privacy policies is not limited to the manner in which they can be composed. In fact, the efficacy of a privacy policy does not depend only on its adherence to the legal

requirements established by the countries of reference – though that is indeed an important and necessary element – but also on how understandable it is by the final recipient of the policy: the concerned website's user.
The capability of users to understand the content of privacy policies is analyzable in more than one way. The majority of the literature focuses on the measure of so-called "readability" [7,11], which can be calculated according to a series of mathematical formulas or characteristics such as length, language complexity and univocal meaning.

Though definitely less studied throughout the years, a significant portion of scholarly investigation related to privacy policies has concerned users' misconceptions about the policies' meaning and function [19,25].

3 A Reference Corpus in English

To start a reference corpus, we used the study by Tang et al. [24], who tried to investigate the understandability of privacy policies from the users' perspective, focusing in particular on specific technical terms commonly used in data use policies in the context of the USA.

To achieve such an aim, the authors of [24] ran three different studies:

1. A qualitative pilot study to identify commonly misunderstood technical terms;
2. A large-scale main study to test the respondents' understanding of the selected terms and their comfort with some data use practices;
3. A small-scale follow-up study to support the main study's results.

To select the 22 terms to be included in the main study, the authors of [24] created a preliminary list of 57 terms obtained from manual analysis of the Alexa top 10 U.S. websites as of June 2020 (a common practice in studies on privacy policies [19]) and some selected apps from the Android Play Store. The list was then validated through an automated analysis of a 3609 English-language policies corpus to verify the frequency of use of the selected terms. The 57 terms were also divided into 11 categories based on the macro-area they belonged to (e.g., crypto, storage, tracking). From the original study's authors, we have obtained a spreadsheet including the technical terms (58 since one of them was missing from the original appendix) and the categories used in the pilot study. We also received the authors' original notes, where the policies in which they found the technical terms, the contexts of use, and a link to (the current version of) the relevant privacy policies are specified.

In the study by Tang et al. [24], the 57 terms, shown randomly based on their category, were included as part of the pilot study. In the pilot study, respondents from Amazon Mechanical Turk were asked to define the term they were shown. Based on the pilot results, 20 technical terms that were misunderstood the most (of which ten belong to the set of high-frequency terms commonly misdefined, while the others belong to the set of terms that the participants significantly misunderstood) were selected for the main study, together with two

mostly well-understood terms. The main study was divided into two sections, asking respondents to answer multiple-choice questions defining some of the 22 terms and rate their comfort with some data use practices on a five-point Likert scale. The ratio of misunderstood to understood technical terms used for the survey is highly unbalanced (20-to-2), which may affect Tang et al. [24] study's validity. From our perspective, we use them only as control elements to make the study comparable.

Due to the uncertainty connected to the definition of comfort, the follow-up study was then used to verify whether users' attitudes to policies changed using technical versus descriptive terms.

4 Methodology

This section describes a usable methodology to build comparable privacy policies' cross-language corpora by mapping the policies of a corpus taken from a reference study into a new one that adopts the language of a future replication study. Figure 1 represents this methodology's diagram summarizing inputs, outputs, main characteristics considered, and transformations steps involved. Indeed, a cross-national replication study needs to consider a new corpus of privacy policies comparable with the original one because it will probably not be available or existent in a hypothetical parallel corpus. Hence, in this case [10], verifying the comparability of the two corpora is a precondition for using the new one.

The comparability check requires at least three different arguments. The first is the object to be compared, the second is another object to compare the first with, and the third is the respect used to compare these two objects. Hence, the third element is crucial, and its choice influences the comparison's result. Therefore, it is necessary to identify some properties concerning which they compare the corpora in comparing two corpora. There is the need to assess the comparability of a corpus with another that is not available since it will be composed of privacy policies used to replace some that are not available in the language of the future replication study. Therefore, before selecting policies based on the appropriate and specific criteria described below, it will be necessary to consider three data features [10]. These are representativeness, homogeneity, and homoscedasticity. The first feature relates to sharing a property between the two privacy policy corpora. The second feature requires that the data be homogeneous concerning the variables relevant to the purpose of the corpus. The third feature is related to the equal variance of the data across all classes.

Before a researcher can replicate a considered study, it is necessary to ensure that the corpora are comparable. Indeed, constructing multilingual corpora of privacy policies requires considering three different corpora: the *original corpus* in the original study and in the original study's language, a *source corpus* in the original study language and *replication corpus* in the replication language. Hence, it is necessary to make three different comparisons between three distinct corpora:

