Requirements Engineering (2019) 24:291–313
https://doi.org/10.1007/s00766-019-00315-y

RE 2018

Identifying incompleteness in privacy policy goals using semantic frames

Jaspreet Bhatia[1] · Morgan C. Evans[1] · Travis D. Breaux[1]

Received: 16 November 2018 / Accepted: 31 May 2019 / Published online: 11 June 2019

Abstract

Companies that collect personal information online often maintain privacy policies that are required to accurately reflect their data practices and privacy goals. To be comprehensive and flexible for future practices, policies contain ambiguity that summarizes practices over multiple types of products and business contexts. Ambiguity in data practice descriptions undermines policies as an effective way to communicate system design choices to users and as a reliable regulatory mechanism. In this paper, we report an investigation to identify incompleteness by representing data practice descriptions as semantic frames. The approach is a grounded analysis to discover which semantic roles corresponding to a data action are needed to construct complete data practice descriptions. Our results include 698 data action instances obtained from 949 manually annotated statements across 15 privacy policies and three domains: health, news and shopping. Therein, we identified 2316 instances of 17 types of semantic roles and found that the distribution of semantic roles across the three domains was similar. Incomplete data practice descriptions undermine user comprehension and can affect the user's perceived privacy risk, which we measure using factorial vignette surveys. We observed that user risk perception decreases when two roles are present in a statement: the condition under which a data action is performed, and the purpose for which the user's information is used.

Keywords Semantic frames · Semantic roles · Privacy risk · Natural language processing · Privacy

1 Introduction

Companies describe their data practices in privacy policies to inform users about how their data will be collected, used and transferred for the purposes embodied by the website or software. US regulators may check these policies for compliance with actual data practices, when a data breach or data misuse arises. Consequently, the statements in policies represent policy and legal requirements for software systems. Unlike requirements engineering for a single system with a narrowly defined system boundary, policy requirements govern a broader collection of systems, some of which are built and maintained independently of one another. For example, a privacy policy can govern multiple types of products, across both physical and virtual stores. The statements in these policies describe data practices, including how personal data collected from users can and will flow into and out of the company's systems. Where statements prohibit a kind of data flow, or require consent, all systems governed by the policy must ensure compliance. In addition, these policies are drafted to account for current practices, as well as to afford flexibility for future practices that the company envisions. In order to ensure that privacy policies are comprehensive and flexible, companies resort to using ambiguity in the data practice descriptions of their policies. This ambiguity includes vague language, such as general or abstract terms and modal verbs such as "may" and "possibly," while still trying to preserve a clear definition of permitted and prohibited data flows. By improving our understanding of how policy requirements are written, we believe we can improve the alignment between policy requirements and requirements for specific systems.

A requirement is incomplete when it does not answer one or more questions a stakeholder might have regarding the requirement. Incompleteness in requirements engineering

✉ Jaspreet Bhatia
jbhatia@cs.cmu.edu

Morgan C. Evans
morganev@cs.cmu.edu

Travis D. Breaux
breaux@cs.cmu.edu

[1] Institute for Software Research, Carnegie Mellon University, Pittsburgh, PA, USA

is known to create misunderstanding among stakeholders who have different interpretations of incomplete information [14]. Incomplete requirements are one of the most critical challenges faced by software companies and are also a frequent cause of project failures [17]. Incompleteness, which is a form of ambiguity, occurs in data practice descriptions when one or more policy statements do not answer all the questions that users or regulators may have regarding the company's data practices. For example, with respect to the data action "share," one could ask: what type of data is shared? With whom will the data be shared? From whom was the data collected? For what purpose is the data shared? Finally, under what conditions will the data be shared? If the data practice description does not answer one or more of these questions, the description can be considered incomplete with respect to the missing information.

Incompleteness in privacy policy requirements can also lead to systemic privacy risk to users whose data is collected and shared. For example, in the summary privacy statement: "We may share your location information," the purpose for which the user's location information is shared is missing, which requires the user to make assumptions about the missing purpose. The user may assume that the sharing is undertaken for a primary purpose for which the data were collected. For example, the purpose to provide services requested by the user leads to underestimating the risk. Alternatively, the user may assume that the shared data is used for an unstated, secondary purpose, either by a first party or third party [6]. Secondary use can lead to overestimation of the privacy risk by users, despite that the third party's data practice remains unknown.

The overestimation of privacy risk is not a favorable situation for a company, because it can lead to either the user not using a service due to fear of data misuse, or it can lead to the regulator concluding that the data practice is not in compliance with a regulation. In 2015, the social networking website and application Snapchat changed its data practice descriptions in their privacy policy concerning collection, use and retention of their user data, stating that "…we may access, review, screen, delete your content at any time and for any reason" and "…publicly display that content in any form and in any and all media or distribution methods." Such statements led users to worry about the ways in which their information could be collected, retained and used, since the policy was extremely permissive. This led some users to report that they had deleted their accounts.[1] In another incident, Google was warned by European regulators about vagueness in their policy concerning data retention practices and about not showing a commitment toward the European Data Protection Directive.[2] Therefore, companies must be careful when using incompleteness to retain flexibility by avoiding unwanted increases in perceived privacy risk.

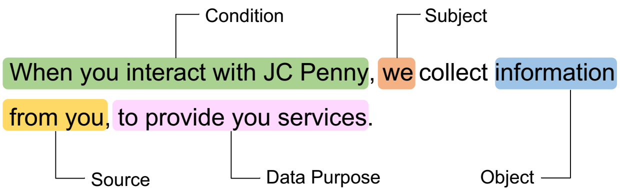

Condition
Subject
When you interact with JC Penny, we collect information
from you, to provide you services.
Source
Data Purpose
Object

Fig. 1 Example statement with annotated semantic roles

In this paper, we identify incompleteness by representing a data practice description namely a data action as a semantic frame. We construct these frames by identifying relevant questions for each data action, which we call semantic roles associated with the action. We propose to develop a network of semantic frames to determine the roles that are expected to complete a data practice description. In so doing, we aim to understand how roles contribute context for an action, and how policy authors choose roles when expressing privacy policies. For example, the following JCPenny privacy policy statement is annotated for semantic roles that describe the data action collected in Fig. 1. The condition on the action collect is "when you interact with JC Penny," the object is "information," the source of the information is "you," and the purpose of collection is "to provide you services."

This paper is organized as follows: in Sect. 2, we review background and related work; in Sect. 3, we describe our approach toward building semantic frames for data practices and our grounded analysis results; in Sect. 4, we present the user study design to measure the perceived privacy risk due to incomplete data practice descriptions and the study results; in Sect. 5, we report the threats to validity, in Sect. 6, we discuss our research questions in light of our results and present the future work and conclusions in Sect. 7.

2 Background and related work

We now review background and prior work related to semantic frames and roles in natural language, semantic frame representations for requirements and privacy risk.

Massey et al. [26] categorize ambiguity in natural language in legal text as follows:

[1] Sally French, "Snapchat's new 'scary' privacy policy has left users outraged," Market Watch, November 2, 2015. http://www.marketwatch.com/story/snapchats-new-scary-privacy-policy-has-left-users-outraged-2015-10-29.

[2] Zack Whittaker, "Google must review privacy policy, EU data regulators rule," ZDNet, October 16, 2012. http://www.zdnet.com/article/google-must-review-privacy-policy-eu-data-regulators-rule/.

- *Lexical* A word or phrase with multiple, valid meanings
- *Syntactic* A sequence of words with multiple valid grammatical interpretations regardless of context
- *Semantic* A sentence with more than one interpretation in its provided context
- *Vagueness* A statement that admits borderline cases or relative interpretation
- *Incompleteness* A grammatically correct sentence that produces too little detail to convey a specific or needed meaning
- *Referential* A grammatically correct sentence with a reference that confuses the reader based on the conduct.

In this paper, we study incompleteness, which occurs when the privacy requirement does not answer any of the questions the stakeholders might ask. We identify incompleteness in data practices by determining which of the expected roles for a data action are missing values in data practice statements. In order to determine the expected roles that will help us better understand a data action, we need to answer questions associated with that action, such as *who performs the action* and *on what data the action was performed*, among other questions [23]. The answers to these questions can be expressed in many different ways in a statement. For example, consider the following data practice statements:

- `We collect user information.`
- `The user information is logged by us.`
- `We gather information about our users.`
- `The user provides us with their information.`

While the above statements use different action words, such as *collect*, *log*, *gather* and *provide* and have different syntax, they also have similar meaning, which is that the user information is collected by the website. One representation that permits comparison among these statements is called semantic roles [23]. Roles are considered shallow representations, because they rely only on the relationship between a given word or role value and other clauses in the statement, and not among all the words in the statement. Using semantic roles, we represent the fact that there is a *collection* action taking place, the action is being performed by the subject *the website company*, and the object of the action is the *user information*. Semantic roles represent the relationship of the different clauses in the statement to the main action, like the subject and object [23]. The context of a data action can be expressed using different semantic roles, such as agent (who initiates and performs an action), patient (what undergoes the action and changes its state), instrument (used to carry out the action), source (where the action originated), among other roles [21].

Semantic roles that are used to describe a data action can be represented together in a knowledge representation technique known as *frames*. Minsky describes a frame as a data structure that is used to represent a stereotyped situation, such as being in a certain kind of living room [27]. Each frame is associated with *slots* or semantic roles, which are filled by *fillers* or semantic role values in specific contexts and which help readers understand a situation in question. The values for these semantic roles can be atomic values, procedures, or pointers to other frames [27]. Frames can be used to represent knowledge in a succinct manner and to reason in an efficient way [18].

According to Fillmore's frame semantics, the meaning of a word cannot be understood in isolation but in conjunction with the related information [22]. For example, the word "share" can be understood when we have knowledge about who is sharing, what is being shared and with whom it is being shared. Fillmore's frame semantics are implemented in the FrameNet project [5]. The FrameNet corpus contains manually annotated, general purpose semantic frames for the English language, with semantic roles specific to a frame. The *frames* are evoked by *lexical units* which are lemmas and their part of speech. The semantic roles associated with each frame are also known as *frame elements*, which provide information about the frame. Consider the following example from the FrameNet database:

`Abby bought a car from Robin.`

In this statement, the *frame* "commerce_buy" is evoked by the *lexical unit* "bought (buy.verb)." The *frame elements* of this frame instantiated in this statement are *buyer* (Abby), *goods* (a car), *seller* (robin). We represent privacy policy statements as frames that consist of semantic roles. These frames help us identify different semantic units in privacy statements that describe different attributes associated with data actions. These frames can also be nested to consolidate semantics across related statements. The technical challenge we address in this paper is to identify the type of frame for a given privacy statement, identify the spans in the statement that consist of words that constitute a semantic unit and finally analyze the extent to which these frames capture the semantic meaning of a privacy statement. Our results described in Sect. 3 show that the data action statements in privacy policies can be encoded using semantic frames. Similar to FrameNet, our frames are evoked by different categories of data actions, which represent a situation where the user's information is being acted upon by a company. We employ semantic roles that are specific to each such frame and are instantiated when that frame is evoked.

