Privacy Dictionary: A Linguistic Taxonomy of Privacy for Content Analysis

Alastair J. Gill±, Asimina Vasalou∗, Chrysanthi Papoutsi†, Adam Joinson∗

±Department of Sociology
University of Surrey
Guildford GU2 7EX UK

∗School of Management
University of Bath
Bath BA2 7AY UK

†Oxford Internet Institute
University of Oxford
Oxford OX1 3JS

Corresponding author: minav@luminainteractive.com

ABSTRACT

Privacy is frequently a key concern relating to technology and central to HCI research, yet it is notoriously difficult to study in a naturalistic way. In this paper we describe and evaluate a dictionary of privacy designed for content analysis, derived using prototype theory and informed by traditional theoretical approaches to privacy. We evaluate our dictionary categories alongside privacy-related categories from an existing content analysis tool, LIWC, using verbal discussions of privacy issues from a variety of technology and non-technology contexts. We find that our privacy dictionary is better able to distinguish between privacy and non-privacy language, and is less context-dependent than LIWC. However, the more general LIWC categories are able to describe a greater amount of variation in our data. We discuss possible improvements to the privacy dictionary and note future work.

Author Keywords

Privacy dictionary, privacy, language, content analysis.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Theory.

INTRODUCTION

While the use of ICTs (Information and Communication Technologies) is recognized to play a beneficial role in our lives, its increasing ubiquity has been met with some concern. One area that is persistently discussed in the field of HCI is that of privacy [16]. Users' privacy perceptions in relation to technology use have become a central question in a variety of contexts such as location tracking in families [35], social network use amongst friends [17, 41] and smart homes for the elderly [4].

A number of methodologies have been used to study privacy. Some researchers probe users to deconstruct prior violations [e.g. 2, 35], an approach that lends itself particularly well to contexts with salient privacy problems. While this method helps identify the source of the offence and also reveals participants' judgments, its limitation is that it does not capture natural, moment-to-moment privacy practices. A second approach builds on the hypothesis that privacy attitudes and concerns motivate privacy-related behaviors [1]. Critical problems however have been noted regarding this approach. First, the reliability of the methods used to measure privacy concerns have been called into question. The leading items included in most attitudinal questionnaires bias participants' responses [14], often resulting in inflated self-reports of privacy concerns that rarely translate to privacy protective behavior [1]. Moreover, when such questionnaires are used in experimental settings, they can prime behaviors. For example, one study found that participants avoided answering sensitive questions after completing a privacy concern measure [18]. These limitations have motivated privacy researchers to contrive methods that will capture nuanced, inclusive and unbiased portrayals of users' concerns, needs and practices. Against this objective, privacy is gauged through neutral questions framed within its wider context [e.g. 6, 24]. Nonetheless, in the absence of a question that can anchor participants' language around the concept of privacy, any subsequent coding and interpretive analyses can be highly subjective while nuanced privacy-related language may end up being ignored in favor of more easily coded themes [29].

In light of these challenges, the development of novel methodologies for the unbiased analysis of privacy have been described as critical in advancing this field [29]. Privacy is frequently studied using qualitative methods whose unit of analysis is language [e.g. 24, 2, 35]. Indeed, natural language is both a reflection and a mediator of internal states, such as personality and emotion, and social situations [32]; words reveal attention patterns, thoughts, feelings, and provide a way of understanding our social worlds [7, 42]. Previous research has developed and used a variety of automated content analysis techniques for the systematic measurement of language, the relevance of which is best understood through comparison to psychometric measures, whose use is established in the field of HCI. Whilst in these latter cases, individual questions are the observed items whose submission to statistical procedures, such as factor analysis, informs the

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

researcher about unobserved latent variables, in content analysis, words and phrases become the observed variables [22]. Some automated content analysis methods count occurrences of words in texts from a set of predefined categories (e.g. LIWC; Linguistic Inquiry and Word Count), while others adopt more sophisticated techniques from computational linguistics (e.g., Coh-Metrix) (for a review see [25]).

The present paper builds on the former approach by contributing a new set of privacy-related categories that can be used with an existing, widely used content analysis program (LIWC; [33]). These new privacy categories constitute a "dictionary" as they encompass a number of words relevant in the semantic analysis of the privacy domain. The dictionary aims to assist social scientists and designers to study longstanding questions in this field while generating comparable results across a large number of different datasets (e.g. interviews, focus groups, open-ended questions) without encountering difficulties relating to different interpretations of coding schemes [25]. Social scientists aiming to bridge the gap between privacy attitudes and actual behavior, may want to use the linguistic features encapsulated by the categories in our dictionary to reveal subtle attitudinal and behavioral differences over time or across contexts as well as relate these to individual characteristics such as gender, personality and social class. Designers concerned that a new version of software pushes the boundaries of privacy may use the dictionary to analyze users' opinions about the system in order to determine the precise privacy issues that users face.

This paper is organized as follows; first, we discuss definitions of privacy developed for both technology and non-technology environments. In ensuring that the dictionary is applicable across contexts, we base it on a wide and context-inclusive definition. We go on to describe the step-by-step procedure we followed to construct and develop the privacy dictionary. To determine whether our new privacy categories are capturing unique aspects of privacy language, not tapped upon by previous tools, we also include in our analysis of texts existing categories from the LIWC dictionary. Next, we present the dataset of interviews and focus groups from seven privacy-sensitive contexts against which we evaluated the dictionary. We then outline the analysis and evaluation of the dictionary. Our main finding is that the privacy dictionary captures unique linguistic features in privacy language. By contrast, LIWC categories are found to measure general contextual differences, but do not reliably distinguish between privacy oriented and non-privacy oriented language. In the final section we discuss future plans for improving the reliability and validity of the dictionary.

