jcls. Journal of Computational Literary Studies

OPEN ACCESS

Article

Need a Good Book about Privacy?

Evaluating Dictionary-Based Corpus Query for Detecting the Topic of Privacy in Literary Texts

Erik Ketzan[1]
Jennifer Edmond[2]
Carl Vogel[3]

1. Department of Digital Humanities, King's College London, London, United Kingdom.
2. Centre for Digital Humanities, Trinity College Dublin, Dublin, Ireland.
3. School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland.

Citation
Erik Ketzan, Jennifer Edmond, and Carl Vogel (2023). "Need a Good Book about Privacy? Evaluating Dictionary-Based Corpus Query for Detecting the Topic of Privacy in Literary Texts". In: *Journal of Computational Literary Studies* 2 (1).
10.48694/jcls.3602

Date published 2024-03-06
Date accepted 2024-02-02
Date received 2023-01-31

Keywords
computational literary studies, privacy, corpus query

Reviewers
Dennis Mischke, Michał Mrugalski

Note
This paper has been submitted to the conference track of JCLS. It has been peer reviewed and accepted for presentation and discussion at the 2nd Annual Conference of Computational Literary Studies at Würzburg University in June 2023.

Abstract. This paper evaluates the usefulness of querying Vasalou et al.'s *Privacy Dictionary* (2011), a dictionary of 600+ words and phrases, in 131 canonical English-language novels from the long 19th century. We evaluate the word frequencies compared with a classification of the novels based on scholarly attention to the topic of privacy in each particular text. We report evidence of low- to low/medium strength of correlation between 3 of the 8 categories of the Privacy Dictionary and this classification. As a final step, by identifying the novels in our corpus which score highest in relative word frequency in these 3 categories, we suggest novels which have not yet received scholarly study on the topic of privacy but which may be promising for such studies. The highest scoring novel by our method, Maria Edgeworth's *Castle Rackrent* (1800), seems indeed to be highly concerned with the topic of privacy, which is discussed in its author's preface and opening pages.

> "Nothing comes up oftener to-day than the question of the rights of privacy." – Henry James (1900, 63)

1. Introduction

This paper evaluates the usefulness of querying a pre-existing dictionary of words relating to the topic of privacy in a large literary corpus, with the goal of uncovering texts which may benefit from further scholarship on the topic of privacy in literary texts. The aim of this paper is thus neither distant reading (Moretti 2000), nor macroscopic literary inquiry (Underwood 2017), nor the tracing of a crisply defined textual feature, such as one might see in e.g. corpus stylistics (Wynne 2006). Such macro- and micro- DH approaches have been analogized to a telescope and microscope (Eve 2019), but here, rather, we employ a digital method – dictionary-based corpus query – as an exploratory spotlight, to shine a light on candidate texts which may be promising for future research on a particular topic. This paper is a preparatory step in an ongoing project on how literary texts may inform the history of discourse relating to Artificial Intelligence (AI), including privacy, identity formation, and anonymity.

How well does a dictionary of words relating to the topic of privacy classify texts in which privacy has been interpreted as a notable feature of the literary work? Despite decades of lexicon i.e. dictionary-based query of texts for computational linguistic, stylistic, and other computational literary studies, the methodology for such an evaluation is unestablished, and the problem is compounded when the topic selected is vague (as "privacy" is) rather than crisp (i.e. it would be far easier to evaluate whether a dictionary of animal names, for instance, correlates positively with novels in which animals feature prominently).

Here, we first report the results of querying Vasalou et al.'s *Privacy Dictionary* (2011), a dictionary of 616 words and phrases for automated content analysis on privacy-related texts, divided into 8 categories, in 131 canonical English-language novels from the long 19th century. We suggest some minor adjustments to Vasalou et al.'s *Privacy Dictionary*, and introduce our methods for evaluation by creating a classification of literary texts on the topic of privacy, namely, a binary classification of novels in our corpus based on one factor: Whether we can identify a scholarly article or monograph which discusses the topic of privacy at length in that novel. We report evidence of low- to low/medium strength of correlation between 3 of the 8 categories of the *Privacy Dictionary* – which Vasalou et al. dubbed "Intimacy," "PrivateSecret," and "NormsRequisites" – and this classification.

As a final step, by identifying the novels in our corpus which score highest in relative word frequency in these 3 categories, we uncover novels about which scholars have not yet commented extensively on the topic of privacy in, but which may be fruitful for extensive research on "Privacy in Novel X". The highest-scoring novel identified by this method, Maria Edgeworth's *Castle Rackrent* (1800), contains extensive discussion of privacy in its author's preface and an early commentary, which suggests that the method may yield good results. Future work could apply this method to larger corpora, e.g. "the great unread" (Cohen 2002, Moretti 2000) to uncover further candidates for the study of privacy in literature. While our method has many limitations due to the application of dictionary-based corpus query to a vague and manifold topic – privacy – at worst, our method can serve as a "tool for thought" (following Rheingold 2000), and at best, uncover texts which can contribute to studies on literature's ability to inform our current robust debates on privacy, as well as ongoing debates on the "contested spaces" of the public and private spheres in the long nineteenth century (Clark 1996).