The FrameNet resource has been used for automatic semantic role labeling [15, 29]. Das et al. report an F1 score of 61.4 and 68.49 for frame identification and semantic role value identification, respectively, for SemEval 2007 data, and F1 score of 80.3 and 79.9 for frame identification and role value identification, respectively, on the FrameNet 1.5 release [15]. Semantic role labeling has been used for improving applications such as question-answering [24], recognizing textual entailment [16], information extraction [32] and in requirements engineering, to extract information from software requirements specifications [36].

In this paper, we identify the expected semantic roles for a given frame and consequently determine when the information provided is incomplete, by identifying roles that are missing values in a given data practice statement. This manuscript extends our previously published conference paper [9]. The *purpose* semantic role has been analyzed in our previous paper [10]. Our analysis in this study is limited to the contextual information provided in a single statement, and we do not combine contextual information from multiple statements. Because incomplete information prevents users from having control over their information and knowing when an entity has access to their information, it can also affect a user's perception of their privacy [3]. In addition, incompleteness prevents users from knowing the potential consequences of such disclosures. Tsai et al. found that users took privacy information into consideration while making decisions about using the services of an online website and were willing to pay to protect their privacy [33]. These findings make it important to identify the privacy risk perceived by a user due to incomplete information. Furthermore, websites can provide more complete information about their data practices to help users make better decisions about using the services provided by the website.

3 Semantic role representation and incompleteness

Our research questions are as follows:

RQ1 What are the different semantic roles associated with different categories of data actions, and how do they vary across website domains?

RQ2 What are the variations in the values of the different semantic roles within and across website domains?

RQ3 What are the different lexical and syntactic triggers that indicate semantic role values within and across website domains?

RQ4 How does the presence or absence of semantic roles and their values affect the user's perception of privacy risk?

To answer the first three research questions, we manually annotated semantic roles in 15 privacy policies from three domains: health, news and shopping. We conducted a survey to identify the types of websites that users most frequently use and found that *news* and *shopping* websites were most frequently used by our survey participants. Most of our participants reported that they read news online several times a day and shopped for products online a few times a week or more [10]. In addition, we chose to study the health domain, since it is a highly regulated domain and deals with sensitive user data. We chose a convenience sample of five policies per domain (see Table 1). For health, we chose companies that provide a diversity of services (DNA testing, online medical records, health clinics, wearable devices and an online symptom dictionary.) For news, we chose websites with a diversity of US viewpoints. Finally, for shopping we chose companies that maintain both online and "brick-and-mortar" stores. These choices were intended to diversify the observed practices.

Table 1 Privacy policy dataset for semantic frame study

Domain	Company name	Last updated
Health	23andMe	10/14/2015
	HealthVault	09/2016
	Mayo Clinic	10/06/2014
	My Fitness Pal	06/11/2013
	WebMD	03/20/2015
News	ABC News	10/18/2016
	Bloomberg	07/15/2014
	CNN	07/31/2015
	Fox News	10/26/2016
	Washington Post	01/01/2015
Shopping	Barnes and Noble	08/05/2016
	Costco	12/31/2013
	JC Penny	09/01/2016
	Lowes	08/20/2015
	Overstock	06/20/2017

3.1 Annotating and extracting semantic roles

The first three research questions concern the different semantic roles and their variations across different data actions, and the lexical and syntactic triggers that indicate semantic role values. We annotated the policies in Table 1 using content analysis, in which an analyst assigns codes to text from a coding frame [30]. Each coded text fragment represents an instance of the code, after which the analyst can review the coded items for insight into the phenomena of interest. In our analysis, we annotate different text fragments in each privacy policy statement. Our analysis is limited to statements about collection, retention, usage and transfer of personal information, which were first studied by Antón and Earp in their seminal paper on privacy goal mining [4].

We prepare the policies for annotation by removing section headers and boilerplate language and itemizing the policy into individual statements. In each statement, we identify the main data action and categorize the statement into one of five categories: *collection*, *retention*, *usage*, *transfer* and *other*. We only analyze the statements which belong to the first four categories, excluding *others*. Statements that belong to the *others* category are of the following kind, shown with examples from the policy named in parentheses:

- *Definitions* (Costco): "Personal information is information that identifies an individual or that can be reasonably associated with a specific person or entity, such as a name, contact information, Internet (IP) address and information about an individual's purchases and online shopping."
- *User actions* (Barnes and Noble): "You may also access, correct or change the personal information in your community profile(s) on SparkNotes.com at any time, except to change your username."
- *Scope of the privacy policy* (Lowes): "This Privacy Statement applies to the US practices of Lowe's Companies, Inc. and its US operating subsidiaries and affiliates except as outlined below."
- *Customer relations* (Overstock): "If you have questions about your order, you should direct them to us and not to the Vendor."

Next, we use the frame-based markup developed by Breaux and Antón to identify semantic roles associated with different data actions [11]. The tool can be used to extract requirements from natural language text. The tool is a good fit for our analysis as it allows us to use first cycle coding [30] and to segment the statement by identifying the phrases that correspond to roles, while accounting for variability in the statement due to logical conjunctions and disjunctions. The markup is then parsed to generate lists of roles based on each action and syntactic cue, which we discuss later. Consider the following example, which annotated statement using the tool and which is from the Lowes privacy policy:

```
[[This information] may be used {to
[provide a better-tailored shopping
experience]}, |and {for [<market
research, | data analytics, | and
system administration>purposes]}.]
```

Each statement from each policy in our dataset was annotated by one of the three authors, and then the annotations were checked by one other author. The guidelines the annotators use to annotate the statements are as follows:

- *Square brackets* are used to denote role fillers that are required to make the statement grammatically correct. For example, in the statement above, the object `[this information]` is required.
- *Curly brackets* are used to denote clauses that can be removed, which typically correspond to optional roles. For example, `{to [value]}` and `{for [value]}` curly-bracketed clauses in the statement above can be removed and the sentence would still be grammatically correct; however, if the words "to" and "for" are present, then the nested role values within the square brackets would be required for the statement to make grammatical sense. For instance, in the statement above, if we remove the roles in the "to" and "for" patterns, the statement would become: "This information may be used." Each statement is enclosed in a square bracket to demarcate sentence boundaries.
- *Angular brackets* are used when a phrase or clause contains alternative sub-clauses among which at most one sub-clause is needed to produce a grammatically correct sentence. For example, the phrase "and for" above applies to all phrases inside the angular brackets.

Annotating text fragments using the frame-based markup tool was an iterative process. For each policy, after the annotator was done annotating the policy, one of the other two authors checked the annotations and marked any disagreement or confusions. The author who annotated the policy and the author who checked the policy annotation then met to discuss any discrepancies. The discussion was continued until the two participating authors reached consensus and built the final version of the annotated dataset that was used for all subsequent studies.

After the annotation process, we code the extracted phrases in curly brackets using open coding [30] to assign semantic role names to these phrases. Example annotation-coded pairs are as follows:

- `[this information]`: object
- `{to [provide a better-tailored shopping experience]}`: data purpose
- `{for [<market research, | data analytics, | and system administration>purposes]}`: data purposes

In this statement, the lexical and syntactic patterns `to [value]` where *value* is "provide a better-tailored shopping experience," and `for [value]` where *value* is "market research, data analytics, and system administration purposes" are used to specify the data purpose role. These annotations were done in a manner similar to the one described above for annotating text fragments.

In order to identify the variations in semantic role values (RQ2), we begin with the coded roles values produced by applying the above method, and then we use open coding

[30] to categorize the role values for the *subject*, *condition*, *source* and *target* roles into different categories. Bhatia and Breaux categorized the purpose role values for the same policies in a prior study [8]. We answer research question RQ3, "what are the different lexical and syntactic triggers that indicate semantic role values within and across website domains?" by extracting all lexical and syntactic patterns from the 15 annotated policies using the frame-based markup tool [11]. Next, we analyze the results to determine how the same pattern, when used with different data actions, indicates different semantic roles and how different patterns lead to the same semantic role. Finally, we analyze the syntax of the observed patterns by identifying categories of prepositions [1] present in the patterns and their associated semantic roles.

3.2 Semantic roles content analysis results

In this section, we describe the results to answer RQ1–RQ3. The first research question RQ1 concerns the identification of different semantic roles associated with different categories of data actions within and across website domains. We identified a total of 17 unique semantic roles across the 15 policies and across the four categories of data actions. We reached saturation in semantic roles after we analyzed the first two privacy policies, Barnes and Noble and Costco. The most frequent semantic roles are defined as follows, with the question answered in parentheses (see "Appendix A" for the complete list of semantic roles):

- *Subject* The entity which acts on the information. (Who is performing the data action?)
- *Object* The data on which the action is being performed. The values of this role were information types in our study. (What is being acted upon?)
- *Purpose* The goal or justification for which the action is performed. (Why is the information being acted upon?)
- *Condition* The states or events under which the data action will be performed on the information. (When will the data action be performed?)
- *Source* The provider of the information in a collection action. (From whom is the information collected?)
- *Target* The recipient of the information in the transfer action. (Who is the data being transferred to?)

Table 2 presents the frequency of semantic role values for each data action category, across all the policies and domains shown in Table 1 (see "Appendix B" for policy wise frequency). Note that some actions have multiple instances of the same semantic role attached to them.

From our analysis, we found that transfer actions had the highest number of semantic roles attached, followed by use actions. Policies across all three domains were least descriptive about retention actions. We also observed that the health policies were the most descriptive with a total of 293 actions 1024 semantic roles across all categories, followed by shopping policies with a total of 281 actions and 878 semantic roles (see "Appendix B"). The news policies were the least descriptive with only 124 actions and 414 semantic roles across the five policies. This could be because the health domain is highly regulated and mostly deals with sensitive user data, as compared to shopping and news domain. The shopping policies had the highest number of collection actions, whereas the health policies had the highest number of retention, use and transfer actions.