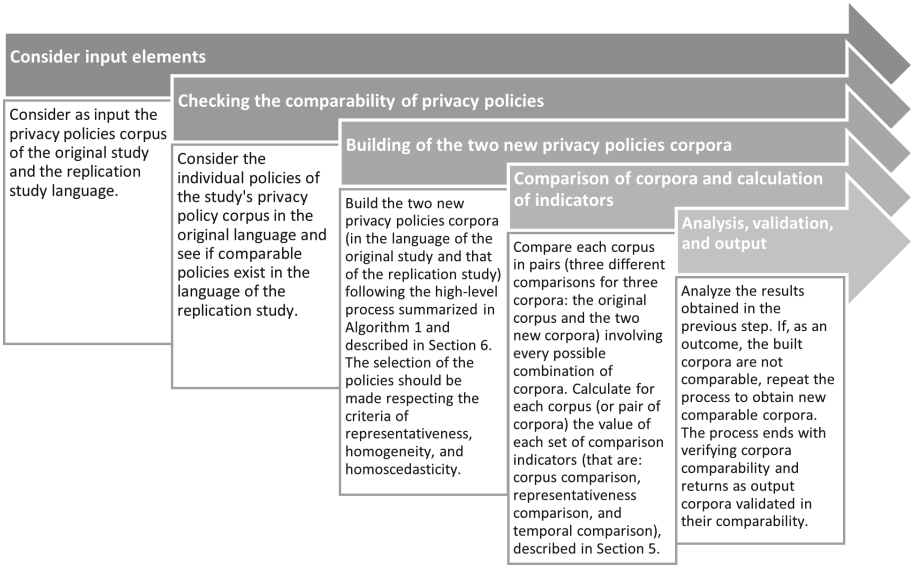

Fig. 1. The methodology to build comparable privacy policies' cross-language corpora

1. between the original study's corpus and the new source corpus in the language of the original study;
2. between the original study's corpus and the new replication corpus in the language of the replication study;
3. between the new source corpus in the original study's language and the new replication corpus in the languages of the replication study.

Identifying a new source corpus in the original language study is necessary because not all policies are usable/present in the replication language, and therefore, there might not be a correspondence between some/several policies in the original language and some policies in the replication language. As we illustrate here, this may happen because a company or institution in the original corpus does not exist in the replication country, or the company does exist but only have a policy in some languages and not the replication language. Only in very particular cases the original corpus and the new source corpus will coincide.

Usually, the availability of comparable corpora is more significant than parallel ones because they have the only requirement that their component documents cover related content in different languages [15]. Many works in the literature have demonstrated the usability of comparable corpora in various areas; for example, in improving CLIR systems [23] or in bilingual vocabulary extraction [4]. [14,22] have shown that an increase in the quality of a comparable corpus offers better performance to applications that refer to it. Moreover, [15] highlighted the need to adopt methods that can qualitatively assess the quality of a corpus. This method allows moving beyond a naive approach based only on reasoning, which can lead to an a priori unpredictable level of performance.

Because of the problems highlighted in comparing different corpora [10], we make use of solutions for measuring the distance between comparable corpora across languages derived from the one described in [17].

In our instance, we built new corpora of privacy policies (in Italian and English) from the privacy policies considered in the study taken as an example. These policies are accessible from the sites of the organizations that make them available through an interlanguage link (i.e., a link that relates documents on the same topic but written in different languages). This reflection allows us to extend to our case the considerations made in [17] regarding retrieving comparable Wikipedia documents. Furthermore, a manual evaluation of the privacy policies retrieved and included in our new corpora (source and replication) showed that, in many cases, the privacy policies in the language of replication were created from a simple translation of the analogous source policies written in English. Indeed, as is evident from Table 1, most companies come from states whose native language is English.

Note that in our case, because the privacy policies are retrieved from the same website (simply choosing the correct language) or from different regional websites of the same company, it is probable that the considered corpora are parallel. However, this is not sure because, due to different legislation (from U.S. and EU), it is not taken for granted that a policy text is the translation of another policy text. Hence, we must assume the only certain fact that, in this situation, the two individually considered corpora are similar in some aspects; that is, they are comparable.

In our case, as suggested in [17], we focused on finding policies with a similar length (difference in word count less than 20%) to increase the probability that they are comparable; that is, they are similar in structure and content. Although most of the policies found have a similar length, there are some exceptions where this condition is invalid. Examples are in the policies of Yahoo, Amazon (IT and U.S., while the state is verified between the IT and NL-EN versions), and the USPS-Poste and Teamsystem-Force pairs (of which there is only one language version of the policies).

A replication gap between the referred example study and the replication study will always exist because some time has passed. This gap has to be manageable and acceptable. To this extent, we identified three groups of indicators to quantify such a gap and which we discuss in more detail in the next section. Then, we analyzed the outcomes provided by these three groups of indicators and qualitatively discussed the real comparability of the corpora.

Finally, having verified the corpora's comparability, we focused on mapping the policies' terms. This activity was made manually and revealed some criticalities described in Sect. 8 which is the ground for future work.

5 Comparison Indicators

Based on the discussion in [10,17], and [15], we propose to use three groups of indicators: *corpus comparison*, *representativeness comparison*, and *temporal comparison*, to qualitatively measure the privacy policies' corpora comparability.