2. Privacy

Conceptual/theoretical studies of privacy are voluminous and diverse, which poses a challenge for our research, but the prevalence of discourse around privacy in our current times suggests that the investigation is worthwhile. One categorization of the vast literature on privacy is proposed by Tavani (2007), who groups theories of privacy into four categories: 1) nonintrusion ("being let alone"); 2) seclusion ("one's being secluded from others"); 3) control ("one has privacy if and only if one has control over information about oneself") and; 4) limitation ("one has privacy when information about oneself is limited or restricted in certain contexts"). While a consistent, uniform theory of privacy has proven elusive (Vasalou et al. 2011, 2095), discourse on privacy has gained new importance with the rise of the so-called "surveillance capitalism," in which large technology companies "challeng[e] social norms associated with privacy" (Zuboff 2015, 85). This growing discourse surrounding online privacy has been met with a slew of NLP work on e.g. privacy violation detection (Silva et al. 2020), privacy language detection in medical data (Alawad et al. 2020), and privacy leaks in social networks (Canfora et al. 2018).

The long nineteenth century is particularly fertile for the study of privacy, as privacy as a concept underwent important evolutions in society, both in its literature, which increasingly explored privacy as a theme, and in law, with the foundation of the recognition of privacy as a legal right separate from copyright or defamation. Our current conceptualisation of privacy in the Anglophone world was born around the early nineteenth century, as Erica Longfellow writes: "[T]he definition of privacy that arouses the most debate for us, 'the state or condition of being alone, undisturbed, or free from public attention, as a matter of choice or right', did not come into use until 1814. That new definition signaled a change in the paradigm of public and private" (2006, 315). Per Koehler, "The invasion of privacy [...] as a theme [...] assumed an especially important role in the Victorian novel" (2016, 64). Koehler interprets this as resulting from shifting class distinctions in nineteenth century Britain: As "[e]xternal manifestations of status became less fixed [...] the newly empowered middle classes fashioned the ethos of respectability. [...] The obsession with respectability, in turn, generated an increased longing for privacy," and that "[i]n response to dazzling economic, social, and economic changes, from the Romantic period onward privacy came to be enshrined as a positive value and claimed as an important individual right" (Koehler 2016, 65). Meanwhile, privacy as a right in common law in the United States and Britain percolated throughout the 19th century, through such cases as *Prince Albert v. Strange* (which distinguished "a breach of trust [or] confidence" from traditional concepts of property, 1849, quoted in Warren and Brandeis). "The injection of private experience into public space [...] became a crucial aspect of American life as the century wore on" (per Ackerman 1997, 2) and Warren and Brandeis' landmark law journal article "The Right to Privacy," was published in 1890, "widely recognized by scholars and judges, past and present, as *the* seminal force in the development of a 'right to privacy' in American law" (Bratman 2001, 624). The contours of privacy were naturally manifold in British, American, and myriad local contexts, but these many vibrant literary and legal discourses around privacy underscore the long nineteenth century as a foundational period in the development of the privacy concept.

3. Experiment: Corpus Query of *Privacy Dictionary*

Here, we explore the results of querying Vasalou et al.'s *Privacy Dictionary* (2011), a dictionary of 616 words and phrases for automated content analysis on privacy-related texts, divided into 8 categories, in our bespoke corpus of 131 canonical English-language novels from the long 19th century. While no selection of corpus-as-canon can be "innocent" (Mark and McGurl 2015, 5), we selected all novels from the 19th and early 20th centuries originally written in English in Clarence Green's *Corpus of the Canon of Western Literature* (Green 2017), which results in 131 texts.

Category	Words / phrases	Description of Vasalou et al.	Sample
NegativePrivacy	143	"words that relate back to privacy concerns and risks as well as judgments about the source and type of violation"	afraid, being watched, deceitful
NormsRequisites	107	"the norms, beliefs, and expectations in relation to achieving privacy"	consent, control of, discreet
OutcomeState	38	"words that describe the static behavioral states and the outcomes that are served through privacy"	anonymous, liberty, security
PrivateSecret	58	"descriptors or words that express the 'content' of privacy", "what aspects people regard as being private"	secret, confidential, sensitive information
Intimacy	117	"words that portray and measure different facets of small-group privacy", "words that refer to the psychological requisites in opening up to another person as well as the emotional closeness that develops between people"	confide, friendship, gave my support
Law	43	"words employed to describe legal definitions of privacy"	illegal, policy, statute
Restriction	150	"the closed, restrictive, and regulatory behaviors employed in maintaining privacy", "the behaviors that people take to protect their privacy"	controls, hidden, lies to
OpenVisible	58	"words that represent the dialectic openness of privacy"	disclosed, posted, reveals

Table 1: Categories of *Privacy Dictionary* by Vasalou et al. 2011.