Table 2 Frequency of semantic role values across data action categories

Semantic role	Collection	Retention	Use	Transfer
Total actions	167	63	241	227
Action location	2	3	12	8
Comparison	0	0	1	4
Condition	66	25	60	106
Constraint	4	3	13	11
Duration	0	4	0	0
Exception	0	1	3	14
Hypernymy	28	3	14	8
Instrument	22	1	6	10
Negation	13	4	17	29
Object	167	63	241	226
Purpose	34	16	190	49
Retention location	0	13	0	1
Retention property	0	2	0	0
Source	50	2	7	4
Subject	154	49	196	196
Target	6	0	2	141
Time of action	4	4	1	3
Total no. of semantic role values	550	193	763	810

In our analysis all of the collection, retention and usage actions across all the three domains have the object role attached, whereas one of the transfer actions is missing the object role in the Costco privacy policy. In our privacy surveys (see Sect. 4.2), we observe that the participants were the least willing to share their information for transfer actions, and not clearly specifying what information is transferred may further increase the perceived risk.

The most frequent role across all actions and across all three domains was object, followed by subject. The other three most frequent semantic roles are purpose, condition and target across all three domains. The purpose role occurred most frequently with use actions and the condition role with transfer actions. The health and news policies did not contain instances of the semantic

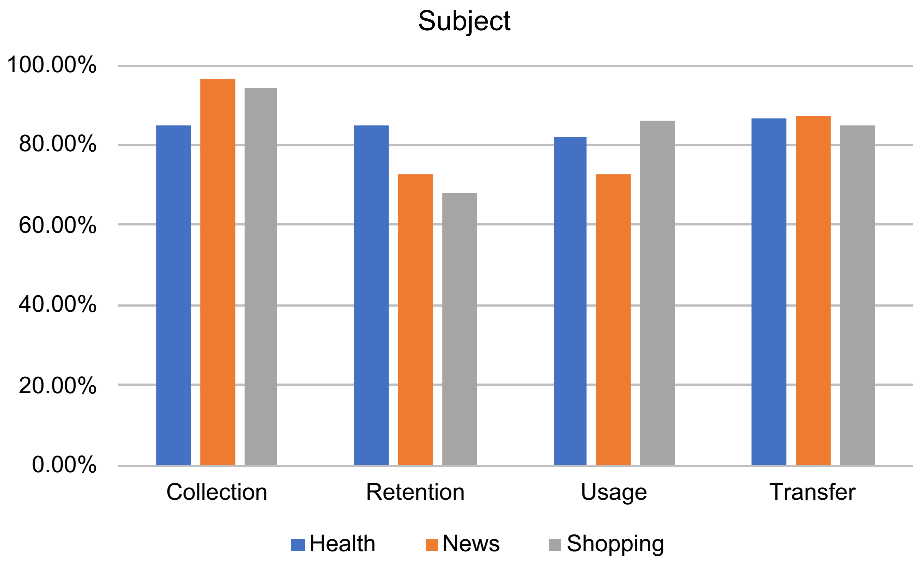


Fig. 2 Frequency of subject role across action categories and website domains

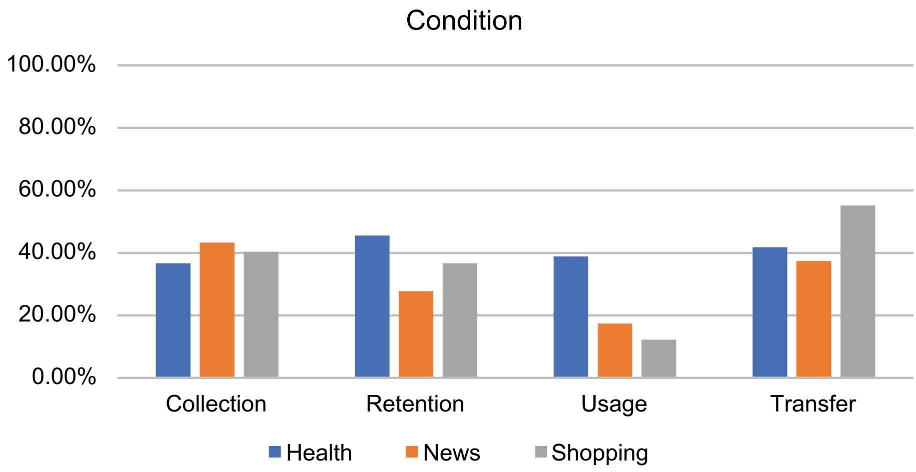


Fig. 3 Frequency of condition role across action categories and website domains

role retention property, which was present in the shopping policies. In addition, the role "comparison" was not found in the news policies.

In Figs. 2, 3 and 4, we show the frequency of semantic roles subject, condition and purpose for each category of data action across the three domains. Most of the actions across the three domains had the subject role: 84.7% of the actions in health domain, 82.4% in news domain and 83.5% in shopping domain had the subject role. The condition role was not as frequent with only 40.6% action in health, 31.4% actions in news and 36.0% actions in shopping domain had a condition role attached. Similarly, the purpose role was also not frequently found, 38.0% of health actions, 41.8% of news actions and only 33.9% of shopping actions had a purpose role.

We observed from our analysis (see Fig. 2) that most of the collection actions, specifically 92.1% on average across the three domains, had the subject role attached. In the shopping domain 94.4% and in news 96.7% of collection actions have the subject role attached, as compared to the health domain where only 85.1% of the collection actions have the subject role. This was closely followed by the transfer actions where 86.5% of all the transfer actions across all domains had the subject role, and 87% (health), 87.5% (news) and 85.1% (shopping) transfer actions had the subject role. On the other hand, only 80.1% of usage and 75.3% of retention actions had the subject role on average across all three domains.

Around 55.2% of the transfer actions in the shopping domain have the condition role attached, whereas only 41.7% of health and 37.5% of news transfer actions have the condition role. On the other hand, 45.5% of the health retention actions have the condition role, as compared to 36.8% of shopping and 27.3% of news retention actions (see Fig. 3).

A large number of usage actions (79.6%) have the purpose role, whereas only a small number of retention (28.0%),

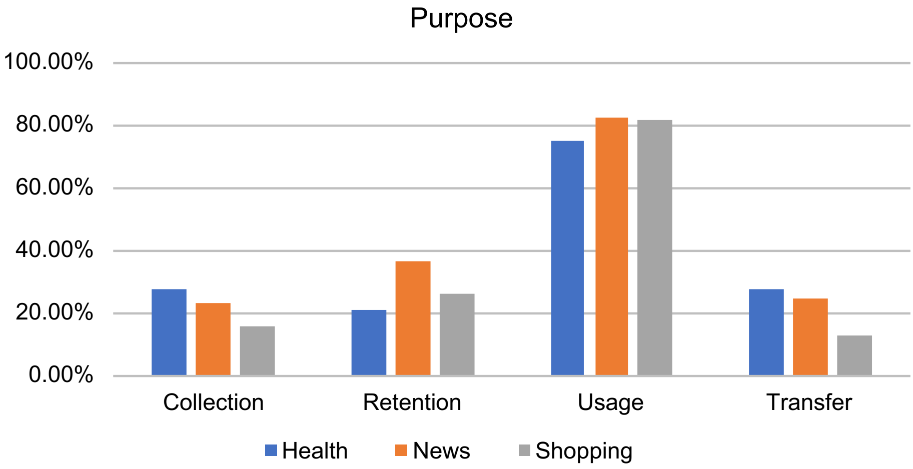


Fig. 4 Frequency of purpose role across action categories and website domains

collection (22.2%) and transfer (21.8%) actions have the purpose role attached, across all the three domains (see Fig. 4).

From our analysis, we observed that on average the actions in health policies had the maximum number of subjects (84.7%) and conditions (40.6%) attached as compared to actions in news policies (subject: 82.4%, condition: 31.4%) and shopping policies (subject: 83.5%, condition: 36.0%). On average, 41.8% of the news actions had the purpose role, as compared to 38.0% health and 33.9% shopping actions.

From our risk surveys (see Sect. 4.2), we observed that the privacy risk perceived by the users decreases if the condition and purpose roles are specified. Incomplete description of data practices with missing role values for condition and purpose could in turn decrease user's willingness to share their information with the website and consequently their use of the services provided by the website.

We further observe that different action words are used to describe data practices belonging to the same data action category. For example, the action words *log*, *submit*, *gather* and *collect* are all used to describe *collection* practices. The action word *log* is often used when the data collection is implicit, or automated, and occurs when the user is browsing or using the website. For example, in the statement, "Like most web sites, our servers log your IP address, the URL from which you accessed our site, your browser type, and the date and time of your purchases and other activities." The action word *submit*, however, is often used when the user submits their information to the website, for example, "When you place an international order, you will submit personal information (e.g., your name, email address, billing address, and shipping address) and other order-related information to JCPenny through and to servers located in the United States." This can include the user's name, address and payment details, in contrast to logged information that includes IP address and browser type. Thus, different action words depict subtle differences in which objects are associated and expected, despite being within the same broader category.

In Sects. 3.2.1–3.2.4, we describe the results from open coding [30] the role values for condition, source, target and subject roles to answer the second research question (RQ2) which concerns the variations in semantic role values. Bhatia and Breaux previously analyzed the role values for purposes in privacy policies; thus, we did not include this role in our analysis [8].

3.2.1 Categories of values for condition role

We identified 280 instances of the condition role across the 15 policies. The condition categories are as follows:

- *First-party action* The data action is conditioned on an action performed by the website company itself.
- *Legal* The data action is performed, if it is required by law.
- *Merger* The data action is performed, if the company is part of a merger or acquisition.
- *Scope* The data action performed is limited by practices described in the privacy policy.
- *Third-party action* The data action is performed in response to an action performed by a third party.
- *User* The data action is conditioned on an action performed by the user, or a property that the user possesses.

Table 3 presents the condition role categories with examples and frequency across all 15 policies. The most frequent condition category across all the three domains is *user*, followed by *first-party action* for news and shopping, and *legal* and *vague* for health. We also noted that health policies have a higher number of third-party actions as conditions as compared to news and shopping.