The first indicators' group (*corpus comparison*) measures whether the policies of the original study corpus are comparable with the corresponding (if they exist) policies in the cross-national study language. To compare corpora, we are mainly interested in quantifying how many policies are comparable, replacement (or complementary), and destructured. The indicator *Number of Comparable Policies* measures the number of privacy policies referred to the same company for which there is both a text in the original and replication language. These texts must have a similar (comparable) structure.

The *Number of Replacement Policies* indicator measures the number of policies in which the company from which the original study's policy is extracted does not exist in the replication country language. In contrast, the *Number of Complementary Policies* indicator measures the number of policies for which does not exist a localized version of the policy the original study considered. In the first case, if a product or service of a comparable company exists from the same industrial sector, the policy can be replaced with a new policy of these comparable companies. In the second case, if a localized privacy policy text of a *comparable* company's product or services exists, this policy will be used.

A privacy policy is considered complementary when it is added to a corpus to reflect any updates to a website rankings list (e.g., Alexa TOP10 U.S. or IT) that have occurred since the date of the original study.

The *Number of Destructured Policies* indicator estimates how many policies where a policy in the target language exists; still, this text is not immediately comparable with the text in the original study language due to a significantly different structure. The more destructured policies, the harder any replication study is. To ensure better replicability across languages of a study, the number of comparable policies must be as high as possible.

The second set of indicators measures the representativeness of the corpora. It provides qualitative criteria to follow in determining how to replace policies in the original corpus that are not usable in the replication study corpora (source corpus and replication corpus).

To measure the corpora's representativeness, we propose to do first a qualitative comparison using a table to compare the rank sources for each corpus. Indeed, it is crucial that the process used to build the new corpora be the same. Analogously, the rank sources used by the different corpora must be distinct from each other (because they are based on the constituent policies' language) and homogeneous (they must use the same source as, for example, Alexa Top10). One possible replacement policy criterion is to investigate if a TOP web ranking used in a study carried out in a particular language also exists in a different language, such that both policies talk about the same terms.

For example, it is possible to consider the Alexa TOP10 website list in both languages. Another example is considering the ranking of top companies in a specific industry (e.g., the top banks) for each state to which the language used in the studies refers.

Finally, to consider the replicability of the original study, it needs to assess the temporal comparability of the corpora (third indicators' group). We evalu-

ated qualitatively three indicators to measure this comparability: the *Temporal Internal Consistency*, the *Temporal Replication Gap*, and the *Qualitative Replication Gap*.

We used the *Temporal Internal Consistency* indicator to ensure that the source and replication corpora policies are all in the same narrow interval. This indicator measures within the same body of privacy policies the number of months between the publication date of the most recently updated policy and the one of the least recently updated policy. The *Temporal Replication Gap* indicator aims to highlight if the policy has changed. It measures the average number of months of the update time of policies in the source and replication corpora vs. the ones in the original corpus. This value is calculated only for the comparable or destructured policies because it has no sense for the other ones (complementary or replacement). Indeed, these policies are added contextually to the new source and replication corpora and are absent from the original one. Since the study may give different results because time has passed and many things have happened (e.g., changes in laws), the *Qualitative Replication Gap* indicator is used to mark if major events happened from the original study and its replication.

6 Reconstruction of Policy Corpus

In the initial phase, we focused on looking for the correspondence between the Italian and English policies of the new cross-national source and replication corpora and the English policies of the original corpus we took from the example study of Tang. et al. [24]. Hence, excluding the case in which the policies are comparable, we identified three cases that require attention.

1. Missing policy: there is no Italian policy; however, there is a corresponding Italian app or website.
2. Missing company: there is no corresponding Italian app or website.
3. Destructured policy: there is a corresponding Italian version of the original privacy policy, but we found that this has an entirely different structure.

Henceforth, a team of two researchers analyzed the original corpus of policies and found correspondence rules to build comparable replication and source corpora of Italian and English ones. After defining our policy replication corpus for the Italian study, we concentrated on analyzing whether its policies were structurally similar to those in English of the original corpus. The high-level process to build the new replication and source corpora is presented in Algorithm 1. All the operations listed in the algorithm are done manually by humans. When in Algorithm 1, the condition *ItalianCompanyMissing* is verified, there is a case of a replacement policy. Analogously, if the condition *ItalianPolicyMissing* is satisfied, there is no Italian policy, but there is a corresponding Italian app or website. Hence, this is a case of a complementary policy. Instead, if the condition *ItalianPolicyDestructured* is verified, an Italian policy exists for a specific company. However, it has an entirely different structure from the original English policy (case of destructured policy).

If we need to substitute the English company with another of the same type (for example, another bank will replace a bank), we choose one that operates globally. Analogously, in case we need to substitute the missing privacy policy, we look for another related to a comparable company for the business sector, market segment, and type of product or service. With these choices, we have favored companies operating in the USA and Italy.