Lexicon- or dictionary-based query in Natural Language Processing (NLP) is "purely descriptive and is to count words according to large lists of words of a specific category" (Schmidt et al. 2021). Querying dictionaries of words as a model of thematics dates to the earliest days of natural language processing, and many of the limitations of querying a large literary corpus selected only for "canonicity" in the English language with a *Privacy Dictionary* created in the 21st century are obvious: finiteness and subjectivity of term selection, semantic change in lexis over time, uneven chronological distribution of lexis, and differences in e.g. British and American English (although in the latter case, the *Privacy Dictionary* often includes both orthographies). Yet dictionary-based query of historical and literary texts remains an established method in digital humanities, either as a first step in more sophisticated methods (e.g. Blanke et al. 2020) or as the experiment itself (Hogenraad 2018).

Vasalou et al.'s *Privacy Dictionary*, intended as "linguistic resource for automated content analysis on privacy-related texts," is divided into 8 categories (Table 1). Vasalou et al. created their dictionary based initially on two datasets: One-on-one interviews based on a number of privacy-related topics and scraping 859 blog posts from a popular blogging platform, Blogger, in which the bloggers discussed issues relating to privacy violations (Vasalou et al. 2011, 2098). After "construct[ing] theoretically sound categories of semantically similar words", as well as checking the semantic relatedness of words in each category using word vectors provided by the LSA website[1] and comparing the frequencies and a number of statistics on a corpus of written descriptions of privacy violations written by university staff and students, and a control corpus unrelated to privacy. Vasalou et al. envisioned the application of their *Privacy Dictionary* to

[1] See: http://wordvec.colorado.edu.

"privacy perceptions as they are expressed in online settings," as well as social science, for instance "comparing technology users' language and the language employed by academics and policymakers" (Vasalou et al. 2011, 2102, 2104). A number of subsequent studies have applied the *Privacy Dictionary* to research in social media (Islam et al. 2014), privacy-aware software systems (Casillo et al. 2022), privacy policies in B2B and B2C e-commerce (Vakeel et al. 2017), and hotel guest reviews (D'Acunto et al. 2021). The only humanistic application of the *Privacy Dictionary* that we are aware of is a study of science fiction film reception, in which Milne et al. performed a basic relative frequency query of the 8 categories of the *Dictionary* in online film reviews, reporting "a dramatic increase of the [privacy] terms used in the media after the movie release than before" (2021, 756). The *Privacy Dictionary* has not previously been applied to explicitly literary texts, but given the rigor with which it was created, and the variety of research questions to which it has been applied, we select it as the best reference dictionary available for our task.

For our experiments, we made some relatively minor modifications to Vasalou et al.'s wordlist. First, we apply a more consistent approach to lemmas (including verb conjugation, tense, and plurals): E.g. Vasalou et al. include *block*, *blocked*, *blocking*, but not *blocks*. Vasalou et al. include *be watched* and *being watched*, but not *been watched*, *am watched*, etc. As literary texts can be written in different tenses, most often past and present, we made 120 such modifications to the word list. Next, we performed manual inspection of results to check for different word senses that may skew results. For a list of hundreds of words in 131 texts, this was no small task, and settled on the pragmatic methodological step of limiting inspection to top 10 results in each *Privacy Dictionary* category query. After this inspection, we ultimately excluded only one word from the *Dictionary* completely: *judge*, as in our literary texts it resulted in more false positives than true positives due to word sense; e.g. *judge* in literary texts more often appears as a noun (a judicial official), than a verb, which would denote a privacy concern. In all following steps, we apply this modified *Privacy Dictionary* which incorporates our changes. As discussed below, we then made a final *post-hoc* inspection of query results for word sense in our top results.

3.1 *Privacy Dictionary* Category: Intimacy

Vasalou et al. describe their category of "Intimacy" as "words that portray and measure different facets of small-group privacy. It includes words that refer to the psychological requisites in opening up to another person as well as the emotional closeness that develops between people" (2011, 2100). In our corpus, the category of intimacy words is highly skewed by only 8 words/lemmas/n-grams, which account for 94.9% of the results: *friend/-s*, *family*, *conversation*, *trust*, *confidence*, *their own*, *group/-s*, *friendship*. *Friend/-s* and *family* alone account for 62.9% of results, so when visualizing the results for the Intimacy sub-dictionary, it is largely a query of these words (Figure 1 and Figure 2).