BACKGROUND

Privacy Theories

The privacy domain covers the entire range of human activity, spanning from how relationships are negotiated within families, to the ways we manage our transactions over the Internet. It is thus not surprising that theoretical and empirical privacy research has taken place within and across a wide range of disciplines. This interdisciplinary work has converged on a number of descriptive features: privacy is achieved behaviorally through actions of control [9, 2, 34, 3, 43], it serves a number of positive psychological functions [31, 45, 3, 43] and it is governed by social norms negotiated through our interactions with others [2, 34].

At the same time, however, research has shown that particular privacy features gain importance in relation to the context under investigation. To give a few examples, whilst privacy entails the selective control over the physical (e.g. one's sensory presence) realm, depending on the interactional context, it can also involve informational (e.g. personal information) or expressive (e.g. one's opinions and values) control [9]. Whereas the static behavioral patterns fostered in some environments have led scholars to propose that privacy is achieved through the withdrawal of a person in a state of solitude or isolation, privacy can also follow from small group intimacy and when among large groups, from a condition of anonymity or reserve [31, 46].

HCI researchers have explored the notion of context by identifying the features of privacy that become most salient as a result of technological affordances [2, 5, 15, 16, 28]. In one such analysis, Palen and Dourish [28] argue that participation in technology requires some degree of disclosure. The dialectics of privacy form a need for balancing what is shared and kept to oneself. The same authors note that technology-mediated information sharing is less amenable to the user's direct control [see also 2, 16], while technologies reconfigure the temporal nature of identity by framing the present against past and future actions shaped, constrained or recorded by technologies [see also 16]. When technological and physical environments intersect, the meaning of context can take further new forms. This fusion activates new privacy features, bringing different sets of norms in conflict to ultimately shape how technology is perceived and used [28].

The discussion so far should illustrate that despite theorists' agreement over some of its features, context inevitably determines much of the way we conceptualize and study privacy. The lack of a unifying theoretical account of privacy creates a challenge for our project: to build a dictionary that is sensitive to the semantics of privacy, our underpinning theoretical framework must encompass a wide and comprehensive set of privacy features. We turn to theories of categorization that help explain the multivariate nature of privacy and go on to propose a feature-based definition that will motivate our work.

Prototypes and the Role of Context

The classic approach to concept definition identifies sufficient and inclusive criteria (e.g. *control* over *information*), and any instance described by these criteria

will be a member of the concept [37]. However, many natural language categories do not share a common set of defining features. For example, card-games, board games and playing tennis bear a 'family resemblance' structure. Members characterized by more features of the family are better exemplars, thus making membership a matter of degree [47]. Prototype theory evolved from this perspective to propose that concepts, such as privacy, are organized through prototypes that represent the average member of a concept. When new situations are encountered, we evaluate their similarity against the prototype to determine whether they belong to the concept and whether they are good or poor exemplars [37]. Context, in particular, has been shown to shift the prototype (and its features) rendering previously good exemplars into poor exemplars [19]. For instance, when outdoors, the word 'games' is more likely to be interpreted as 'sports'. At a Halloween party, however, 'role-playing', generally regarded to be a poor exemplar of games, becomes a more viable interpretation.

Privacy researchers [45, 40] have proposed that the multifaceted nature of privacy can be explained through prototype theory. Recently, this claim was empirically established [45]. When a concept is organized by a prototype, a wide range of features are reported, none of which are shared across all reports [11]. In [45], 146 participants reported an average of 6.6 features, a process that yielded a total of 82 privacy features. It was then determined whether participants could reliably rate the features' importance or centrality with regards to the concept. Once it is shown that features of the concept vary in their degree of centrality, exemplars of the concept can be directly derived from the features [11]. Using a 9-point scale (9-extremely good feature, 1-extremely poor feature), 118 participants were able to reliably rate the privacy features' centrality. In a final step, 62 participants evaluated vignettes that contained either more central or peripheral features of the privacy concept. The vignettes containing more central privacy features were recognized as better exemplars.

Privacy Linguistics

The findings reviewed above [45] reconcile the different views on privacy, while the features reported by participants provide a solid basis for constructing the privacy dictionary. Four reasons motivate this approach:

(1) From a theoretical perspective, the privacy prototype covers the entire gamut of psychological and behavioral components discussed in the literature [45]. Advantageously, a dictionary built on this foundation will not be representative of a single theoretical view. This means that the dictionary could be used to provide a fair test of theory, without the danger of influencing results through theory-based methods.

(2) The privacy features used in building the dictionary directly address the issue of contextual influence. Context is explicitly woven into the features reported, reflecting environmental (e.g. *personal space*), informational (e.g. *having control over one's information*) and expressive concerns (e.g. *concealing embarrassing details*).

(3) During the analysis of qualitative data, researchers may use a different cognitive reference point to privacy than participants. This can lead researchers to overlook important privacy language. By including language that expresses both central and peripheral privacy features it is possible to address the danger of biased coding and interpretation [29].

(4) The finding that privacy is organized by a prototype that affects human perception [45] has implications to how language is understood; human coders use the network of privacy features (perceived through language) as a heuristic for recognizing privacy exemplars. An automated linguistic method built on the same heuristics can be faster than laborious human coding [see 25].