Algorithm 1. Look for the corresponding Italian privacy policy entry

```
1: procedure ITALIAN POLICY CHECK(EnglishPolicyURL)
2:     Retrieve English privacy policy text from EnglishPolicyURL
3:     Read English privacy policy text
4:     Retrieve English company from EnglishPolicyURL
5:     Read English company
6:     Search for the correspondent Italian company
7:     if ItalianCompanyMissing = True then
8:         Choose a new company with the same type as the original company and a privacy policy in Italian and English
9:         Retrieve URL and text of Italian policy of new company
10:        Retrieve URL and text of English policy of new company
11:        Add new company's English policy to new English corpus as replacement
12:        Add new company's Italian policy to Italian corpus as replacement
13:    else
14:        Search for the corresponding Italian privacy policy of the original Company
15:        if ItalianPolicyMissing = True then
16:            Choose another privacy policy in Italian related to a company comparable to the original company
17:            Retrieve the URL and the text of the Italian policy of the new comparable company
18:            Add the original company's English policy to the new English corpus
19:            Add the new comparable company's Italian policy to the Italian corpus
20:        else
21:            Retrieve the URL and the text of the original company's Italian policy
22:            Add the original company's English policy to the new English corpus
23:            Add the original company's Italian policy to the Italian corpus
24:            Analyze the Italian privacy policy structure
25:            Compare the Italian policy structure with the English one
26:            if ItalianPolicyDestructured = True then
27:                Operate a manual measures activity of the privacy policy
28:            else
29:                Classify the Italian privacy policy as comparable
```

We performed a manual analysis of the Alexa top 10 Italy websites as of November 2021, and analogously, we analyzed selected apps that, in the same period, had ranked better in the "most profitable games" category of the Play Store for Italy. After that, we compared these lists with the analogous ones (that refer to June 2020) in the original study for the U.S. privacy policies. To overcome the 17-month time gap between the establishment of the original corpus and the replication corpus, we made a further comparison by considering the content of the Alexa Top 10 U.S. ranking in the language of the original study as of November 2021.

From the original study, we considered 58 URLs obtained from 17 different companies' privacy policies. After this review phase, we removed six companies' privacy policies; however, we added twelve. This addition allowed us to improve the size of our corpus of policies, especially its specialization in the Italian reality

and its focus on EU GDPR [18] concepts. We summarized these activities in Table 1. Our process in detail was as follows

a) the privacy policies of the apps *Infinite Word Search* and *Woody Puzzle* have been replaced by those of the apps *Coin Master* and *Empires & Puzzles* that, in addition to having an Italian privacy policy, respectively (on November 23, 2021) ranked first and fourth in the "most profitable games" category of the Play Store for Italy;
b) the privacy policy of the app *Signal* is replaced respectively by the policy and by the FAQ documentation of the apps *Whatsapp* and *Telegram*. The *Signal* app has an Italian version but not an Italian privacy policy, so we have chosen to replace it with other messaging apps with similar characteristics, such as support for end-to-end encryption. Moreover, because for the app *Telegram* there is not an Italian privacy policy but only an Italian FAQ documentation section in which the considered technical terms are present, we referred to this documentation;
c) the privacy policy of the website *Bank of America* is replaced by the *Unicredit* one, that is a company of the same category and operating worldwide;
d) the privacy policies of the *Reddit* and *Verizon* websites are replaced by the *Vodafone* one, that is a company that provides analogous services;
e) starting from one of the top 10 Alexa U.S. websites on November 2021, we added both the privacy policy of the related company and the one connected to a comparable Italian company (in particular the *Force* U.S. company and the *Teamsystem* Italian one);
f) starting from one of the top 10 Alexa Italian websites on November 2021, we added both the privacy policy of the related company and the one connected to a U.S. comparable company (in particular the *Poste* Italian company and the *USPS* U.S. one);
g) we added the privacy policies of the companies *Zoom* and *Microsoft* which have a website that appears both in Alexa top 10 Italian and in Alexa top 10 U.S. We chose policies from websites that appeared in the Alexa top 10 rankings for IT and U.S. to maintain equivalence between the original corpus, the new source corpus of English policies, and the new replication corpus of Italian policies. The original study's authors built their English technical terms corpus by manually analyzing the Alexa top 10 U.S. websites.

7 Results and Comparison

In our example, we made three different comparisons between three different corpora to determine their comparability level. The first corpus is the English-language corpus, the original used in the study by Tang et al. [24]. The other two are cross-language corpora built (one, the source corpus, in English, and the other, the replication corpus, in Italian, which is the language of a potential replication study) from the first corpus. We aim to use them for a potential replication study on how humans understand privacy policies.

Table 1. Composition of privacy policies corpus (This table describes the compositions of the original privacy policies corpus and the new English and Italian source and replication corpora. Columns named *Orig. Tang et al.*, *New US*, and *New IT* refer to different privacy policies corpora. The first entry refers to the English-language privacy policies corpus analyzed in the original Tang. et al. [24]study, while the second and third entries refer to the new privacy policies source and replication corpora in English or Italian, respectively. The *Referred company* field shows the company (and eventually their product) whose privacy policy we considered. Finally, the *Notes* field summarizes the rationale behind our choices in selecting the privacy policies.)