This observed distribution – with the total frequency mostly due to a small percentage of very frequent words – follows Zipf's Law (Brezina 2018, 44). These visualizations underscore how, despite the 600+ words and phrases in the modified *Privacy Dictionary*, when applied to literary texts, its model results in a very small number of highly frequent words, a concern for its application to our task. Having set out to locate literary texts in

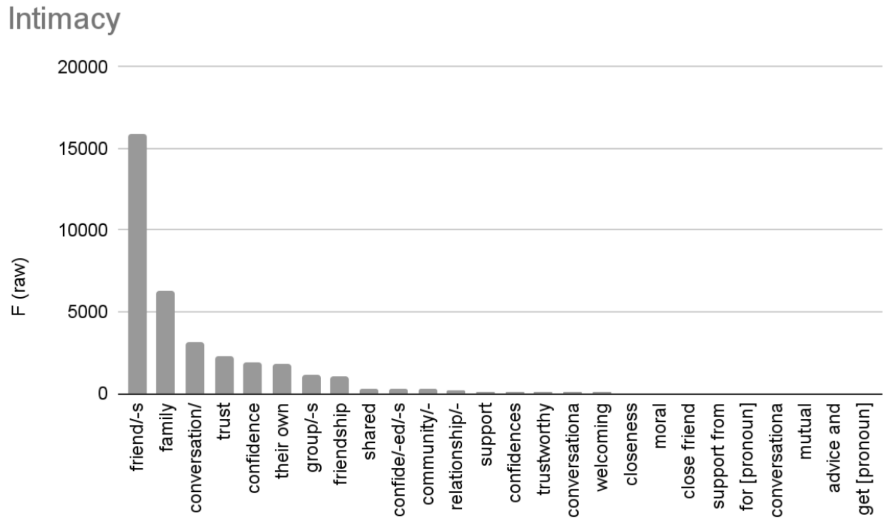

Figure 1: Raw frequency of Intimacy category of modified *Privacy Dictionary* in 131 canonical English novels.

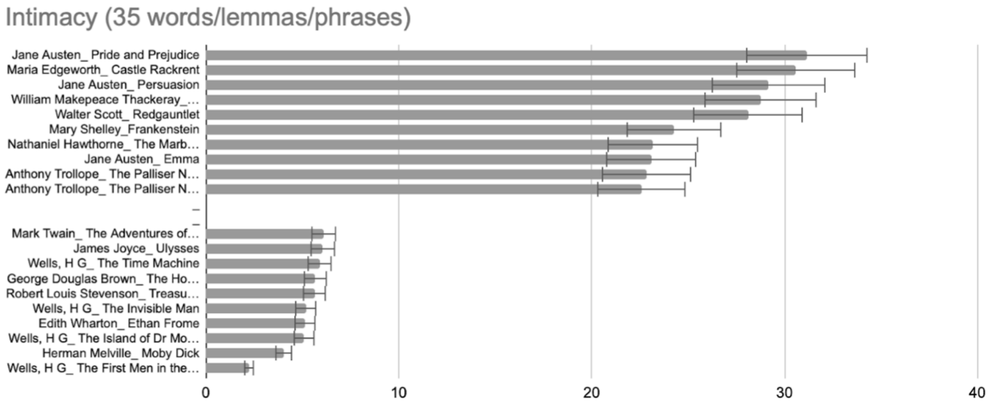

Figure 2: Texts with highest and lowest relative frequency of words in Intimacy category of modified *Privacy Dictionary*, per 10k word tokens.

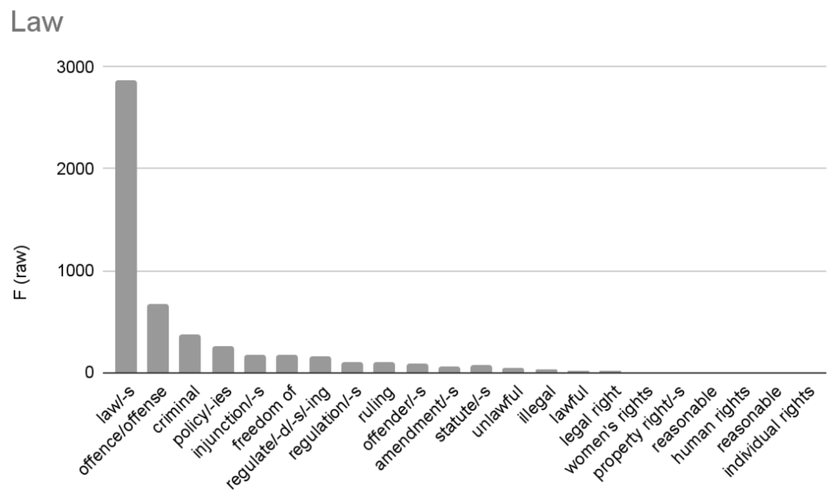

Figure 3: Raw frequency of Law category of modified *Privacy Dictionary* in 131 canonical English novels.

which "privacy" is a topic, the first sub-model is largely an extremely crude query of *friend/-s* and *family*.