Table 3 Condition categories

Category	Examples	% Frequency		
		Health	News	Shopping
First-party action	Only if we identify a biometric match to our database of known shoplifters, in the receipt of automatically collected information	7.9%	19.2%	12.9%
Legal	If we believe we are required to do so by law, or legal process, as we deem appropriate in response to requests by government agencies	13.5%	7.7%	5.9%
Merger	As part of any merger or sale of company assets or acquisition, if some or all of our business assets are sold or transferred	1.6%	5.8%	8.9%
Scope	As permitted by this privacy policy	3.2%	1.9%	1.0%
Third party	If any of these service providers need access to your personal information, when they no longer need it	9.5%	1.9%	2.0%
User	if you choose to connect your mobile device to the free in-store Wi-Fi available at Lowe's stores, if you are under 18	50.8%	55.8%	61.4%
Vague	As necessary	13.5%	7.7%	7.9%
Total number of condition instances		126	52	102

Table 4 Source categories

Category	Example role values	% Frequency		
		Health	News	Shopping
Technology	Your computer and mobile device, third-party cookies	22.2%	0.0%	22.6%
Third party	Third-party sources, public sources	16.7%	14.3%	38.7%
First party	WebMD website	5.6%	7.1%	0.0%
User	You, children under the age of 13	55.6%	78.6%	35.5%
Vague	Various sources	0.0%	0.0%	3.2%
Total number of source instances		18	14	31

3.2.2 Categories of values for source role

The source role describes the information provider. We identified 63 source role instances across all 15 policies, which were categorized using open coding as follows:

- *Technology* The source of collected information is a device or technology.
- *Third party* The information about the user is collected from a third-party.
- *First party* The information about the user is collected from a first party.
- *User* The information is collected from the user.
- *Vague* The source of information is present, but unclear.

Table 4 presents the source categories with examples and their frequencies across all policies and domains in our dataset. Users were the main source of information for health and news policies, whereas for shopping websites the source of information was equally likely to be third-party sources or the users themselves.

The collection of information from technology or from third parties is generally automated, and the user may be unaware that the collection is taking place. In contrast, information collected from the user can be explicit collection, when the user provides their information to the company directly through a website.

3.2.3 Categories of values for target role

We identified a total of 150 instances of the target role, which describes the information recipient in a transfer action, and categorized these instances as follows:

- *First party* The information is transferred to the first-party website company.
- *Third party* The recipient of the information is a third party.
- *Location* The target is the location where the information is being transferred.
- *Technology* The information is being transferred to a technology.

Table 5 Target categories

Category	Example role values	% Frequency		
		Health	News	Shopping
First party	JC Penny, us	4.2%	4.5%	7.0%
Third party	Third parties, issuer of the Mastercard	90.1%	90.9%	80.7%
Location	Countries, globally	1.4%	0.0%	3.5%
Technology	Servers, mobile devices	2.8%	0.0%	5.3%
User	You	1.4%	4.5%	0.0%
Vague	Others, anyone	0.0%	0.0%	3.5%
Total number of target instances		71	22	57

- *User* The recipient of the information is the user.
- *Vague* The target of the information is present, but unclear.

Table 5 presents the target categories, examples and frequencies across the 15 policies in our dataset (see Table 1). Most of the information was transferred to third parties for all the three domains. Health and news websites were not vague about the target of shared information, when specified.

3.2.4 Categories of values for subject role

We identified 595 instances of the subject role across the 15 policies. The subject categories are as follows:

- *First-party action* The data action is performed by the website company itself.
- *Third-party action* The data action is performed by a third party.
- *User action* The data action is performed by the user.
- *Vague* It is not clear who performs the action.

Table 6 presents the subject role categories with examples and frequency across all 15 policies. Most of the actions across all domains are performed by the first-party companies. It is interesting to note that none of the subjects in the shopping domain were vague.

3.2.5 Lexical and syntactic patterns

Lexical and syntactic patterns are used to coordinate role values in a role phrase or clause. Pattern consists of one or more role keyword(s) followed by a slot value. The keywords correspond to the role semantics, e.g., the keyword “before” indicates a temporal relationship between the data action and another event, whereas the keywords “as long as” indicates persistence with an event. The patterns are both lexical and syntactic: lexical, because they bind words or lexemes to the role; syntactic because the arrangement of the words and lexemes distinguish the role with respect to the data action. For example, the pattern “to [value]” when used with the action “provide” indicates a target of the action, whereas changing the value to another action transforms this role into a purpose for which the action is performed, e.g., “provide information to a third party” wherein the pattern is instantiated as “to [a third party].” Alternatively, the “to [value]” pattern can be used to specify the purpose semantic role as follows: “To fill your prescription, we collect your name,” wherein the pattern is instantiated as “to [fill your prescription].” Although both examples use the “to [value]” pattern, the syntax of role phrases and clauses (where they are positioned in the sentence, and what phrases appear in their respective values) determine whether the semantic role is a target or a purpose.

Lexical and syntactic patterns describe how keywords attach to different data actions, and as part of syntactically different statements, they specify similar or different semantic role values. To answer RQ3, we identified 74 patterns, with 504 instances across health policies, 235 instances across news policies, and 380 instances across shopping policies in our dataset. Each pattern occurred one or more times in our dataset. Table 7 presents the five of the most frequent patterns, with example consisting of the semantic role name, followed by a colon and an example role phrase from the policy. The rest of the patterns are reported in “Appendix C.” For each frequent pattern, we also present the pattern frequency across the 15 policies. As the last row of the table, we present the total instances of all the 74 patterns identified across each domain. For example, a total of 504 patterns were identified in the health domain, among which 31.5% were `to [value]`.

Table 6 Subject categories

Category	Examples	% Frequency		
		Health	News	Shopping
First party	We, some of our tools	79.4%	83.3%	76.3%
Third party	Research contractors, third parties	11.7%	8.8%	18.8%
User	You, user	7.3%	6.9%	4.9%
Vague	Whoever has the access code, programs	1.6%	1.0%	0.0%
Total number of subject instances		248	102	245

Table 7 Lexical and syntactic patterns

Pattern	Semantic roles and values	%Frequency		
		Health	News	Shopping
to [value]	*Purpose*: to provide location-based services *Target*: to servers *Object*: to personally identifiable information	31.5%	29.4%	28.4%
if [value]	*Condition*: if Barnes and Noble becomes involved in a merger	6.0%	6.4%	8.2%
with [value]	*Condition*: with your consent *Object*: with other information *Target*: with other companies	5.2%	8.1%	7.9%
when [value]	*Condition*: when you interact with JC Penney	5.2%	8.5%	7.6%
from [value]	*Source*: from you, action *Location*: from our files	4.8%	5.5%	7.6%
Total number of instances		504	235	380

Table 8 Keywords used to specify different semantic role values

Semantic role	Keywords used
Object	along with, in conjunction with, to, with
Condition	according to, as, as part of, as long as, as well as, along with, at, based on, before, by, depending on, each time, even if, from, if, if and only if, if and when, in connection with, in the good faith belief that, in the event that, in, provided that, once, only if, when, with, without, unless, upon, until
Purpose	as, allowing, in, in an effort to, in order to, for, only as, to, that, so, so that, that, where
Target	among, between, in, only with, outside, to, with
Source	across, from, that, through

Another frequent pattern in the health domain is `for [value]` (8%), in the news domain is `such as [value]` (5.5%) and in the shopping domain is `as [value]` (3%). We observe that the same lexical and syntactic pattern is used to specify different semantic roles, when attached to different data actions and across different statements. The semantics conveyed by these patterns changes when attached to different data actions and in different contexts. For example, the syntactic pattern with the keyword `to [value]` can be used to introduce different semantic roles in the context of different data actions:

- `to [data purpose]`
  "We will *store and use* this information to administer the programs and services in which you choose to participate, and as permitted by this Privacy Policy."
- `to [target]`
  "In addition, we *disclose* certain personal information to the issuer of the MasterCard in connection with the administration of the Barnes and Noble MasterCard program."

  In addition, different syntactic patterns can be used to introduce the same semantic role. For example, the syntactic pattern `if [value]` and `depending on [value]` can be used to specify the condition role.
- `if [condition]`
  "If Barnes and Noble becomes involved in a merger, acquisition, restructuring, reorganization, or any form of … some or all of its assets personal information and your transaction history may be provided to the entities …"
- `depending on [condition]`
  "Depending on how you choose to interact with the Barnes and Noble enterprise we may collect personal information …"

In our dataset, we observed that although the patterns `if [value]` and `depending on [value]` both represent the role condition, they cannot be used interchangeably. This is because in our dataset the semantic role values that occur with `if` are specific and the values occurring with `depending on` are comparatively generic set of conditions, which can take one of many possible values.

Table 8 presents the keywords for each of the most frequent roles across the 15 policies.

Across the three domains, we observed that similar patterns were used to specify the conditions, source and target semantic roles. The most frequent patterns used to specify the condition role for health is `if [value]`, `when [value]` and `in [value]`; for news is `when [value]`, `if [value]` and `unless [value]` and for shopping is `if [value]`, `when [value]` and in [value]. The pattern that was used to specify most of the source roles across all the three domains is from [value]. The patterns used to specify most of the target roles across the three domains are to [value] and with [value].

We noticed from our analysis that the semantic role specified by a pattern is also dependent on the action category with which it occurs. For example, in the shopping policies, the pattern to [value] occurs 58 times with usage actions, and in 57/58 times, this pattern coincides with the purpose role. When the pattern is attached to transfer actions, it occurs 36 times and 31/36 times it coincides with a target role. Some of the patterns such as if [value], depending on [value], and when [value] are only used to specify the condition role.

We further evaluate our 74 unique patterns under the assumption that the majority of patterns share the syntactic quality of beginning with a preposition. Leveraging this observed pattern quality, we employ preposition categorization [1]. In the category examples below, the variable [value] represents an instance of a semantic role value that occurs with the prepositions. The categorization is based on the properties of the preposition and is independent of the instantiation of the variable [value].