Despite the many motivations driving a prototype approach to privacy linguistics, as noted earlier, content analysis techniques have already been developed, albeit in other research contexts. Using these techniques it has been possible to identify emotional states [12], predict deception [13] and detect differences in personalities [27], to give just a few examples. A widely-used tool has been the LIWC dictionary that contains 2,300 words and word stems in 74 categories, ranging from basic descriptives of the text (e.g. word count) and grammatical categories (e.g. articles and pronouns), through cognitive and affective words (e.g. "Psychological Constructs"), to words describing time, space and motion (e.g. "Relativity"), to more topic related categories referring to people, leisure activities, and physical and metaphysical concerns (e.g. "Personal Concerns"). Alongside, the custom privacy categories we created (described below), we also apply these established methods to compare with our own approach.

To summarize our objectives, this research sets out to understand whether new categories based on the privacy prototype or existing linguistic categories from the LIWC dictionary, reveal specific characteristics within privacy language. This is achieved by *comparing privacy to non-privacy language*. To determine how stable these categories are across different contexts, our analysis also focuses on *the interactions between privacy language and context*. The next section describes the procedure used to construct the dictionary from the 82 privacy features. It then details the choice of existing LIWC dictionaries used.

DICTIONARIES

Privacy Dictionary Construction

The privacy dictionary is developed for use with the 2001 version of the LIWC software, since this is the most widely used dictionary for content analysis [25]. To construct the dictionary, iterative techniques, similar to those applied in the development of the LIWC dictionaries, were used [33]. After collecting a panel of relevant words, groups of judges decided whether they should be included in or excluded

Table 1: Summary of category groupings.

Category name (number of words)	Description	Example dictionary words
NegativePrivacy (77)	antecedents and consequences of negative privacy experiences	*judgmental, troubled, interfere*
Restriction (91)	restrictive and regulatory behaviors for maintaining privacy	*conceal, lock, exclude*
NormsRequisites (38)	norms, beliefs and expectations in relation to achieving privacy	*consent, respect, discrete*
OutcomeState (52)	behavioral states and the outcomes that are served through privacy	*freedom, separation, alone*
OpenVisible (71)	open and public access to people	*post, display, accessible*
PrivateSecret (20)	the 'content' of privacy, i.e., what is considered private	*secret, intimate*, *data*
Intimacy (14)	small group privacy marked by group inclusion and intimacy	*accept, belong, intimacy*
SafetyProtect (25)	feeling safe and protecting or guarding oneself	*guard, protect, safe*

from the dictionary, and how they should be grouped into categories.

A first version of the privacy dictionary was developed on the basis of the prototype feature list discussed earlier [45]. This list included phrases or words generated by 146 participants. Features that were phrases were reduced to single words if possible so as to ensure maximal compatibility with the LIWC software [33]. For example, "having control over one's information" was broken down into two linguistic units: *control* and *information*. Several features had to be omitted, as they were not reducible to single words (e.g. "keeping to oneself"). This process yielded a total of 72 unique words. These revised prototype words were then used as "seed words" over several iterations to generate additional synonyms and antonyms using traditional and computational semantic dictionaries and thesauri. In a first step, two judges evaluated the consistency of the additional synonyms and antonyms with the original words, with consensus between judges determining a word's inclusion or exclusion. This resulted in the selection of 573 dictionary words.

In a second step, 'key word in context' analysis (KWIC) was conducted on the dictionary words on a sample of data (see dataset section). The output of this analysis provided contextual information of the occurrence of the dictionary words. This step was necessary to ensure that the reduction of multi-word prototype features to single words did not capture unintended meanings from those originally envisaged by the judges in Stage 1. Words regarded as inconsistent with the original intended meanings were excluded. For example, the word *company* was intended to capture the state of "having or not having company", but instead the analysis of the context in which this word was used revealed that it was more frequently used to refer to a business organization, and, therefore, it was removed from the privacy dictionary. This process led to 185 words being excluded.

The final stage in the dictionary development was to construct theoretically sound categories of semantically similar words, which would form the basis of the output of the analyses carried out using the privacy dictionary. This is necessary to enable the measurement of consistent and reliable categories that can provide theoretically meaningful results. Two judges, one familiar with privacy theory and the other with linguistics and content analysis, worked together on this task. The eight high-level categories presented in Table 1, are the result of discussion and unilateral consensus between these two researchers.

Linguistic Inquiry and Word Count

As noted earlier, the LIWC dictionary contains a large number of semantic categories with possible relevance to privacy research. Therefore, rather than building new categories from scratch which have substantial overlap with the standard LIWC categories, we adopted 16 of the LIWC categories, which corresponded to the following prototype features: 'Sexual life' (*Sexual* category), 'Body' (*Body, Groom* categories), 'With people you feel close to' (*Family, Friends* categories), 'Personal space' (*Space* category), 'Financial information' (*Money* category), 'Concealing one's emotions' (*Positive Emotions, Negative Emotions* categories), 'Personal' (*Religion* category), 'Involves a group of people and no one else' (*I, you, we, other, Other References* categories), 'At home' (*Home* category).

DATASET

In this section we detail the methodology applied to collect a dataset of one-to-one interviews and focus groups against which we evaluated the dictionary. We chose seven offline and online contexts that previous research suggests are sensitive to privacy issues. These were: (1) criminal offences and imprisonment [26, 30] (2) children and the Internet [23]; (3) financial exclusion; [21] (4) sexuality and

Table 2: Summary of dataset.