Referred company	Orig. Tang et al.	New US	New IT	Notes
Yahoo	√	√	√	We maintained it from initial study.
King Games (app Candy Crush)	√	√	√	
LinkedIn	√	√	√	
Amazon	√	√	√	
Google	√	√	√	
Apple	√	√	√	
Facebook	√	√	√	
Wikipedia	√	√	√	
Twitter	√	√	√	
Ebay	√	√	√	
Firefox (documentation)	√	√	√	
Random Logic Games (app Infinite Word Search)	√			We substituted policies because they do not exist in their Italian version.
Athena FZE (app Woody Puzzle)	√			
Moon Active (app Coin Master)		√	√	Companies added from the PlayStore's rank on Nov. 2021.
Zinga (app Empires & Puzzles)		√	√	
Signal	√			We substituted the policy because it does not exist in its Italian version
Telegram (FAQ doc.)		√	√	Companies added to replace the *Signal* one.
WhatsApp		√	√	
Bank of America	√			We replaced the policy because the respective Italian companies do not exist
Unicredit		√	√	Company added to replace the *Bank of America* one
Reddit	√			We substituted companies with new ones focused on Italian reality.
Verizon (cookie policy)	√			
Vodafone		√	√	Company added to replace the *Reddit* and *Verizon* ones
Zoom		√	√	Policies added from the top 10 Alexa Italian and U.S. on Nov. 2021.
Microsoft		√	√	
USPS (only US)		√	√	It was added because it is a U.S. company with similar features and products to the Italian *Poste*
Poste (only IT)		√	√	Company added from the top 10 Alexa Italian on Nov. 2021
Teamsystem (only IT)		√	√	It was added because it is an Italian company with similar features and products to the U.S. *Force*
Force (only US)		√	√	Company added from the top 10 Alexa U.S. on Nov. 2021

Comparing the English-language original corpus from Tang et al. [24] with the source corpus built by us in the same language, we have 52,38% comparable

policies, 38,10% complementary policies, 9,52% replacement policies, and no destructured policies. The results are worst when comparing the original English-language corpus and the new replication one in Italian. In that case, we have 47,62% comparable policies, 38,10% complementary policies, 9,52% replacement policies, and 4,76% destructured policies.

In contrast, comparing the two new source and replication corpora in Italian and English, we find these are well aligned. We have 85,72% comparable policies, 9,52% complementary policies, 4,76% destructured policies, and no replacement policies. In all comparisons, the number of destructured policies is low, if not wholly absent.

Hence, rather than using Tang et al.'s [24] privacy policies' source corpus for a cross-language comparison, we compare the two new source and replication corpora (built from the original one) that are more closely aligned.

Moreover, internal temporal consistency exists in the new source and replication corpora policies. Indeed, these policies are all in the same narrow interval because their release varied from June 2020 to September 2021. Furthermore, with few exceptions, the Italian and English versions of our policies were published simultaneously. The exceptions to this are Amazon (where the Italian version is December 2020 while the English version is February 2021), Facebook (with versions varying from August 2020 to January 2021), WhatsApp (varying from January 2021 to November 2021), or Vodafone (varying from July 2021 to November 2021).

For a correct calculation of the value of the "*temporal replication gap*" indicator, it is essential to know the version and date of each privacy policy considered (both in the original corpus in the original study and those in the source and replication corpora in the replication study). Knowing the number of intermediate versions of privacy policies for each company is also helpful. In our example, there are many problems, and this indicator is not helpful. The original study [24] did not specify the version and date of policy updates. We deduced the date of the considered policies by cross-referencing the updating date of the policy with the date of consultation of the TOP 10 of Alexa U.S. (June 2020). We know that this date is incorrect, but it still gives us an indicative idea. Moreover, considering that not all companies expose the historical archive of policy versions, we can obtain a policy date to calculate the indicator in only 7 out of 11 cases. Other companies have made some intermediate versions of the privacy policy disappear from their website (for example, on the Facebook website, the previous versions are no longer available between September 9, 2016, and January 4, 2022). This fact generates another interesting problem because users have given their consent based on a no longer existing policy. We have not considered the case of Facebook because it presents an outlier value. Therefore, we used 6 out of 11 policies in calculating the indicator. Interestingly, with only one exception (Google, for which there were four different versions in the face of a 17-month time gap), between the two studies, there was at most one version gap in the privacy policies considered. We calculated the temporal replication gap value only for the comparable or destructured policies because it has no sense for the other

ones (complementary or replacement). Indeed, these policies are added contextually to the new source and replication corpora and are absent from the original one. The final value obtained for this indicator shows that the average number of months of the update time of the policy (comparable or destructured) is 16,5. This result is good because it shows that from the policies version available in the example study's original corpus and the new ones (source and replication corpora) built by us, there is only one version difference and a time lag between updates of just over a year.