Our analysis across the patterns can be characterized by the following properties of prepositions:

- *Transitive* A single preposition that takes a noun phrase, an adjective phrase, an adverb phrase, a prepositional phrase, or a clause as a complement, e.g., to [value]
- *Intransitive* A single preposition that does not require a complementing phrase or clause, e.g., when [value]
- *Deverbal* A preposition that takes the form of a participle, e.g., during [value]
- *Complex* A preposition that consists of two or more words, e.g., as part of [value]
- *None* Pattern does not contain a preposition

The 74 patterns contain 28 transitive preposition patterns, four intransitive preposition patterns, five deverbal preposition patterns, 35 complex preposition patterns and 2 patterns without prepositions. We refer to the transitive, intransitive and deverbal categories as single-preposition patterns. Our analysis shows that the majority of our patterns can be characterized by having complex or simple syntactic structure: 47% of patterns fall in the complex category, single-preposition patterns comprise 50%, leaving 3% of patterns without a preposition. We then examined patterns across categories with shared prepositions, specifically complex preposition and single-preposition patterns that end with the same preposition, for example, as [value] and except as [value]. We found that the last preposition in a complex preposition pattern can in fact diversify the semantic role value from that of the parallel, single-preposition pattern. For example, if we consider the complex pattern in a manner similar to [value] and the single-preposition pattern to [value], we know they will both contain prepositional phrases beginning with the preposition “to.” Examining the aforementioned “to” patterns in the policies from the health domain, we find that out of the four complex patterns that end in “to,” the semantic role value occurs once as a constraint using the pattern in a manner similar to [value] and once as a constraint using the pattern in addition to [value] and as twice as a purpose using the pattern in order to [value]. We find that out of the 159 to [value] patterns from the health domain, the semantic role value is 68% purpose, 30% target and infrequently as object and source. This example suggests that the complex patterns semantic role value is dependent on a noun phrase within the complex preposition. In this example, we note that “a manner similar” and “addition,” which both invoke a constraint role value, contrast with “order,” which implies a purpose similar to that of the majority of the single-preposition semantic roles for “to” patterns in the health domain.

4 Semantic roles and privacy risk

In this section, we describe the study designs and results for measuring the effect of semantic roles on privacy risk.

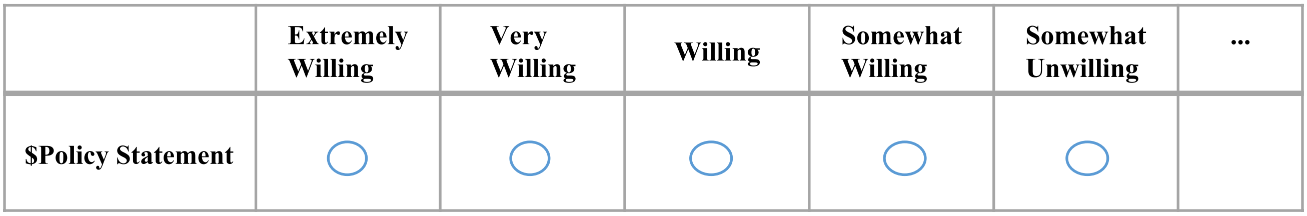
4.1 Privacy risk study design

Research question RQ4 asks, “how does the presence or absence of different semantic roles affect the user’s perception of privacy risk?” Fischhoff et al. [19] describe risk as the individual’s willingness to participate in an activity. To answer RQ4, we modified the empirical framework developed by Bhatia et al., which uses factorial vignette surveys and multilevel modeling to measure the change in perceived risk due to different factor levels [7, 10]. The modifications include introducing factors that correspond to semantic roles, noting that some sentences will include these factors while others will exclude these factors. Multilevel modeling is a statistical regression model with parameters that account for multiple levels in datasets. In addition, the model limits the biased covariance estimates by assigning a random intercept for each subject [20]. This random intercept allows us to account for subject-to-subject variability. Our surveys have multiple independent factors that affect a user’s perception of privacy risk. Multilevel modeling analysis helps us determine how each factor individually and in combination with other factors affects the privacy risk perception.

In each vignette, we present participants with a scenario that consists of multiple factors, also called independent

Table 9 Study 1 vignette factors and their levels

Factors	Factor level
Risk likelihood (`$RL`) Between subject	Only one person in your family
	Only one person in your workplace
	Only one person in your city
	Only one person in your state
	Only one person in your country
Data actions (`$DA`) Within subject	(C) Collection: collect
	(R) Retention: retain
	(U) Usage: use
	(T) Transfer: share
Semantic Role (`$SR`) Within subject	(DP) Data Purpose: to provide you services
	(Cond.) Condition: when you create an account with us
	(Source) Source: from you



	Extremely Willing	Very Willing	Willing	Somewhat Willing	Somewhat Unwilling	...
$Policy Statement	○	○	○	○	○	

Fig. 5 Template used for vignette generation (fields with $ sign are replaced with values selected from Tables 9 and 10)

variables. In addition, the vignette consists of a risk likelihood level and a risk acceptance scale [7]. The risk likelihood scale developed by Bhatia et al. is based on construal level theory, which shows that a privacy violation affecting *only one person in your family* is considered psychologically closer and more salient than *only one person in your country* [7, 34]. The privacy risk framework measures the privacy risk as the user's willingness to share their data, which is the dependent variable for the factorial vignette surveys, *willingness to share* (`$WtS`), and is estimated from participant ratings on an eight-point, bipolar semantic scale, labeled at each anchor point: *1 = Extremely Unwilling*, *2 = Very Unwilling*, *3 = Unwilling*, *4 = Somewhat Unwilling*, *5 = Somewhat Willing*, *6 = Willing*, *7 = Very Willing and 8 = Extremely Willing*. In a posttest, participants answer demographic questions, including their gender, age range, education level, ethnicity and household income.

We conducted three studies to measure the effects of the presence or absence of different semantic roles on privacy risk. The survey participants were recruited from Amazon Mechanical Turk, had completed ≥ 5000 Human Intelligence Tasks and had an approval rating of 97% or greater. The nature of crowdsourcing is that some people can choose to see the survey task, but not take the survey. Similar to using third-party mailing lists where one does not have access to the complete list, we could not measure the non-response rate to our survey. However, we can compare the demographic characteristics of respondents to those of the general Internet user population. As compared to the 2015 PEW Internet and American Life Survey data of US Internet users, the participants that we recruited from Amazon Mechanical Turk had less reported Asian, Black and Hispanic participants [28].

The surveys were published on Survey Gizmo. We recruited 80 participants for each of the three surveys. Participants were allowed to take each survey only once, and the same participant was allowed to take all three surveys. The participants of the first survey were paid $3, and those of the second and third survey were paid $2.

We now describe privacy risk survey modifications.

4.1.1 Semantic roles and privacy risk

These studies aim to measure the effect of the presence or absence of different semantic roles across all four data action categories on the perceived privacy risk. To that end, we fixed the values of the *subject* role and *object* role to be "we," and "personal information," respectively. Table 9 presents the factors and corresponding factor level values. Figure 5 presents the factorial vignette survey text.

The baseline policy statement for our survey was "We `$DataAction` your personal information," which includes the semantic roles *subject* and *object* associated with the data action. The policy statements `$Policy`

Table 10 Study 2 vignette factors and their levels

Factors	Factor level
Semantic role ($SR) within subject	(Cond.) Condition: with your consent
	(Source) Source: from you
	(Target) Target for the data action Transfer: third parties

Statement for each of the four actions are generated by adding one or more of the semantic roles from Table 9 to the baseline statement. For this survey, we have three different semantic roles, and therefore a total of eight policy statements for each action including the baseline statement, with all combinations of one or more of the semantic roles. For example, the *collection* statement with the roles *data purpose* and *condition* would be: "When you create an account with us, we collect your personal information to provide you services."

The second study has the same three independent variables: risk likelihood, data action and semantic roles. However, the levels and the values of the independent variable semantic roles in Study 2 are different than in Study 1, as shown in Table 10. The levels for the *risk likelihood* and *data action* variables are the same for Study 1 and 2. Table 10 presents the additional factors and factor levels for the semantic roles used in Study 2.

4.1.2 Semantic role value categories and privacy risk

In the grounded study, we categorized the role values for the *condition*, *source* and *target* roles (see Sects. 3.1 and 3.2). The semantic role value categories can affect a user's perception of privacy risk. A user may be more willing to share their information, if the data action is *required by law*, as compared to if the action is performed *as necessary*, which is a vague condition. The most frequent roles in our policy statements after the subject and object roles were condition, source and target. The third study has three pages with all the role value categories for a particular semantic role on each page. Table 11 presents the factor (a semantic role), the breakout for each semantic role category, followed by the factor levels, which is the semantic role value per category.

Reusing the survey design from Fig. 5, the $Policy Statement is generated by adding the semantic role value category to the baseline statement, "we transfer your personal information" for the *condition* and *target* roles, and "we collect your personal information" for the *source* role.

The risk perception could in addition be affected by other semantic roles such as purposes or subjects. We leave those studies to future work.

4.2 Privacy risk survey results

We now describe our results from three studies described in Sect. 4.1 to answer RQ4, which concerns the effect of presence or absence of semantic roles and their values on the user's perceived privacy risk. We report the survey results in two separate series: the first series measures the effect of the four data action categories and the condition, source, purpose and target roles on perceived privacy risk;

Table 11 Study 3 vignette factors and their levels

Factors	Category	Factor level
Condition ($Cond) Within subject	first-party action	As part of your member profile
	legal action	If we are required to do so by law
	merger action	As part of a merger
	scope	As permitted by this privacy policy
	third-party action	If third-party service providers need access to your information
	user	With your consent
	vague	As necessary
Source ($Source) Within subject	technology	From your computer and mobile device
	third party	From third-party sources
	user	From you
	vague	From various sources
Target ($Target) Within subject	first party	To us
	third party	To third parties
	location	Globally
	technology	To servers
	vague	To others

Table 12 Study 1 multilevel modeling results

Term	Coeff.	SE
Intercept (DataAction-collect)	4.588***	0.378
Risk: only 1 person in your workplace	−0.242	0.524
Risk: only 1 person in your city	−0.697	0.524
Risk: only 1 person in your state	0.197	0.524
Risk: only 1 person in your country	0.021	0.524
Data Action: retain	0.097	0.068
Data Action: transfer	−0.413***	0.068
Data Action: use	0.039	0.068
Baseline + condition	0.006	0.096
Baseline + condition + purpose	0.397***	0.096
Baseline + condition + purpose + source	−0.444***	0.096
Baseline + condition + source	0.016	0.096
Baseline + purpose	0.478***	0.096
Baseline + purpose + source	0.313***	0.096
Baseline + source	−0.794***	0.096

*$p \leq .05$ **$p \leq .01$ ***$p \leq .001$, 4 = somewhat unwilling

Table 13 Study 2 multilevel modeling results

Term	Coeff.	SE
Intercept (DataAction-collect)	3.795***	0.354
Risk: only 1 person in your workplace	0.078	0.496
Risk: only 1 person in your city	1.340	0.481
Risk: only 1 person in your state	0.791	0.488
Risk: only 1 person in your country	0.088	0.488
Data Action: retain	−0.222	0.088
Data Action: transfer	−1.341	0.088
Data Action: use	−0.328	0.088
Baseline + condition	0.744***	0.088
Baseline + source	0.081	0.088
Baseline + target	−0.141	0.149
Baseline + condition + source	0.784***	0.088
Baseline + condition + target	0.684***	0.149
Baseline + source + target	−0.104	0.149
Baseline + condition + source + target	0.659***	0.149

*$p \leq .05$ **$p \leq .01$ ***$p \leq .001$, 4 = somewhat unwilling

and the second series measures the changes in privacy risk due to different role values for the condition, source and target roles.