Project title	Name	Available participant demographics	Data source	Method	Total words (privacy)
Co-operation or Contest? Inter-Agency Relationships in Police Custody Areas [39]	CRI	5 detainees (4 male) in two police custody areas (Age range: 21-27)	UK Data Archive	Interviews	22,904 (3,509)
United Kingdom Children Go Online [23]	CHI	20 (6 male) secondary school and sixth form students (Age range: 13-17)	UK Data Archive	Focus groups	35,994 (8,638)
Delivering Financial Services in the Home [20]	FIN	3 managers in door-to-door financial services firms	UK Data Archive	Interviews	18,188 (3,552)
Cultural Context of Youth Suicide: Identity, Gender and Sexuality [36]	CUL	5 (2 male) school or college students or in unemployment (Age range: 16-19)	UK Data Archive	Interviews	28,471 (8,455)
Social Network Sites and Identity [6]	SNS	5 (2 male) undergraduate and graduate students (Age range: 21-40)	With permission from primary investigator	Interviews	30,373 (12,066)
Technology and Natural Death: a Study of Older People [38]	QOL	9 (5 male) participants over 65 living in deprived/mid-deprived areas (Age range: 65-84)	UK Data Archive	Focus groups and interviews	62,477 (14,156)
Health and Experiences with Illness	HEA	6 participants (4 Male) with diverse health and illness experiences (Age range: 22-66)	With permission from Health Experiences Research Group (University of Oxford)	Interviews	74,729 (12,249)

self-harm [36]; (5) sharing in social network sites [17, 41]; (6) experiences of elderly people with medical care and dying [8, 44]; (7) health experiences within medical practices [10, 15]. We then searched through the UK Data Archive (hosted on the ESDS site: www.esds.ac.uk) or contacted researchers who had worked on these topics to identify previously collected datasets. The aim was to find qualitative data rich in privacy content, which had been generated by asking questions unrelated to privacy, in order to avoid methodological problems of priming in the responses. Table 2 presents more details of the datasets included in our study. The data included fully abided with participants' informed consent and the institutions' ethics approval procedures.

A team of five researchers who were knowledgeable in privacy theory selected appropriate transcripts using the following procedure. Two researchers worked on each context. The first surveyed the entire panel of transcripts made available in order to identify a maximum of five transcripts per context that involved a diversity of privacy-related issues. Focusing on one transcript at a time, the same assistant identified areas in the text where participants expressed privacy-related issues. These segments were examined by the second assistant who raised any disagreements concerning the inclusion of a given privacy text. Disagreements between coders were resolved through discussion and only privacy texts that yielded bilateral agreement were included. Table 2 summarizes the dataset, which comprised 273,136 words, out of which 62,625 were coded as belonging to the privacy condition and 210,511 to the non-privacy condition.

RESULTS

Privacy dictionary

A series of three-way (Privacy, Context, Speaker) Restricted Maximum Likelihood (REML) mixed effect models were calculated with the privacy dictionary categories as dependent variables. This method was used since it is well suited to data consisting of repeated measures. *Privacy* had two levels (Privacy/Non-Privacy): The Privacy condition included utterances assessed by dataset coders to express privacy issues. The Non-privacy condition contained all other language captured during each interview (see dataset section). *Context* comprised of seven levels, one for each context in the data panel (CRI, CHI, FIN, CUL, SNS, QOL, HEA). To control for the functionally different conversational role of interviewers and interviewees, a third independent variable, *Speaker*, was coded with two levels (Interviewer/Interviewee); we do not consider these speaker results in detail in the current paper. Since this is a repeated measures analysis, each speaker identifier (*SpeakerID*) was included as a random

Table 3: Mean usage of Privacy categories and LIWC categories.

	Privacy		Data Context						
Privacy categories	Privacy	Non-Priv.	CHI	CRI	CUL	FIN	HEA	QOL	SNS
NegativePrivacy	0.16	0.04	0.08	0.10	0.20	0.10	0.10	0.06	0.07
OpenVisible	0.23	0.07	0.07	0.21	0.08	0.10	0.41	0.02	0.16
NormsRequisites	0.13	0.02	0.03	0.11	0.03	0.00	0.12	0.22	0.03
OutcomeState	0.07	0.02	0.03	0.10	0.06	0.03	0.03	0.01	0.05
Restriction	0.12	0.05	0.14	0.09	0.13	0.10	0.03	0.04	0.06
PrivateSecret	0.41	0.02	0.32	0.23	0.08	0.28	0.26	0.05	0.30
Intimacy	0.07	0.01	0.02	0.00	0.05	0.00	0.12	0.03	0.06
SafetyProtect	0.13	0.06	0.17	0.02	0.11	0.09	0.10	0.11	0.08
LIWC categories									
I	4.22	4.36	3.83	6.39	5.49	2.15	3.74	3.77	4.64
We	0.48	0.69	0.53	0.15	0.37	1.44	0.59	0.55	0.48
You	5.78	5.33	6.33	6.03	5.20	3.19	6.70	4.40	7.04
Other	4.32	3.28	2.67	3.94	3.51	5.01	3.37	5.42	2.69
Posemo	2.28	2.22	2.46	2.52	2.83	1.20	2.35	1.82	2.58
Negemo	0.95	0.93	0.70	0.94	1.44	0.69	0.95	1.27	0.57
Othref	11.09	9.64	10.13	10.37	9.61	10.00	11.06	10.61	10.70
Friends	0.26	0.27	0.48	0.04	0.37	0.01	0.07	0.05	0.84
Family	0.31	0.31	0.64	0.03	0.43	0.06	0.35	0.57	0.11
Space	2.22	2.33	1.96	2.03	2.51	2.53	2.55	2.22	2.21
Home	0.52	0.42	0.64	0.30	0.31	0.08	0.74	1.01	0.20
Money	0.22	0.36	0.24	0.22	0.03	1.26	0.10	0.09	0.09
Relig	0.02	0.04	0.00	0.02	0.08	0.00	0.05	0.04	0.02
Body	0.36	0.45	0.28	0.31	0.30	0.12	0.87	0.86	0.06
Sexual	0.15	0.15	0.11	0.02	0.46	0.00	0.35	0.04	0.05
Groom	0.03	0.05	0.00	0.03	0.05	0.08	0.01	0.05	0.04

effect in the model. Table 3 displays the mean differences across conditions for the eight privacy categories.