How can we evaluate this query – and the queries of other sub-categories of the *Privacy Dictionary* – as evidence that the literary text is concerned with the topic of privacy? Without some gold standard to evaluate these queries against, the scholar can all too easily conjure the kinds of "leaps" from data to interpretation that Stanley Fish criticized in digital stylistics (1980, 89-90). As tentative observations, we could mention that the beloved classic on the themes of family and love, *Pride and Prejudice*, scores highest in these intimacy words, and the very title of Thackeray's *Vanity Fair* (scoring third) has become a metaphor for societal interaction. Amongst the lowest scoring in this model of intimacy, meanwhile, is Melville's *Moby-Dick*, the narration of Captain Ahab's lonely quest of revenge, amidst sailors cramped on a whaling ship with precious little privacy. These general observations, however, require some comparison. While comparison corpora would often fill this methodological step, there is no obvious comparison corpus in which the topic of "privacy" is high. We create one below, but first, we present some more of the dictionary queries in our experiment.

3.2 *Privacy Dictionary* Category: Law

Vasalou et al. describe this category as "words employed to describe legal definitions of privacy" (2011, 2100). In our literary corpus, results are again extremely skewed by a few highly frequent words: 67.5% of results are due to only two words out of 21, *law/-s* and *offense/offence*, while 93.3% of results are due to only 9 words and their plurals: *law*, *offense/offense*, *criminal*, *policy*, *injunction*, *freedom of*, *regulate* (verb lemma), *ruling*, *regulation* (Figure 3 and Figure 4).

With this query mostly based on the frequency of *law/laws* and *offense/offence*, it is likely that this is little indication of privacy as a topic, but rather, law and legal matters which figure strongly in the fabula of the novels. And indeed, law is central to the plot in Scott's *The Heart of Midlothian* (Ward 1997), John Galt's *The Entail* (1823) is a multi-generational drama involving a legal inheritance (the "entail"), while Stevenson's *Weir of Hermiston*

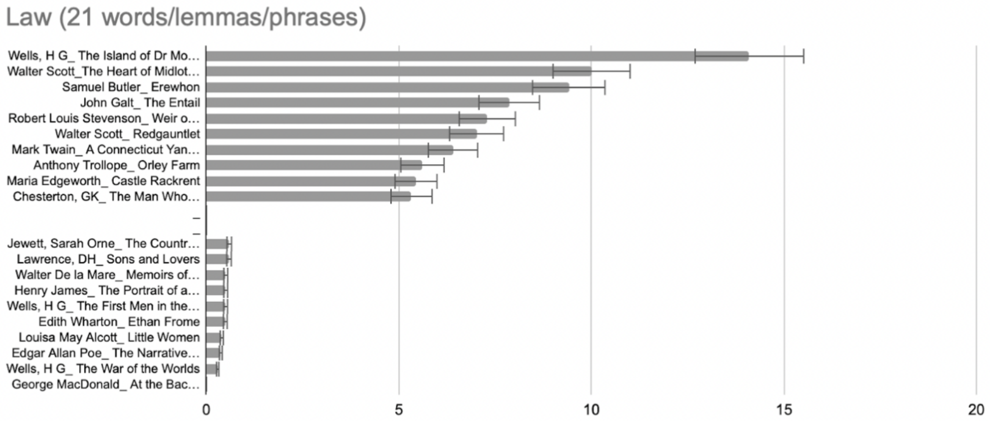

Figure 4: Highest and lowest relative frequency of words in Law category of modified *Privacy Dictionary*, per 10k word tokens.

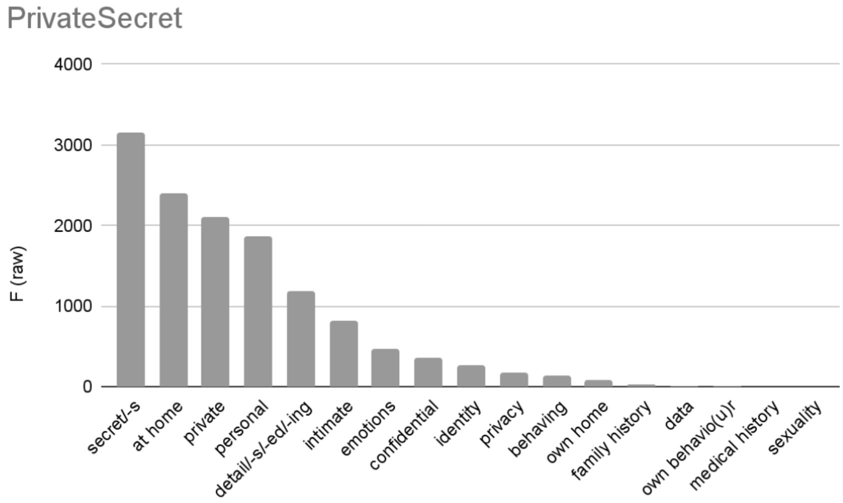

Figure 5: Raw frequency of PrivateSecret category of modified *Privacy Dictionary* in 131 canonical English novels.

features a conflict between high-born Scot Archie Weir and his cruel father, a judge. *The Island of Dr. Moreau* scores highest due to arguably false positives, as *Law* is an oft-repeated proper noun in the text (frequency: 61), meaning the system of rules created by Dr. Moreau that his creature-men must follow, in such dramatic passages as, "'Evil is he who breaks the Law,' chanted the Sayer of the Law." So far, the *Privacy Dictionary* is proving a crude method to query the topic of privacy in literary texts, especially with the sub-dictionary of Law.