4.2.1 Data action categories and semantic roles

The first and second studies described in Sect. 4.1.1 measure the effect of the presence and absence of the condition, source, purpose and target roles on the participant's willingness to share their information.

Equation 1 is our main additive regression model for studies 1 and 2 with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood of a privacy violation, `$DA`, which is the data action, and `$SR`, which is the semantic role (see Tables 9 and 10). The additive model formula defines the dependent variable `$WtS` (willingness to share) in terms of the intercept α and a series of components, which are the independent variables. Each component is multiplied by a coefficient (β) that represents the weight of that variable in the formula. The formula in Eq. 1 is simplified as it excludes the dummy (0/1) variable coding for the reader's convenience.

$$\texttt{\$WtS} = \alpha + \beta_{\mathrm{R}}\texttt{\$RL} + \beta_{\mathrm{DA}}\texttt{\$DA} + \beta_{\mathrm{DA}}\texttt{\$SR} + \epsilon \quad (1)$$

Tables 12 and 13 present the results for the baseline statement "We `$DataAction` your personal information." In Tables 12 and 13, the row baseline + semantic role(s) presents the value of the coefficient for the statement which is constructed by adding the semantic role(s) to the baseline statement. A positive coefficient signifies an increase in `$WtS`, and a negative coefficient represents a decrease in `$WtS` over the baseline.

We observe that adding the source role to the baseline statement (e.g., from you) decreases the participant's willingness to share. In addition, specifying the purpose role in any situation increases the willingness to share. Participants were less willing to provide their information when their data can be transferred as compared to when their data is collected by the website. Table 13 presents the modeling results for Study 2.

In Study 2, we observe that adding the condition role, which concerns seeking consent from the user before their data are acted upon, considerably increases the participant's willingness to share their information. In both surveys, we did not observe any statistically significant difference among the levels of the factor *risk likelihood*.

4.2.2 Semantic role value categories

We now report results from Study 3 to measure the effect of role values on perceived privacy risk. The policy statements for this survey were generated by adding the role value category to the baseline statement, "we transfer your personal information" for the condition and target roles, and "we collect your personal information" for the source role.

In Eqs. 2.1, 2.2, and 2.3, we present our main additive regression models for study 3, with a random intercept grouped by participant's unique ID (ϵ), the independent within-subjects measure `$RL`, which is the likelihood

Table 14 Study 3 multilevel modeling results

Term	Coeff.	SE
Semantic role: condition, baseline: "first-party action"		
Intercept (first party)	3.113***	0.355
Condition: legal	1.788***	0.196
Condition: merger	−0.188	0.196
Condition: scope	0.775***	0.196
Condition: third party	−0.875***	0.196
Condition: user	2.213***	0.196
Condition: vague	−0.150	0.196
Semantic role: source, baseline: "technology"		
Intercept (technology)	2.325***	0.399
Source: third party	0.100	0.173
Source: user	2.000***	0.173
Source: vague	0.163	0.173
Semantic role: target, baseline: "first party"		
Intercept (first party)	3.245***	0.330
Target: location	−1.775***	0.159
Target: technology	−0.050	0.159
Target: third party	−1.438***	0.159
Target: vague	−1.525***	0.159

*$p \leq .05$ **$p \leq .01$ ***$p \leq .001$, 4 = somewhat unwilling

of a privacy violation, and `$DA`, which is the data action, and `$Cond` which is the condition role, `$Source` which is the source role, `$Target` which is the target role, (see Table 11).

$$\texttt{\$WtS} = \alpha + \beta_{\mathrm{R}}\texttt{\$RL} + \beta_{\mathrm{DA}}\texttt{\$DA} + \beta_{\mathrm{DA}}\texttt{\$Cond} + \epsilon \tag{2.1}$$

$$\texttt{\$WtS} = \alpha + \beta_{\mathrm{R}}\texttt{\$RL} + \beta_{\mathrm{DA}}\texttt{\$DA} + \beta_{\mathrm{DA}}\texttt{\$Source} + \epsilon \tag{2.2}$$

$$\texttt{\$WtS} = \alpha + \beta_{\mathrm{R}}\texttt{\$RL} + \beta_{\mathrm{DA}}\texttt{\$DA} + \beta_{\mathrm{DA}}\texttt{\$Target} + \epsilon \tag{2.3}$$

The baseline for the condition category is "first party," the baseline source is "technology," and the baseline target is "first party." The results appear in Table 14.

We observe from Table 14 that when information will be transferred on condition of a user consent action, as required by law, or as permitted by the policy, elsewhere, the user's willingness to share increases above the baseline. On the other hand, third-party condition ("if third-party service providers need access to your information") decreases the willingness to share below the baseline, whereas the differences between merger and vague condition as compared to the baseline condition are not statistically significant. We observed that the user's willingness to share increases when the information is collected from the user, directly, as compared to when it is collected from their computer or mobile device. With respect to the target role, the user's willingness to share decreases when the information is transferred to third parties, or the target role value is vague.

5 Threats to validity

Construct validity addresses whether what we measure is actually the construct of interest [37]. To mitigate threats to construct validity, the annotations were performed by one author and then checked by the other author for all the policies in our dataset. The privacy risk framework we use for our studies assumes that a person's willingness to share their information corresponds to their acceptance of the risk [7], which was also used in other studies by Acquisti and Knijnenburg to measure risk related to privacy [2, 25]. As noted by Bhatia et al. [7], semantic scale anchor labels used for the dependent variable `$WtS` in the risk surveys could be interpreted differently by participants [13]. To mitigate this threat, we designed our independent factors `$RL`, `$DA`, `$SR`, `$Cond`, `$Source` and `$Target` as within-subject factors, such that all the participants see and respond to all levels of the independent variables. Subject-to-subject variability is accounted for in our analysis by the random intercept.

Internal validity concerns whether our correlation of the independent and dependent variables is valid [37]. The selection of the number of vignettes to be rated by a participant must take into account multiple factors, including fatigue experienced by the participant, which affects internal validity [31, 35]. We therefore conducted two studies, wherein participants rated different semantic roles and had to rate 32 and 20 statements, respectively, rather than a single study where they had to rate more than 45 statements at one time. In our risk perception studies, we randomized the order of vignettes and the order of questions in each vignette to mitigate confounding effects. We conducted the privacy risk surveys using statements constructed by adding and removing different semantic roles to a baseline statement with the same subject, action, and object. Even though these statements were grammatically correct, they sometimes lacked coherence due to missing contextual information. For example, the statement, "We transfer your personal information, if you are an executive member," is grammatically correct, however, it lacks context to understand executive membership. We limited the context, because additional context can become a confounding factor and affect the risk perception measurements.

The extent to which we can generalize results refers to external validity [37]. In this study we analyzed 15 privacy policies across three domains. We reached saturation in semantic roles after we analyzed the first two privacy policies, Barnes and Noble and Costco. Barnes and Noble policy contained fourteen out of the 17 total semantic roles we identified across all 15 policies, and Costco contained three additional semantic roles (instrument, retention location, retention property) not present in Barnes and Noble policy. We did not identify any new semantic roles in the other thirteen policies. Policies not in our dataset and in different domains could contain new semantic roles and syntactic patterns that we did not observe. Similarly, requirements from other domains could contain additional semantic roles. We believe that the list of semantic roles, their categorization and the list of syntactic patterns that we discovered is only complete for our dataset, whereas new policies or requirements documents could require additional analysis. For our risk surveys, our target population is the average US Internet user. As compared to the 2015 PEW Internet and American Life Survey data of US Internet users, the participants that we recruited from Amazon Mechanical Turk had less reported Asian, Black and Hispanic participants [28]. In our risk surveys, 58–80% of the participants reported their ethnicity as White. Privacy risk perceptions that are affected by ethnicity might therefore be skewed in our study.

6 Discussion

In this paper, we manually annotated and analyzed fifteen privacy policies across health, news and shopping domains to identify the different semantic roles and their values attached to the four different categories of data actions: collection, retention, use and transfer. From a total 698 instances of data actions, we identified 17 unique semantic roles which occur 2316 times. The health policies were the most descriptive of the three domains, with 293 actions and 1024 semantic roles, followed by shopping with 281 actions and 878 semantic roles. And the news policies were least descriptive with 124 data actions and 414 semantic roles instances across all actions.

The expected roles for the four categories of data action were subject, information, condition and purpose. In addition, collection actions frequently have the source role to indicate *from where* the information was collected, and transfer actions have the target role to indicate *to where* the information was transferred. Missing values for these roles in a data practice statement leads to incompleteness in the data practice description and thus become a source of ambiguity. From our analysis, we observe that all the three domains had similar distribution of semantic roles. The health policy actions had the most subjects (85%) and conditions (41%) attached, whereas the news policies had the most purposes (42%) attached. In our dataset, on average 25% of the retention statements across all three domains were incomplete with respect to the subject role. In addition, 55% of transfer statements were incomplete with respect to the condition role, and 20% of usage statements were incomplete with respect to the purpose role. Most of the actions across all domains were performed by first-party companies, followed by third-party companies. We also observed that the most frequent source of user information were the users themselves in health (56%) and news (77%) policies, whereas for shopping policies the source was equally likely to be user (36%) and third parties (39%). When the information is being sourced from third parties, the user might not be aware of the actions being performed on the user's information and thus feel at greater risk. This was also evident from our risk study, wherein participants perceived greater privacy risk when the information was being collected from them, as compared to the information being collected from third parties (see Table 14).

We believe that transfer actions have the most number of semantic role values because transfer actions frequently describe a wide breadth of possible future events, as compared to data actions collection, retention and use, which typically describe events that are more well defined. As such, transfer actions are modified using conditions almost 100% more often than any other data action. The presence of a condition provides a specific property which aims to define the scope of the transfer action. The target semantic role further defines the nature of the transfer. It defines the receiver of the transfer, which is why target is almost exclusive to the transfer action. Unsurprisingly, target is present in over half (141/227) of the total occurrences of the transfer actions because explicitly stating the target is necessary in avoiding ambiguity when describing a physical transfer of data.