R^2 values show that the model explained the following variance for each linguistic category: *NormsRequisites* (0.49), *NegativePrivacy* (0.35), *PrivateSecret* (0.19), *OutcomeState* (0.16), *SafetyProtect* (0.16), *Restriction* (0.10), *Intimacy* (0.02), *OpenVisible* (-0.06). The categories *Intimacy* and *OpenVisible* explained little variance within this dataset; indeed the negative R^2 found for *OpenVisible* indicates that the fit curve of this model explains less variance in the data than a straight line placed at the mean. We thus retain only the remaining six categories in the subsequent analyses and revisit the two excluded categories in the discussion.

Words belonging to six linguistic categories of the privacy dictionary were used significantly more frequently in the Privacy than the Non-privacy condition: *PrivateSecret, NegativePrivacy*, *NormsRequisites*, *OutcomeState, Restriction* and *SafetyProtect*. This indicates that our privacy dictionary categories captured differences between privacy and non-privacy language.

Turning now to the role of Context: For the *NormsRequisites* category, we note a significant interaction between Privacy and Context. Tukey post-hoc tests revealed that when talking about privacy issues, participants in the QOL condition used more words from the *NormsRequisites* category compared to the Privacy condition of four other contexts (CHI, FIN, CUL and SNS). We also note a main effect of Context on *NormsRequisites*, with Tukey post-hoc tests showing a similar pattern of Context more generally: *NormsRequisites* category words were used significantly more in the QOL context than in CHI, FIN, CUL and SNS contexts. Taken together, the interaction and main effect for *NormsRequisites* demonstrate that the QOL context had an influence over participants' general language, an effect that was stronger when they spoke about privacy specifically. We therefore note that the category *NormsRequisites* is to some extent sensitive to contextual issues. Table 4 reports the main effects for Privacy, Context and Privacy*Context interactions. We note that findings relating to *Speaker* are not presented in detail here, but instead are left to future work[1].

LIWC dictionary

A second set of three-way REML mixed effect models were calculated, this time with the 16 LIWC categories, deemed

[1] There are Speaker main effects and a Speaker * Privacy interaction for *NormsRequisites* (and Speaker * Context interactions for *NegativePrivacy*, *NormsRequisites*, and *Restriction*).

Table 4: Main effects and interactions.

	Privacy		Privacy*Data Context			Data Context		
Privacy categories	F	*p*	F	*p*	*Post-hoc analysis*	F	*p*	*Post-hoc analysis*
NegativePrivacy	14.84	***	0.86		-	1.50		-
NormsRequisites	12.86	***	3.01	*	P_QOL > (NP_QOL, NP_HEA, NP_CRI, P&NP_FIN, P&NP_CUL, P&NP_SNS, P&NP_CHI)	3.01	*	QOL > (CHI, FIN, CUL, SNS)
OutcomeState	7.89	**	0.89		-	1.24		-
Restriction	6.51	*	2.59	*	P_CUL > NP_CUL	1.31		-
PrivacySecret	28.97	****	1.30		-	1.32		-
SafetyProtect	4.22	*	0.34		-	0.97		-
LIWC categories								
I	0.73		5.77	**	P_CRI > (NP_CRI, P_CUL, P&NP_HEA, P&NP_QOL, P&NP_CHI, P&NP_FIN, P&NP_SNS); NP_CUL > (P_CHI, P_QOL, P&NP_HEA, P&NP_FIN); NP_CRI > (P_CHI, P&NP_FIN); (P_CUL,NP_SNS) > P&NP_FIN	10.79	****	CRI > (SNS, CHI, QOL, HEA, FIN); CUL > (CHI, QOL, HEA, FIN); SNS > FIN
We	9.88	**	1.19		-	5.58	***	FIN > (HEA, QOL, CHI, SNS, CUL, CRI)
You	5.15	*	3.78	**	P_CHI > (NP_CHI, NP_QOL, P&NP_FIN); (P_HEA, NP_CRI, P&NP_SNS) > (NP_QOL, P&NP_FIN); (NP_HEA, P_CRI) > P_FIN	9.44	****	SNS > (CUL, QOL, FIN); (HEA, CHI) > (QOL, FIN); CRI > FIN; CUL > FIN
Other	26.95	****	0.38		-	3.63	**	QOL > (SNS, CHI)
Posemo	0.11		1.62		-	2.06		-
Negemo	0.07		1.25		-	5.92	****	CUL > (CHI, FIN, SNS); QOL > (CHI, SNS)
Othref	23.54	****	1.36		-	0.39		-
Friends	0.12		1.41		-	12.58	****	SNS > (CUL, HEA, QOL, CRI, FIN); CHI > (HEA, QOL, CRI, FIN)
Family	0.00		4.07	**	(NP_CHI, P_QOL) > (NP_QOL,P&NP_HEA, P&NP_SNS, P&NP_FIN, P&NP_CRI); P_CHI > P&NP_CRI; P&NP_CUL > P_CRI	14.15	****	CHI > (HEA, SNS, FIN, CRI); (QOL, CUL) > (SNS, FIN, CRI); HEA > (SNS, FIN, CRI); HEA > CRI
Space	0.32		1.00		-	1.01		-
Home	1.06		4.89	**	P_QOL > (NP_HEA, P_CHI, P&NP_CRI, P&NP_CUL, P&NP_SNS, P&NP_FIN); P_HEA > (P_CUL, NP_CRI, P_SNS, P_FIN); NP_CHI > P_SNS	5.97	****	QOL > (CUL, CRI, SNS, FIN)
Money	31.28	****	11.70	***	NP_FIN > P_FIN > (P&NP_CHI, P&NP_CRI, P&NP_QOL, P&NP_HEA, P&NP_SNS, P&NP_CUL)	24.76	****	FIN > (CHI, CRI, HEA, QOL, SNS, CUL)
Relig	7.10	**	0.45		-	2.67	*	CUL>CHI
Body	2.29		6.84	****	(NP_QOL, NP_HEA) > (P_QOL, P&NP_CHI, P&NP_CRI, P&NP_CUL, P&NP_FIN, P&NP_SNS)	12.05	****	(HEA, QOL) > (CRI, CUL, CHI, FIN, SNS)
Sexual	0.00		1.13		-	6.58	****	CUL > (CHI, SNS, QOL, CRI, FIN); HEA > (SNS, QOL, CRI, FIN)
Groom	1.48		1.92		-	1.88		-