3.3 *Privacy Dictionary* Category: PrivateSecret

Vasalou et al. describe this category as words which "express the 'content' of privacy. This category can be used to understand precisely what aspects people regard as being private" (2011, 2100). We expand their list of 58 words and phrases, which includes *alone*, *secure**, *prevent**, and *protect**, to include more word forms (e.g. not only *details*, but *detail**). This modified query is presented in Figure 5 and Figure 6.

Skew is a concern, as only 5 words and lemmas account for 82.5% of results: *secret/-s*,

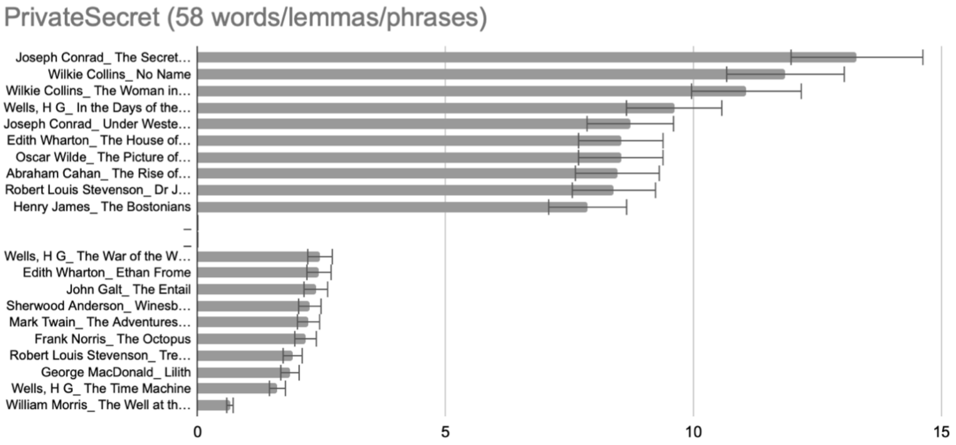

Figure 6: Texts with highest and lowest relative frequency of words in PrivateSecret category of modified *Privacy Dictionary*, per 10k word tokens.

at home, *private*, *personal*, *detail**. While this remains a fairly crude model of a topic, the query is based on a wider range and distribution of words. In a category of words centered on the concept of secrets, it makes sense that Conrad's *The Secret Agent*, a spy novel which contains *secret* in its title, scores highest, while two texts by Wilkie Collins, proto mystery novels, follow.

Many more observations could be made about each of these queries, discussing the "secrets" in more canonical texts or hypothesizing why "friends" and "family" are high in others. But, rather than interpreting the output of the word queries at greater length, we add two steps of statistical evaluation which, we suggest, are more useful for our task. Visualizations of the rest of the sub-dictionary word frequency results are available at the GitHub page linked in Data Availability below.

4. Evaluation of Query with Scholarly Publications as Classification

Evaluation of a method is often ultimately a classification task. Here, we wish to see whether the *Privacy Dictionary* reveals literary texts which scholars have discussed at length with regard to the topic of privacy. But what gold standard could be used to classify texts which do and do not relate to privacy? Is a numerical score, such as 1-10 for "low and high privacy as a topic" even possible? While ever-shifting definitions present an essential challenge of comparing privacy discussions across time and texts, this extant analog scholarship on privacy in literary texts provides an opportunity to compare with the results of our corpus query.

We suggest a method based on existing scholarly publications. We created one binary metric for our 131 novels: Whether a scholarly article or book explores "Novel X and Privacy" at substantial length.[2] For instance, querying academic publication databases such as JSTOR and Google Scholar, we found that there is an article titled "The British

[2] The methodology for this step includes scholarly commentary on "privacy" and Text X, the "private sphere," and "surveillance" and "secrets" insofar as aspects of privacy are also discussed.

Postal Service, Privacy, and Jane Austen's *Emma*" (Wheeler 1998). After reading this article to confirm a substantial discussion of privacy in the text, we assigned a score of 1 (rather than 0) to the specific text, Austen's *Emma*. This often required some digging; for instance, in "Sexuality, Shame, and Privacy in the English Novel" (Yeazell 2001), around four pages of discussion are devoted to issues of privacy in Elizabeth Gaskell's *North and South*; we thus include this as a score of 1 for *North and South*.