Similar to the transfer actions, the use actions displayed the second largest number of semantic roles. We attribute this to the nature of the action. More specifically, when a company describes how they intend to use user data, we can expect to find answers to a number of initial questions. After all, if users learn that they do not agree with what information is being used (object), who is using the information (subject), and the reason for using the information (purpose), they can decide to opt-out of any relationship. In the event that this happens, the company will not even get the opportunity to collect, retain, or transfer user information, and it is this reason that we believe use actions present a large number of semantic role values.

In addition to analyzing data actions and their associations with semantic roles, we also consider the effect of domain on semantic roles. We found that the health domain had the most semantic roles, likely due to the fact that the nature of the data they handle is extremely sensitive. Additionally, we believe that health information consists of more unique objects than the other domains. For each type of information, there may be multiple associated semantic roles. Conversely, shopping and news domains handle arguably less variations of data, which may result in companies feeling less obliged to repeat the roles associated with credit card information, for example, which users may be required to provide.

We found that the shopping domain had proportionally more instances of the object semantic role (280/878) as compared to health (293/1042) and news (124/414). This could be because while the information that a shopping website collects is typically considered common knowledge, e.g., credit card information, shopping websites have the opportunity to retain, use and transfer this information. To promote transparency in their data practices, shopping websites might make a concerted effort to restate the information, namely the semantic role object, that is being not only collected, but also retained, used or transferred.

In our analysis, we identified a total of 74 unique lexical and syntactic patterns that occurred a total of 1119 times in our dataset and can be used to specify semantic roles. We also observed that multiple lexical and syntactic patterns can be used to specify the same semantic role, for example the `if [value]` and `depending on [value]` pattern, among other such patterns, can be used to specify the condition semantic role. In other instances, we found that the same pattern can be used to specify different semantic roles, for example, the pattern `to [value]` can be used to specify the purpose, target, object and constraint roles. We also observed that in some cases, the semantic role specified by a pattern can be predicted from the action category it occurs with. For instance, in the shopping policies, the pattern `to [value]` specifies a data purpose in 98.3% of instances when attached to a usage action and specifies a target in 86.1% of instances when attached to a transfer action. Other patterns, such as `if [value]` and `when [value]`, are used to specify a condition, irrespective of the action category to which they are attached. It was also interesting to note that same patterns were used frequently across all three domains to specify the semantic roles. For example, the most frequent patterns used to specify condition semantic role across all three domains were `if [value]` and `when [value]`, and the patterns `to [value]` and `with [value]` were used to specify the target roles. Finally, we used preposition categorization to analyze the 74 unique patterns and observe the relationship between the syntax-based category and a pattern's observed semantic role value.

We conducted three studies to measure the effect of semantic roles and role values by category on perceived privacy risk. From these studies, we observe that describing the purpose for which the user's data will be acted upon considerably increases the user's willingness to share their information. Similarly, specifying that the user's data will be acted upon only under the condition that the user has consented, increases the willingness to share information. In Study 1, adding the source role with the value "from you" decreased the user's willingness to share their information. In this survey, there was no other value of the *source* role. One explanation may be that participants assume that the source suggests the collected information is more sensitive or personal, or that it is collected automatically without user consent. In Study 2, we observed that adding the condition role, which concerns seeking consent from the user before their data is acted upon, considerably increases the participant's willingness to share their information. In Study 2 we also saw an increase in participant's willingness to share their information when the source was added to the baseline statement, as compared to Study 1 where the condition was "when you create an account with us." The participants see multiple statements on the same page in the survey which includes the statements with conditions. The condition value in Study 2 "with your consent" could have primed participants to think about the other statements more positively.

In Study 3, we observe that participant's willingness to share increases when the information is collected from the user directly, as compared to when the information is collected from third parties, or when the source of the information is vague. Participants were also shown multiple sources from which their information could be collected, including from their devices, third parties and instances where the source role value is vague. These additional sources may have implied that "from you" excludes automated sources in which participants would not be directly involved in the collection process, in other words, there was an anchoring effect. By comparing the sources from which their information is collected, the users may have felt that they have more control over their information, when they directly provide it to the website, as compared to information about them that can be collected by the website from other sources outside their control. Participants were most willing to share their information when they consented to the transfer, or when the transfer was required by law. In addition, participants perceived the least risk when the information was being transferred to the first-party company, compared to other targets.

7 Future work and conclusions

7.1 Future work

In future work, we envision using the annotation technique and the findings from this paper to build a corpus of semantic frames for data practices, and then studying ways to develop an automatic role labeling system for privacy policies. The semantic roles that we identified in this study can be used as a starting point to annotate roles in other privacy policies, and to determine when a role dataset reaches saturation. In addition, we believe that the lexical and syntactic patterns that we identified in this paper can be used as features to automate role labeling.

In the sections below, we present additional recommendations for future work in light of the observations we made during this study.

7.1.1 Multistatement analysis

In our study each frame is constructed from a single privacy policy statement. However, information about the same information type can be elicited from multiple policy statements. This elicitation can help us construct a more complete representation of the actions being performed on that information type. To address this issue, our current frames can be extended to construct a network of frames which can then be used for analysis. For example, consider the two statements "Depending on your interaction with our website, we collect your personal information." and "We use your personal information to provide your personalized ads." The frames for these two statements can be linked together since they are describing actions performed on the same object "personal information." During this process, we elicited different semantic role values attached to different data actions to answer more questions regarding the information type, than could be answered using a single statement.

7.1.2 Permitted actions

In our dataset, we observed multiple statements use a permission verb on the action of interest. For example, consider the following statement from the Lowes privacy policy: "We reserve the right to transfer personal information …" In this statement the main verb is "reserve" which constraints the action of interest "transfer." In our analysis we annotate "transfer" as the verb of interest and annotate semantic roles that are associated with the action "transfer." Other permission verbs in our dataset include: authorized to, allow, intend to, require to, consent to, ask to, obligated to, right to and need to among other such actions. Notably, Breaux et al. [12] found similar phrases, some of which they refer to as delegations, because an authority was assumed to confer the right or obligation onto another party. In the policies we studied, some of the situations the policy authors are referring to are: third parties are allowed by first party or user, third parties are required by first party, first party is required by law, and third parties or first party has the right to or need to perform a data action. Two potential research questions are: (a) how does the use of permission verbs change the semantics of a data practice statement? (b) Under what conditions do policy authors prefer to use such actions along with the collection, retention, use and transfer data actions?

7.1.3 Annotation levels

In this study, we annotate only the semantic roles that are associated with the main action of interest, which is also the root verb in the policy statement. We define this annotation as the first level of annotation. Some of the semantic role values can further contain one or more actions and the associated semantic roles. Annotating the actions and their semantic roles that are part of the semantic role value of the main verb is the second level of annotation. Consider the following statement from MyFitnessPal policy:

```
[{If […]}, [MyFitnessPal] will use
[Your login/ID]

{to [access and collect the informa-
tion …]}]
```

In this statement, the main verb is "use," and we annotate semantic roles that are associated with this action. For example, the purpose of using the user's login/ID is specified using the `to [value]` pattern, "`to [access and collect the information…]`." We call this *first-level* annotation. The purpose value in the pattern `to [value]` is "access and collect…," which also has data-related actions that are of interest to us. For example, the action "access" is a use action and "collect" is a collect action. In this analysis, we limit our annotations to the first level and do not annotate "access" and "collect" actions, which we call second-level actions, or their semantic roles. However, our approach could be extended to annotate actions and semantic roles for more than one level.

7.1.4 Implied semantic role values

We observed in our analysis that some of the missing semantic roles values could be inferred from the statement context. For example, consider the following statement from the 23andMe policy: "If you allow sharing, Genetic and Self-Reported Information may be displayed in other users' accounts." The action of interest in this statement is "display" and the subject for this action is missing. It can be inferred from the statement that the information may be displayed by the service, referring to the first party in this case. Similarly, consider this statement from the WebMD policy: "If you sign up for one of our newsletters on a third-party site, the site will provide us with…" In this statement the condition "if you sign up for one of our newsletters on a third-party site" has an implied purpose which is that the user information in being used for purposes of signing up and consequently for receiving the newsletter. In this study, we do not identify implied semantic roles. However, the semantic role values we identify can be analyzed in second cycle coding to identify implied role values. When developing an automated system to analyze policies, it becomes important to determine what semantic roles values can be implied from the information present in the statement. One way to determine the implied semantic role values is by using textual entailment.

7.2 Conclusion

In this paper we present a semantic frame-based representation for privacy statements that can be used to identify incompleteness in data action context. In this study we analyzed 15 policies from three domains and identified 17 unique semantic roles that are associated with the four categories of data action: collection, retention, usage and transfer. We also categorized the semantic role values of *condition*, *source*, *subject* and *target* roles. In addition, we conducted privacy risk surveys to measure how presence or absence of these semantic roles affects privacy risk. We observed from our survey that specifying the *condition* and *purpose* roles decreases user's perception of privacy risk up to 20%.

The software being developed should meet all the different types of requirements. Some of these requirements concern the privacy of the users and can be extracted from privacy policies. Our semantic frames approach can be used in such scenarios by requirements engineers to automatically extract privacy requirements from the privacy policies of the company and align it with the software's design and also with government regulations. Question-answering systems can also be developed that use these semantic frames as intermediate representation. The questions answering systems could be used by requirements engineers or software designers to query the privacy policy. Example questions can include "What personal information can be shared with third parties?" "When is the browsing history logged?" among other such questions. In addition, software designers can use the findings from our risk surveys to identify high-risk configurations of the software system and take steps to ensure that user's privacy is protected consequently leading to a less risky situation. These findings can also be used to identify default settings that are privacy preserving for the software.

Acknowledgements We thank the CMU RE Lab for their helpful feedback. This research was funded by NSF Frontier Award #1330596 and NSF CAREER Award #1453139.

Appendix A: extracted semantic roles

We identified 17 total semantic roles in our analysis, six of which are described in Sect. 3.2. The remaining roles are as follows:

- *Action location* The location where the action is performed.
- *Comparison* Comparison of the action with other action(s).
- *Constraint* The restrictions on the action.
- *Duration* The duration for which the action will be performed.
- *Exception* Describes an exception to the action.
- *Retention property* This role describes how the information is retained. Example role value from Costco policy: *separately from other member databases*.
- *Hypernymy* A more generic semantic role value with specific values.
- *Instrument* The medium with which the action is performed.
- *Negation* The presence of this role signals that the action will not be performed.
- *Retention location* The location at which the object of the retention action is retained.
- *Time of action* The time at which the action is performed.