Note: Statistical significance indicated as follows: *=p<.05, **=p<.01, **=p<.001, ****=p<.0001.

relevant to privacy, entered as dependent variables. Again, *SpeakerID* was included as a random effect in each model. The 16 LIWC categories explained a larger amount of variance as shown by the R^2 reported below. Thus, all variables were retained: *I* (0.96), *Money* (0.94), *You* (0.94), *Sexual* (0.92), *Other* (0.88), *Friends* (0.86), *Other references* (0.85), *We* (0.81), *Body* (0.79), *Family* (0.74), *Religion* (0.72), *Positive emotions* (0.69), *Negative emotions* (0.54), *Home* (0.47), *Space* (0.33), and *Groom* (0.22). Table 3 displays the mean differences across conditions for the 16 LIWC categories.

We found main effects of Privacy to be significant for *Other*, *Other references*, and *You* which were used more frequently in the Privacy than the Non-privacy condition. By contrast, words belonging to *Money, Religion* and *We* categories were used more frequently in the Non-privacy than the Privacy condition. Thus, from a total of 16 categories, six revealed differences between privacy and non-privacy language, three of which identified patterns specific to non-privacy discussions.

Findings relevant to Context are the interaction of Privacy and Context for *Money, I, Home*, *Family,* and *You* categories. Main effects were found for 12 of the 16 LIWC variables with the following *not* showing a main effect of context: *Posemo, Other references, Space,* and *Groom*. Given the complex relationships between these linguistic categories and the different contexts (and interactions with Privacy for *Money*, *I*, *Home*, *Family*, and *You*), we do not describe these in detail. Significant differences revealed by Tukey post-hoc tests are found in Table 4 along with the main effects for Privacy, Context and Privacy*Context

interactions[2]. These findings suggest that despite their relevance to privacy, these categories are more sensitive to measuring general contextual features that frame privacy than discriminate specific privacy language. This apparently ubiquitous usage of many LIWC categories is further demonstrated by the significant findings relating to speaker role, which we do not report in any further detail.

DISCUSSION

This research developed and evaluated a privacy dictionary whose objective is to assist researchers in conducting automated content analysis of texts and transcripts, providing a valuable addition to the arsenal of tools available for the study of privacy. The integrative theoretical approach used to build the dictionary can help reduce bias introduced by theory-based methods. It can also provide a shared and common platform that allows meaningful comparisons of cross sectional data. In practical terms, it can precede qualitative thematic coding by pre-identifying language of interest that would otherwise require laborious coding [25].

The dictionary offers new prospects for future theoretical development. The categories can be used to track privacy perceptions over time and changing conditions; for instance, users' privacy perceptions before and after the introduction of mandatory and possibly threatening technologies can be measured through language. Furthermore, some contexts are governed by strong privacy norms, an aspect the privacy dictionary is able to capture, raising stakeholder awareness to prevent privacy threats from escalating into violations. The dictionary can also drive the development of new measures. The cumulative use of dictionary words can serve as a general privacy metric while analyses at the individual word level can inform the development of new psychometric measures by identifying context-specific privacy behaviors, to give one example. A further question of interest for the field of HCI is whether technology users are cognizant of the dangers raised by technologists [e.g. 16, 28]. People's language across comparable technological and non-technological privacy-sensitive contexts may assist in answering this question.

To summarize our findings, of the six privacy categories that were retained for analysis, the variance explained by privacy versus non-privacy language and context ranged from between 10 and 49 percent. Out of these six categories, all exposed context-independent linguistic patterns in privacy language when compared to non-privacy language. These categories are: *NegativePrivacy, NormsRequisites, OutcomeState, SafetyProtect, Restriction* and *PrivateSecrets*. By contrast, the privacy-related LIWC categories included in our analysis explained more variance overall but showed a stronger relationship to context than to privacy. Thus, the greater explanatory power of these categories comes from their relative frequency and consistent use in language.

The privacy categories forming the output of our analysis found that participants within the Privacy condition described more negative privacy experiences, such as, *feeling intruded upon, embarrassed, threatened,* (*NegativePrivacy* category); they made references to the norms and expectations of privacy, e.g., *discretion, respect* (*NormsRequisites* category) and used language that expressed the realms that privacy protects e.g., *data, secrets* (*PrivateSecret* category). Participants detailed concerns about safety and protection e.g., *security, safeguard* (*SafetyProtect* category), as well as discussed the various behavioral states through which they achieve privacy and its outcomes e.g., *alone, quiet* (*OutcomeState* category) or the behaviors applied to manage privacy e.g., *control, hide* (*Restriction* category).