The main benefit of this method is that it provides some kind of classification to compare the *Privacy Dictionary* queries against. The limitations of this method are many. While we experimented with speeding up the process through automatic queries of titles and abstracts, such as through JSTOR's Constellate API,[3] the ability of this API to query titles and abstracts only led us to complete this step manually, which was a considerable task. Technical considerations and scholarly labor aside, these articles and books are heterogeneous in discipline, methodology, and way in which they explore various aspects of privacy ("an elusive and multidimensional concept whose meaning is culturally and historically contingent," per Bennett 2010, xi) in texts from vastly different time periods and genres, united only in their canonicity. Most of the studies we found are typical of contemporary literary criticism: They investigate some aspect of privacy in the text, with the goal of adding knowledge to author-/genre-/time period-specific studies, e.g. "Fierce Privacy in *The Wings of the Dove*" (Lescinski 1990) or "Feminism and the Public Sphere in Anne Brontë's *The Tenant of Wildfell Hall*" (Carnell 1998). A number of other studies come from American law journals, as privacy is, among other things, a legal concept that has generated considerable jurisprudence and commentary. For our purposes, we include law journal articles that have substantial discussions of privacy as it relates to the literary text itself. For instance, "The Huck Finn Syndrome in History and Theory: The Origins of Family Privacy" (Macias 2010) introduces the competing interests in the law of family privacy through an example in Twain's *Adventures of Huckleberry Finn*; while Macias's aim is an analysis of American lawmaking and court cases, he devoted substantial discussion to literary scholars' interpretations of Twain's text. For our purposes, we thus consider Macias's article as evidence that the topic of privacy has been interpreted in Twain's text. Finally, there are scholarly commentaries which consider the issue of privacy as it impacts the privacy of the author, e.g. the reading and scholarship of a deceased author's private letters and unauthorized texts – we have not included such studies in our classification table, but rather, academic texts which consider privacy within the diegetic world of the text.

The result is a classification table with scores for each author and text in our corpus of 131 canon novels, a sample of which is in Table 4 and the full table available at the GitHub page linked in Data Availability below.

As a first experiment in comparing the *Privacy Dictionary* word queries in our corpus of canonical novels with this new classification based on scholarly attention, we first looked at the word frequencies per category in novels with scholarly attention to the topic of privacy (1) and without (0) using log likelihood, a well-established metric in corpus linguistics (Brezina 2018). These results, calculated by Rayson's Log Likelihood Calculator (Rayson 2003),[4] are in Table 3.

[3] See: https://constellate.org.
[4] See: https://ucrel.lancs.ac.uk/llwizard.html.

Text	Privacy Scholarship	Citation
Charles Dickens, *David Copperfield*	1	Bulman, Jessica. "Publishing Privacy: Intellectual Property, Self-Expression, and the Victorian Novel." *Hastings Communications and Entertainment Law Journal* 26, no. 1 (2003): 73-118.
Charles Dickens, *Hard Times*	0	–
Charlotte Bronte, *Jane Eyre*	1	Spacks, Patricia Meyer. "The Privacy of the Novel." *NOVEL: A Forum on Fiction* 31, no. 3 (1998): 304-316.
Chesterton, GK, *The Man Who Was Thursday*	0	–
David Lindsay, A *Voyage to Arcturus*	0	–
Edgar Allan Poe, *The Narrative of Arthur Gordon Pym*	0	–
Edith Wharton, *Ethan Frome*	0	–
Elizabeth Gaskell, *North and South*	1	Yeazell, Ruth Bernard. "Sexuality, Shame, and Privacy in the English Novel." *Social Research* 68, no. 1 (2001): 119-144.
EM Forster, *Howard's End*	0	–

Table 2: A sample of our classification of literary texts in our corpus by: (1) we have identified scholarly literature which discusses the topic of privacy in this specific text at length, or (0) we have not identified such scholarship.

Privacy Dictionary Category	Novels WITH commentary (rel. F)	Novels WITH NO commentary (rel. F)	LL	p value	Significance
Intimacy	169.7	135.5	440.04	< 0.0001	very highly significant
Law	21.8	23.0	-3.51	> 0.05	not significant
OpenVisible	114.8	115.0	-0.03	> 0.05	not significant
NormsRequisites	69.2	55.2	180.5	< 0.0001	very highly significant
NegativePrivacy	168.1	173.1	-8.38	< 0.01	moderately significant
Restriction	127.8	126.9	0.43	> 0.05	not significant
OutcomeState	131.6	134.7	-4.11	< 0.01	moderately significant
PrivateSecret	61.8	52.0	98.19	< 0.0001	very highly significant

Table 3: Log likelihood comparison between modified *Privacy Dictionary* word frequencies in novels with and without scholarly commentary on privacy (rel. F = relative frequency per 100k word tokens).

These results show that 3 of the 8 bags of words in the modified *Privacy Dictionary* showed no statistically significant difference between the words in novels we marked 1 (scholarly attention to privacy in that text) and 0 (the absence of). If there is no significant difference in the word queries, then it can be inferred that word frequency cannot correlate with either 1 or 0 in these text samples. We thus conclude that these categories have no value for our task, again, which is to use the *Privacy Dictionary* to identify literary texts in which privacy might be a substantial topic. Also, 2 of the 8 categories show a significant correlation, but a negative one. This means that the *fewer* of the presumed privacy words are present, the more likely it is that a novel is marked as 1, and has a scholarly study article that explores privacy at length in that novel. We conclude that these categories – NegativePrivacy and OutcomeState – also cannot aid our task, because our aim is to find a bag of words positively relating to privacy, not the inverse. This leaves us with 3 categories of the modified *Privacy Dictionary* which are, so far, promising for our task: Intimacy, NormsRequisites, and PrivateSecret.