Appendix B: semantic roles frequency

The following table presents statistics, including the total number of data actions identified in each data action category (Total Actions); the number of role value instances for the most frequent roles and the total number of roles attached to each data actions category (Total Roles), for each policy (Tables 15, 16 and 17).

Table 15 Frequency of semantic roles across health policies

Policy	Category	Total actions	Subject	Object	Condition	Purpose	Total roles
HealthVault	C	7	5	7	2	3	23
	R	9	8	9	4	0	28
	U	14	13	14	4	11	48
	T	9	8	9	4	3	35
Mayo Clinic	C	1	0	1	0	1	4
	R	1	0	1	0	1	2
	U	17	16	17	11	11	64
	T	40	34	40	20	14	137
MyFitness	C	6	6	6	2	3	21
	R	0	0	0	0	0	0
	U	13	10	13	2	12	13
	T	19	18	19	7	8	19
WebMD	C	14	14	14	5	3	52
	R	8	8	8	3	2	26
	U	21	19	21	4	16	80
	T	15	13	15	6	2	57
23andMe	C	19	15	19	8	3	65
	R	15	12	15	8	4	47
	U	40	28	40	20	29	126
	T	25	21	25	8	4	89
Total		293	248	293	118	130	1024

C Collection, *R* retention, *U* usage, *T* transfer

Table 16 Frequency of semantic roles across news policies

Policy	Category	Total actions	Subject	Object	Condition	Purpose	Total roles
ABC News	C	5	5	5	3	1	16
	R	1	1	1	0	0	2
	U	2	2	2	1	2	7
	T	6	6	6	1	3	22
Bloomberg	C	2	2	2	0	0	7
	R	2	1	2	0	0	5
	U	9	6	9	0	9	24
	T	4	4	4	0	0	14
CNN	C	5	5	5	1	2	17
	R	0	0	0	0	0	0
	U	18	13	18	4	16	57
	T	6	5	6	3	2	20
Fox News	C	7	7	7	5	0	22
	R	7	5	7	2	4	24
	U	12	10	12	4	9	43
	T	9	8	9	3	2	33
Washpost	C	11	10	11	4	4	43
	R	1	1	1	1	0	4
	U	10	6	10	0	6	27
	T	7	5	7	5	1	27
Total		124	102	124	37	61	414

C Collection, *R* retention, *U* usage, *T* transfer

Table 17 Frequency of semantic roles across shopping policies

Policy	Category	Total actions	Subject	Object	Condition	Purpose	Total roles
Barnes and Noble	C	30	29	30	16	6	89
	R	7	6	7	4	3	24
	U	22	20	22	4	17	69
	T	24	18	24	12	1	76
Costco	C	16	13	16	4	2	38
	R	4	1	4	0	0	10
	U	16	14	16	5	12	49
	T	28	24	27	20	4	97
JC Penny	C	20	19	20	9	2	69
	R	1	1	1	0	0	2
	U	19	13	19	0	17	51
	T	12	10	12	4	3	40
Lowes	C	14	14	14	3	2	52
	R	5	3	5	2	2	13
	U	12	10	12	0	10	34
	T	15	14	15	10	2	52
Overstock	C	10	10	10	4	2	32
	R	2	2	2	1	0	6
	U	16	16	16	1	13	46
	T	8	8	8	3	0	29
Total		281	245	280	102	98	878

C Collection, *R* retention, *U* usage, *T* transfer

Appendix C: lexical and syntactic pattern

The following table presents all the unique lexical and syntactic patterns we discovered in our dataset (Table 18).

Table 18 All lexical and syntactic patterns discovered

according to [value]	consistent with [value]	if and when [value]	like [value]	so [value]
as [value]	depending on [value]	including [value]	on [value]	then [value]
as part of [value]	due to [value]	in [value]	once[value]	to [value]
as long as [value]	during [value]	in accordance with [value]	only if [value]	that [value]
as well as [value]	each time [value]	in conjunction with [value]	only as [value]	through [value]
across [value]	even if [value]	in connection with [value]	only with [value]	when [value]
along with [value]	examples of [value]	in a way that can be [value]	only on [value]	with [value]
allowing [value]	except as [value]	in any way [value]	over [value]	where [value]
among [value]	except as noted below [value]	in an effort to [value]	outside [value]	within [value]
at [value]	except [value]	in the course of [value]	prior to [value]	without [value]
based on [value]	for [value]	in order to [value]	provided that [value]	unless [value]
before [value]	for example [value]	in addition to [value]	separately from [value]	upon [value]
between [value]	from [value]	in a manner similar to [value]	subject to [value]	until [value]
beyond [value]	if [value]	in the good faith belief that [value]	such as [value]	via [value]
by [value]	if and only if [value]	in the event that [value]	so that [value]	

References

1. Aarts B (2011) Oxford modern english grammar. Oxford University Press, Oxford
2. Acquisti A, Grossklags J (2012) An online survey experiment on ambiguity and privacy. Commun Strateg 88(4):19–39
3. Acquisti A, Gritzalis S, Lambrinoudakis C, di Vimercati S (2007) Digital privacy: theory, technologies, and practices. CRC Press, Boca Raton
4. Antón AI, Earp JB (2004) A requirements taxonomy for reducing web site privacy vulnerabilities. Requir Eng J 9(3):169–185
5. Baker CF, Fillmore CJ, Lowe JB (1998) The Berkeley FrameNet project. In: Proceedings of the 36th annual meeting of the association for computational linguistics and 17th international conference on computational linguistics—volume 1 (ACL'98), vol 1. Association for Computational Linguistics, Stroudsburg, pp 86–90
6. Bellman S, Johnson EJ, Kobrin SJ, Lohse GL (2004) International differences in information privacy concerns: a global survey of consumers. Inf Soc 20(5):313–324
7. Bhatia J, Breaux TD, Reidenberg JR, Norton TB (2016) A theory of vagueness and privacy risk perception. In: IEEE 24th international requirements engineering conference (RE'16), Beijing, China, 2016
8. Bhatia J, Breaux TD (2017) A data purpose case study of privacy policies. In: 25th IEEE international requirements engineering conference, RE: Next! Track, Lisbon, Portugal, 2017
9. Bhatia J, Breaux T (2018a) Semantic incompleteness in privacy policy goals. In: 2018 IEEE 26th international requirements engineering conference (RE), Banff, AB, Canada, 2018, pp 159–169. https://doi.org/10.1109/re.2018.00025
10. Bhatia J, Breaux T (2018) Empirical measurement of perceived privacy risk. ACM Trans Hum Comput Interact (TOCHI) 25(6):34
11. Breaux TD, Antón AI (2007) Impalpable constraints: framing requirements for formal methods. Technical report technical report TR-2006-06, Department of Computer Science, North Carolina State University, Raleigh, North Carolina, February 2007
12. Breaux TD, Vail MW, Antón AI (2006) Towards compliance: extracting rights and obligations to align requirements with regulations. In: Proceedings of IEEE 14th international requirements engineering conference (RE'06), Minneapolis, Minnesota, pp 49–58
13. Clark LA, Watson D (1995) Constructing validity: basic issues in objective scale development. Psychol Assess 7(3):309–319
14. Dalpiaz F, van der Schalk I, Lucassen G (2018) Pinpointing ambiguity and incompleteness in requirements engineering via information visualization and NLP. In: Requirements engineering: foundation for software quality 2018, pp 119–135
15. Das D, Chen D, Martins AFT, Schneider N, Smith NA (2014) Frame-semantic parsing. Comput Linguist 40:1
16. de Salvo Braz R, Girju R, Punyakanok V, Roth D, Sammons M (2005) An inference model for semantic entailment in natural language. In: National conference on artificial intelligence (AAAI), pp 1678–1679
17. Fernández DM, Wagner S (2015) Naming the pain in requirements engineering: a design for a global family of surveys and first results from Germany. Inf Softw Technol 57:616–643
18. Fikes RE, Kehler T (1985) The role of frame-based representation in knowledge representation and reasoning. Commun ACM 28(9):904–920
19. Fischhoff B, Slovic P, Lichtenstein S, Read S, Combs B (1978) How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sci 9:127–152
20. Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge
21. Gruber JS (1965) Studies in lexical relations. Ph.D. thesis, MIT
22. Fillmore CJ (1976) Frame semantics and the nature of language. Ann N Y Acad Sci 280:20–32
23. Jurafsky D, Martin JH (2000) Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall PTR, Upper Saddle River
24. Kaisser M, Webber B (2007) Question answering based on semantic roles. In: Proceedings of the workshop on deep linguistic processing (DeepLP'07). Association for Computational Linguistics, Stroudsburg, PA, USA, pp 41–48
25. Knijnenburg B, Kobsa A (2014) Increasing sharing tendency without reducing satisfaction: finding the best privacy-settings user interface for social networks. In: 35th international conference on information systems, pp 1–21
26. Massey A, Rutledge RL, Antón AI, Swire PP (2014) Identifying and classifying ambiguity for regulatory requirements. In: 22nd IEEE international requirement engineering conference, pp 83–92
27. Minsky M (1981) A framework for representing knowledge. In: Haugeland J (ed) Mind design. MIT Press, Cambridge
28. Perrin A, Duggan M (2015) Americans' internet access: 2000–2015. In: PEW internet and American life project, June 26, 2015
29. Roth M, Lapata M (2015) Context-aware frame-semantic role labeling. Trans Assoc Comput Linguist 3:449–460
30. Saldaña J (2012) The coding manual for qualitative researchers. SAGE Publications, Thousand Oaks
31. Shadish WR, Cook TD, Campbell DT (2002) Experimental and quasi-experimental designs for generalized causal inference. Houghton, Mifflin and Company, Boston
32. Surdeanu M, Harabagiu S, Williams J, Aarseth P (2003) Using predicate-argument structures for information extraction. In: Proceedings of 41st annual meeting on association for computational linguistics—volume 1 (ACL'03), vol 1. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 8–15
33. Tsai JY, Egelman S, Cranor L, Acquisti A (2011) The effect of online privacy information on purchasing behavior: an experimental study. Inf Syst Res 22(2):254–268
34. Wakslak C, Trope Y (2009) The effect of construal level on subjective probability estimates. Psychol Sci 20(1):52–58
35. Wallander L (2009) 25 years of factorial surveys in sociology: a review. Soc Sci Res 38(3):505–520
36. Wang Y (2015) Semantic information extraction for software requirements using semantic role labeling. In: 2015 IEEE international conference on progress in informatics and computing (PIC), Nanjing, 2015, pp 332–337
37. Yin RK (2013) Case study research: design and methods, 5th edn. Sage Publication, Cambridge