In the LIWC analysis, with the exception of just a few variables, the majority of categories occurred more frequently between contexts, irrespective of whether participants were talking about privacy. Thus, although these categories were considered by us to be directly relevant to privacy, they appear more suitable for capturing general contextual differences. The three categories that captured patterns in privacy language showed that when talking about privacy, people made more references to others and third-person pronouns. These categories reflect the relational and social nature of this construct [34, 28]. However, in answering more specific theoretical questions about privacy, the applicability of these categories seems to be limited. Thus, the remainder of the discussion explores in more detail the privacy dictionary. We consider why some of the categories explained little variance, and discuss ways for the future development and validation of the dictionary.

Despite the possible concern that the category *SafetyProtect* represents a narrow and specific component of privacy (containing only 25 dictionary words), it was nevertheless used relatively frequently in privacy related discourse. A small number of dictionary words in a category may not necessarily be an issue (e.g., in the case of *SafetyProtect*). Indeed, the relatively low frequency of words within the privacy dictionary categories did not impair its function in general. Nonetheless, in the development of the LIWC dictionary categories, that also formed the basis of our approach, safeguards were taken to ensure that words within categories achieved an acceptable frequency of usage (in that case 0.005%). Given the very specialized nature of the words within the privacy categories, we have to be more sympathetic to less-frequently occurring words. One particular problem of using low-frequency words however is that of sparse data which is more likely to lead to skewed distributions. Although this is to be expected, given that content analysis categories both in the privacy dictionary and LIWC often deal with words in the 'long

[2] Speaker main effects for *I, We, You, Other, Posemo, Negemo, Othref, Family*; Speaker * Context interactions effects for *I, We, You, Other references, Family, Money*.

tail' (i.e. those words which form the vast majority of language, but which are used very infrequently), this was a particular problem for *OpenVisible* and *Intimacy* categories. In both cases, examination of the distribution of these categories revealed data sparsity.

Future Work

In future dictionary iterations, the issue of data sparsity due to the inclusion of low frequency items in dictionary categories must be balanced against the requirement to capture a narrow and focused portrayal of privacy. Whilst the inclusion of a greater number of conceptually related words may improve the reliability and power of dictionary categories, further work will also need to test the validity of the privacy dictionary categories. Therefore, we propose further internal and external evaluation of the privacy dictionary. For example, internal validity can be improved by combining information from psychometric measures of privacy and multidimensional semantic space measures from computational linguistics [32, 25]. We also expect to use a much larger privacy dataset collected from a greater variety of contexts and situations, for example, a more inclusive dataset that explicitly recruits group privacy contexts (e.g. online support communities). Finally, the explanatory power of our dictionary could also be improved (based upon e.g., n-grams and probabilistic information) through the inclusion of basic contextual rules enabled in more recent versions of the LIWC software, which could aid word-sense disambiguation and more accurate classification of words within dictionary categories.

SUPPLEMENTARY MATERIAL

The dictionary is available under the Creative Commons license and can be obtained through correspondence with Asimina Vasalou (minav@luminainteractive.com). Information about the dictionary development can be found at www.privacydictionary.info.

ACKNOWLEDGEMENTS

We thank Fadhila Mazanderani, Anne-Marie Oostveen and Sacha Brostoff whose assistance was invaluable with the transcript analysis. We gratefully acknowledge Cristina Soriano and the reviewers of this paper whose comments improved our work. This research was funded by the EPSRC through the Privacy Value Networks project (EP/G002606/1) and partially supported by the European Commission Future and Emerging Technologies programme FP7-COSI-ICT (QLectives project, grant 231200).

REFERENCES

1. Acquisti, A. and Grossklags, J. Privacy attitudes and Privacy behavior: Losses, Gains, and Hyperbolic Discounting. In Camp and Lewis: The Economics of Information Security Kluwer, 2004.
2. Adams, A. and Sasse, M.A. Privacy in multimedia communications: protecting users not just data. In *Proceedings of IMH HCI'01*, 2001, 49-64.
3. Altman, I. The environment and social behavior. Brooks/Cole, Monterey: CA, 1975.
4. Beach, S., Schulz, R., Downs, J., Matthews, J., Barron, B. and Seelman, K. Disability, Age, and Informational Privacy Attitudes in Quality of Life Technology Applications: Results from a National Web Survey. ACM Trans. Access. Comput., 2, 1 2009, 1-21.
5. Boyle, M. and Greenberg, S. The language of privacy: Learning from video media space analysis and design. ACM Trans. Comput.-Hum. Interact., 12, 2 2005), 328-370.
6. Christidi, S. and Rosenbaum-Elliott, R. Shared spaces and personal corners in social networking websites: the contracted and the cryptically revealed self. *European Advances in Consumer Research*, 9 2010).
7. Chung, C. and Pennebaker, J. W. The Psychological Function of Function Words. In: Fiedler (ed.): Social Communication, Psychology Press, 2007.
8. Costello, J. Nursing older dying patients: findings from an ethnographic study of death and dying in elderly care wards. Journal of Advanced Nursing, 35, 1 2001), 59-68.
9. DeCew, J. W. In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Cornell University Press, Ithaca, NY, 1997.
10. DeCew, J. W. The Priority of Privacy for Medical Information. Social Philosophy and Policy, 17, 2 2000, 213-234.
11. Fehr, B. Prototype Analysis of the Concepts of Love and Commitment. *Journal of Personality and Social Psychology*, 55, 4 1988, 557-579.
12. Gill, A. J., French, R. M., Gergle, D. and Oberlander, J. *The Language of Emotion in Short Blog Texts*. ACM, New York, NY, 2008.
13. Hancock, J. T., Curry, L., Goorha, S. and Woodworth, M. T. On lying and being lied to: A linguistic analysis of deception. Discourse Processes, 45 2008, 1-23.
14. Harper, J. and Singleton, S. With a Grain of Salt: What Consumer Privacy Surveys Don't Tell Us. 2001. http://ssrn.com/abstract=299930.
15. Introna, L. D. and Pouloudi, A. Privacy in the information age: Stakeholders, interests and values. *Journal of Business Ethics*, 22, 1 1999, 27-38.
16. Joinson, A. N. and Paine, C. Self-disclosure, Privacy and the Internet. Oxford University Press, Oxford, 2007.
17. Joinson, A. N. Looking at, looking up or keeping up with people? Motives and use of facebook. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (Florence, Italy). ACM, 2008.
18. Joinson, A. N., Paine, C., Buchanan, T. and Reips, U. D. Measuring self-disclosure online: Blurring and non-response to sensitive items in web-based surveys. *Comput Hum Behav*, 24, 5 2008, 2158-2171.