Considering the qualitative replication gap indicator, it shows that in the time between the construction of the original corpus and of the pair source and replication corpora (June 2020 - November 2021), there have been no significant regulatory developments. For example, both the *GDPR* and the *California Consumer Privacy Act (CCPA)* predate the construction of the original corpus, while other laws, such as the *Connecticut Data Privacy Act (CDPA)*, the *Utah Consumer Privacy Act (UCPA)*, and the *Virginia Consumer Data Protection Act (VCDPA)* was not yet in effect. Despite this, there have been some changes at the corporate level (e.g., the corporate change from *Facebook* to *Meta*) or at the policy level (e.g., additional information added regarding data retention in the Yahoo policy).

8 Open Problems: Technical Terms May Not Match

After building a privacy corpus in a different language, the next step is replicating different studies. For example, the authors in [24] arranged a survey based on 22 key terms derived from the pilot study's results to investigate the understanding of technical terms.

We experienced many cases of technical terms that do not match and occur with extremely different frequencies.

The most egregious case are the English terms "*personal information*" and "*personally identifiable information (PII)*" (that were distinct in the considered study). Hence, we used a syntactic criterion both to be sure of their overlapping and overcoming it in a single Italian term, "*dati personali*." In practice, we used a syntactic rule to identify the possible syntactic distinction of these terms. Hence, they are distinguishable if there is no correspondence between the two policies (English and Italian) or if there are orphans or widows. Otherwise, it is possible to conclude that they are the same thing and are codifiable with the same term in Italian.

To understand if the terms PII@US (*personally identifiable information*) and PI@IT (*informazioni personali*, *personal information*) are equivalent, we manually looked for widows' or orphans' presence related to these technical terms. An *orphan* is a term that appeared in the Italian policy without a match in the corresponding clauses or sections of the English policy. A *widow* is any occurrences in the English policy in which the term PII@US (or the term SPI - *sensitive personal information* -) appears without the term PI@IT in the Italian policy.

We also looked for implicit terms or pronouns that refer to technical terms. We found very few occurrences of those cases.

Other critical issues occur when technical terms do not appear or are rare in our corpora. For example, the term *fingerprinting* does not appear in the Italian corpus and is useless for an Italian language survey. Instead, *public information* and *browser web storage* terms are rare in our corpora, but for an Italian survey, they are significant and can be used.

We illustrate these issues by considering the 22 terms from [24]:

- two high-frequency terms that most users in the pilot study correctly defined;
- ten high-frequency terms that the study's participants commonly misdefined;
- ten additional terms for which the study's participants exhibited significant misunderstandings.

Using our new source and replication corpora of Italian and English policies, we investigate the differences between the frequency of occurrences in privacy policies of the most common Italian and English technical terms (the TOP 10 and the TOP 22 lists). More details are listed in Table 2.

The correspondence between the TOP 10 English list and the TOP 10 Italian one is generally reasonable. 80% of the terms included in the TOP 10 English list also belong to the TOP 22 Italian one (with a peak of 70% of terms also included in the Italian TOP 10). Hence, only 20% of terms in the TOP 10 English list do not belong to the TOP 22 Italian one. The correspondence between the TOP 22 English list and the TOP 22 Italian list is acceptable. 58,33% of terms in the TOP22 English list belong to the Italian one, and 16,66% of them also belong to the TOP10 Italian list. Hence, only 41,67% of terms in the TOP22 English list are outside the TOP22 Italian one.

Out of 22 terms used in the Tang et al. [24] example study's privacy policies English original corpus, ten of these (45,46% of the total) appear in one of the TOP English lists of our new English source corpus (specifically, five in the TOP 10 and the other five in the TOP 22). While the remaining 12 terms (54,54% of the total) are outside our source corpus TOP 22 English list. This outcome means that referring to our source corpus of privacy policies, more than half of the technical terms considered in the study taken as an example are rare. This result is crucial because our source and replication corpora are more focused on EU reality.

The situation is worse, considering the terms used in Italian policies. Only 7 (31,82% of the total) of these terms appear in one of the TOP Italian lists (in particular, five in the TOP 10 and the other two in the TOP 22). The remaining 15 terms (68,18% of the total) are outside the TOP 22 Italian list.

Hence, the technical terms used in the example study's survey by Tang et al. [24] are rarely used in the privacy policies in Italian. This consideration is a crucial point that shows the need first to map technical terms and, secondly, to adapt the survey for replicating it to Italian-speaking respondents. These activities will be the subject of future research work.

Table 2. TOP 10 and TOP 22 lists of more frequent terms in the new Italian source corpus of privacy policies (This table describes the TOP10 and TOP22 lists of terms that appear most frequently in the new Italian source corpus of privacy policies. We used boldface to represent the terms belonging to the TOP 10 list. The order (*Rank IT* column) of technical terms in the table follows the terms that appear most frequently in Italian language privacy policies. The table also shows the ranking of the most frequent technical terms in English language privacy policies (*Rank U.S.* column). Corresponding to the Italian-language term (*Term IT*) is the corresponding English-language term (*Term US*). Finally, the table shows the cardinality with which we detected the term in the new replication and source corpora of Italian-language (*#Freq IT*) and English-language (*#Freq US*) privacy policies, respectively.)

Rank IT	Rank US	Term IT	Term US	#Freq IT	#Freq US
1	1	**Dati Personali**	**Personal information**	1523	1430
2	12	of which	**Personally identifiable inform.**	1012	79
3	2	**Cookie/Marcatori**	**Cookies**	645	784
4	4	**Terzi/Terze parti**	**Third parties**	618	360
5	3	**Informativa sul trattamento dei dati personali**	**Privacy policy**	471	540
6	14	**Rettificare**	**Correct**	170	60
7	7	**Indirizzo IP**	**IP address**	136	153
8	26	**Disattivare**	**Deactivate**	128	25
9	9	**Cifratura/Crittografia**	**Encryption**	86	113
10	5	**Dati sull'account**	**Account information**	60	324
11	56	Hash crittografico	Cryptographically hashed	52	2
12	36	Transferimento di dati	Data transfer	50	15
13	10	Affiliate	Affiliates	48	87
14	6	Pubblicità mirata	Targeted ads	47	157
15	15	Identificatori univoci	Unique identifiers	45	54
16	20	API/SDK	API/SDK	44	44
17	16	Dati relativi all'ubicazione	Location-related information	42	50
18	18	Informazioni sui pagamenti	Payment information	41	47
19	40	Resi anonimi	Anonymize	40	13
20	46	Operazioni del dispositivo	Device operations	39	6
21	29	Categorie particolari di dati personali/Dati sensibili	Sensitive personal information	34	19
22	19	Dati di utilizzo	Usage data	28	45

9 Conclusion and Future Works

The first conclusion of our work is that a strict replication study in a different language is possible only up to a certain extent because some elements do not transfer from the original study to the new one.

Therefore, our contribution is to discover and describe how elements (for example, the original corpus used in the survey) can be adapted from an original study to a new one to understand what is needed to maintain relative comparability of the components of the study, such as the privacy policy corpora.

Still, the results obtained in our case study confirm that by taking due care, it is possible to build cross-language source and replication corpora of privacy policies usable for research purposes, such as investigating how humans understand their content.

The outcomes of this work will be the base for further research activities. Future interesting work is to investigate the validity of automated analysis approaches, i.e. whether just looking for term occurrences in a policy would yield a correct analysis. Precision and recall values achievable in the automated

coding of technical terms should be compared from those derived by manual coding. A second aspect is mapping technical terms in different language corpora privacy policies. Our preliminary analysis has already shown that this is not so obvious. Future work could define some indicators to identify the irrelevant, infrequent, overlap, and unreliable technical terms that must be removed or replaced according to the case. This information will help, for example, to replicate the understandability questions from Tang et al. [24] to Italian-speaking people who interact with Italian privacy policies.

Acknowledgement. The authors would like to thank Eleanor Birrell and Ada Lerner for providing us their raw privacy corpus used in their paper [24]. Without their time and expertise this paper would not have been possible. This work was supported in part by the EU under the H2020 Leadership in Enabling and Industrial Technologies program under grant agreement 830929 (CyberSec4Europe).

Data Availability. The datasets generated and analyzed during the study are archived on Zenodo at https://doi.org/10.5281/zenodo.7729546.

CRediT statements. *Conceptualization:* FC, FM, SV; *Methodology:* FC, SV; *Software:* SV, FC; *Validation:* FC, SV, FM; *Investigation:* FC, SV; *Data Curation:* FC, SV; *Writing - Original Draft:* FC, SV; *Writing - Review & Editing:* FC, SV, FM; *Visualization:* FC; *Supervision:* FM; *Project administration:* FM, SV; *Funding acquisition:* FM.

References

1. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., Mayer, J.: Privacy policies over time: curation and analysis of a million-document dataset. In: WWW 2021: Proceedings of the Web Conference 2021, pp. 2165–2176 (2021)
2. Cecere, G., Le Guel, F., Soulié, N.: Perceived internet privacy concerns on social networks in Europe. Technol. Forecast. Soc. Change **96**, 277–287 (2015)
3. Chatzipoulidis, A., Tsiakis, T., Kargidis, T.: A readiness assessment tool for GDPR compliance certification. Comput. Fraud Secur. **2019**(8), 14–19 (2019)
4. Chebel, M., Latiri, C., Gaussier, E.: Bilingual lexicon extraction from comparable corpora based on closed concepts mining. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10234, pp. 586–598. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57454-7_46
5. Ciclosi, F., Massacci, F.: The data protection officer: a ubiquitous role that no one really knows. IEEE Secur. Privacy **21**(1), 66–77 (2023)
6. Del Alamo, J.M., Guaman, D.S., García, B., Diez, A.: A systematic mapping study on automated analysis of privacy policies. Computing **104**, 2053–2076 (2022)
7. Ermakova, T., Fabian, B., Babina, E.: Readability of privacy policies of healthcare websites. In: Wirtschaftsinformatik Proceedings 2015 (WI 2015) (2015)
8. Fabian, B., Ermakova, T., Lents, T.: Large-scale readability analysis of privacy policies. WI 2017: Proceedings of the International Conference on Web Intelligence, pp. 18–25 (2017)