19. Labov, W. The boundaries of words and their meaning. In Bailey and Shuy (eds.): New Ways of Analyzing Variation in English. Georgetown University Press, Washington DC, 1973.
20. Leyshon, A., Knights, D. and Burton, D. Delivering Financial Services in the Home. Colchester, EsCUL: UK Data Archive, 2002-2004.
21. Leyshon, A., Signoretta, P., Knights, D., Alferoff, C. and Burton, D. Walking with Moneylenders: The Ecology of the UK Home-collected Credit Industry. Urban Studies, 43, 1 2006, 161-186.
22. Lowe, W. Content analysis and its place in the (methodological) scheme of things Qualitative Methods 2, 1 2004, 25-27.
23. Livingstone, S. and Bober, M. United Kingdom Children Go Online. Colchester, EsCUL: UK Data Archive, 2003-2005.
24. Mazanderani, F. and Brown, I. Making things private: exploring the relational dynamics of privacy. In *Proceedings of the Computers, Privacy and Data Protection* (Brussels, Belgium), 2010.
25. Mehl, M.R. and Gill, A.J. Computerized Content Analysis. In: Gosling and Johnson (eds.), Advanced Methods for Behavioral Research on the Internet. American Psychological Association Publications, Washington, DC.
26. Murphy, T. and Whitty, N. Risk and Human Rights in UK Prison Governance. British Journal of Criminology, 47, 5 2007, 798-816.
27. Oberlander, J. and Gill, A. J. Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes*, 43, 2 2006, 239-270.
28. Palen, L. and Dourish, P. Unpacking "privacy" for a networked world. In Proceedings of the SIGCHI conference on Human factors in computing systems (Ft. Lauderdale, Florida, USA). ACM, New York, NY, 2003.
29. Patil, S., Romero, N. and Karat, J. *Privacy and HCI: methodologies for studying privacy issues*. In Proceedings of the SIGCHI conference on Human factors in computing systems (Montreal, Canada). ACM, New York, NY, 2006.
30. Pattenden, R. and Skinns, L. Choice, Privacy and Publicly Funded Legal Advice at Police Stations. The Modern Law Review, 73, 3 2010, 349-370.
31. Pedersen, D. M. Model for types of privacy by privacy functions. Journal of Environmental Psychology, 19, 4 1999, 397-405.
32. Pennebaker, J. W., Mehl, M. R. and Niederhoffer, K. G. Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology, 54 2003, 547-577.
33. Pennebaker, J. W., Francis M.E., Booth R.J. (2001). Linguistic Inquiry and Word Count (LIWC): LIWC2001. Mahwah: Lawrence Erlbaum Associates.
34. Petronio, S. Boundaries of privacy: Dialectics of disclosure. State University of New York Press, Albany: NY, 2002.
35. Raento, M. and Oulasvirta, A. Designing for privacy and self-presentation in social awareness. *Personal and Ubiquitous Computing*, 12, 7 2008, 527-542.
36. Roen, K., Scourfield, J. and McDermott, E. Cultural Context of Youth Suicide: Identity, Gender and Sexuality. Colchester, EsCUL: UK Data Archive, 2006.
37. Rosch, E. *Principles of categorization*. Erlbaum, Hillsdale, NJ, 1978.
38. Seymour, J. Technology and Natural Death: a Study of Older People. Colchester, EsCUL: UK Data Archive, 2001-2002.
39. Skinns, L. Co-operation or Contest? Inter-Agency Relationships in Police Custody Areas. Colchester, EsCUL: UK Data Archive, 2007.
40. Solove, D. J. A taxonomy of privacy. University of Pennsylvania Law Review, 154, 3 2006, 477-564.
41. Stutzman, F. and Kramer-Duffield, J. Friends only: examining a privacy-enhancing behavior in facebook. In Proceedings of the 28th international conference on Human factors in computing systems (Atlanta, Georgia, USA). ACM, New York, NY, 2010.
42. Tausczik, Y. R. and Pennebaker, J. W. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology, 29, 1 2010, 24-54.
43. Tavani, H. T. Philosophical theories of privacy: Implications for an adequate online privacy policy. *Metaphilosophy*, 38, 1 2007, 1-22.
44. Terry, W., Olson, L. G., Wilss, L. and Boulton-Lewis, G. Experience of dying: concerns of dying patients and of carers. Internal Medicine Journal, 6, 3 2006, 338-346.
45. Vasalou, A., Joinson, A., Houghton, D. A prototype analysis of privacy (under review).
46. Westin, A. Privacy and freedom. Athenaeum, New York, 1967.
47. Wittgenstein, L. *Philosophical Investigations*. Blackwell, Oxford, 2001.