9. Hosseini, M.B., Breaux, T.D., Slavin, R., Niu, J., Wang, X.: Analyzing privacy policies through syntax-driven semantic analysis of information types. Inf. Softw. Technol. **138**, 106608 (2021)
10. Köhler, R.: Statistical comparability: methodological caveats. In: Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. (eds.) Building and Using Comparable Corpora, pp. 77–91. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-20128-8_4
11. Krumay, B., Klar, J.: Readability of privacy policies. In: Singhal, A., Vaidya, J. (eds.) DBSec 2020. LNCS, vol. 12122, pp. 388–399. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49669-2_22
12. Layton, R., Elaluf-Calderwood, S.: A social economic analysis of the impact of GDPR on security and privacy practices. In 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), pp. 1–6. IEEE (2019)
13. Leicht, J., Heisel, M.: A Survey on privacy policy languages: expressiveness concerning data protection regulations. In: 2019 12th CMI Conference on Cybersecurity and Privacy (CMI), pp. 1–6 (2019)
14. Li, B., Gaussier, E.: Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 644–652 (2010)
15. Li, B., Gaussier, E.: Exploiting comparable corpora for lexicon extraction: measuring and improving corpus quality. In: Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. (eds.) Building and Using Comparable Corpora, pp. 131–149. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-20128-8_7
16. National People's Congress of the People's Republic of China. Personal Information Protection Law of the People's Republic of China (2021)
17. Paramita, M.L., Guthrie, D., Kanoulas, E., Gaizauskas, R., Clough, P., Sanderson, M.: Methods for collection and evaluation of comparable documents. In: Sharoff, S., Rapp, R., Zweigenbaum, P., Fung, P. (eds.) Building and Using Comparable Corpora, pp. 93–112. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-20128-8_5
18. Parliament EU and Council EU. Consolidated text: 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) (2016)
19. Reidenberg, J.R., et al.: Disagreeable privacy policies: mismatches between meaning and users' understanding. Berkeley Technol. Law J. **30**(1), 39–68 (2015)
20. Robillard, J.M., et al.: Availability, readability, and content of privacy policies and terms of agreements of mental health apps. Internet Intervent. **17**, 100243 (2019)
21. Sarne, D., Chler, J., Singer, A., Sela, A., Bar Siman Tov, I.: Unsupervised topic extraction from privacy policies. In: WWW 2019: Companion Proceedings of The 2019 World Wide Web Conference, pp. 563–568 (2019)
22. Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D., Gornostay, T.: Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In: The 5th Workshop on Building and Using Comparable Corpora, p. 17. CiteSeer (2012)
23. Talvensaari, T., Laurikkala, J., Järvelin, K., Juhola, M., Keskustalo, H.: Creating and exploiting a comparable corpus in cross-language information retrieval. ACM Trans. Inf. Syst. (TOIS), **25**(1), 4-es (2007)
24. Tang, J., Shoemaker, H., Lerner, A., Birrell, E.: Defining privacy: how users interpret technical terms in privacy policies. Proceedi. Priv. Enhan. Technol. **3**, 70–94 (2021)

25. Turow, J., Hennessy, M., Draper, N.: Persistent misperceptions: Americans' misplaced confidence in privacy policies, 2003–2015. J. Broadcast. Electron. Media **62**(3), 461–478 (2018)
26. Vail, M.W., Earp, J.B., Antón, A.I.: An empirical study of consumer perceptions and comprehension of web site privacy policies. IEEE Trans. Eng. Manage. **55**(3), 442–454 (2008)
27. Wilson, S., et al.: The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 1330–1340 (2016)
28. Zaeem, R.N., et al.: PrivacyCheck v2: a tool that recaps privacy policies for you. CIKM 2020. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 3441–3444 (2020)
29. Zeadally, S., Winkler, S.: Privacy policy analysis of popular web platforms. IEEE Technol. Soc. Mag. **35**(2), 75–85 (2016)
30. Zimmeck, S., Bellovin, S.M.: Privee: an architecture for automatically analyzing web privacy policies. In: 23rd USENIX Security Symposium (USENIX Security 14), pp. 1–16, San Diego, CA. USENIX Association (2014)
31. Zimmeck, S., et al.: MAPS: scaling privacy compliance analysis to a million apps. Proceed. Priv. Enhan. Technol. **2019**(3), 66–86 (2019)