To evaluate correlation between our sub-dictionary queries and our binary classification, we use a non-parametric rank biserial correlation, noting the approximation of the Cureton test (Cureton 1956; also Glass 1966) with Spearman's rank correlation statistic rho (which, in turn, is equivalent to Pearson correlation of ranks of scores instead of the scores themselves). When the score ranking yields no ties, as in the data here, the Spearman approximation is exact. Thus, we assess Spearman correlations between relative frequencies and the binary classifications of works in which the frequencies are noted. As a non-parametric test, there are no particular requirements beyond sample size (we take 20 observations as a heuristic minimum; Berk 1978), and here 53 items are classified in category 1 and 77 items classified in category 0. To illustrate the distributions of relative frequencies for each of the word lists in relation to these categories, we provide tables of median relative frequency and also rank-sums for the two classification categories (rank sums being the values obtained from considering the relative frequencies in rank order, with the least relative frequency assigned the least rank, and summing the ranks that fall into each of the binary classification categories; note that it is "easier" for the rank sum of 77 items to exceed the rank sum of 53 items). The result is that three categories of the modified *Privacy Dictionary* positively correlate at a statistically significant level with scholarly attention to privacy in individual texts (Table 4).

Table 4 reports Spearman's rho for the rank correlations between the relative frequencies of observations of dictionary items and the binary text categorization described above. Adopting the convention alpha = 0.05, and adjusting alpha by dividing by the number of tests (alpha' = 0.00625), the correlation is significant for the lists associated with Intimacy, NormsRequisites and PrivateSecret. As rho can range from -1 to 0 to 1, the rho values for these categories, from 0.298 to 0.341, can be interpreted as a low to low/medium strength of correlation.

These experiments now allow us to turn to our task: Identifying literary texts in which the topic of privacy may be a notable concern and has gone unnoticed by scholars. Having identified 3 sub-dictionaries of the modified *Privacy Dictionary* as most promising for the task, this allows us to potentially identify novels which rank highest in terms of "privacy potential", as a simple sum of the relative word frequencies in these 3 sub-dictionaries

Category	Median rel. F		Rank sum		rho	p
	0	1	0	1		
Intimacy	12.31	15.44	4380	4266	0.316	0.0002
Law	1.71	1.70	5153	3493	-0.002	0.9814
NegativePrivacy	17.21	16.95	5324	3322	-0.072	0.4113
NormsRequisites	4.89	6.55	4424	4222	0.298	0.0006
OpenVisible	11.03	11.20	5033	3613	0.047	0.5916
OutcomeState	12.61	13.15	5057	3589	0.037	0.6713
PrivateSecret	4.73	5.88	4318	4328	0.341	0.000066
Restriction	12.33	12.28	4981	3665	0.069	0.4357

Table 4: Descriptive statistics and rank correlation coefficients (Spearman's rho) with p values (rel. F = relative frequency). The text category corresponds to 0 (no evident scholarly attention in relation to privacy; n=77) and 1 (evidence of scholarly attention in relation to privacy; n=53).

Text	**rel. F**	**Scholarship on privacy in text**
Jane Austen, *Persuasion*	48.72	1
Jane Austen, *Pride and Prejudice*	47.50	1
William Makepeace Thackeray, *Vanity Fair*	45.07	1
Maria Edgeworth, *Castle Rackrent*	44.09	0
Walter Scott, *Redgauntlet*	42.86	0
Jane Austen, *Mansfield Park*	41.19	1
Jane Austen, *Emma*	39.74	1
Walter Scott, *Waverley*	39.32	1
Wilkie Collins, *No Name*	38.24	0
Anthony Trollope, *The Palliser Novels 3*	37.98	1
Anthony Trollope, *The Palliser Novels 4*	37.94	1
Anthony Trollope, *The Palliser Novels 2*	37.46	1
Mary Shelley, *Frankenstein*	37.04	1
Wilkie Collins, *The Woman in White*	36.65	1
William Makepeace Thackeray, *The History of Henry Esmond*	36.12	0
Henry James, *The Ambassadors*	35.24	1
Anthony Trollope, *The Palliser Novels 1*	35.06	1
James Hogg, *The Private Memoirs and Confessions of a Justified Sinner*	34.89	1
Nathaniel Hawthorne, *The Marble Faun*	34.78	1
Anthony Trollope, *Orley Farm*	33.84	0

Table 5: Combined relative F (per 10,000 word tokens) of Intimacy, NormsRequisites, and PrivateSecret categories in modified *Privacy Dictionary*.

(see Table 5). Many of the highest results are simply less well-known novels by authors whom scholars have already investigated on the topic of privacy, such as Walter Scott and Anthony Trollope. So, to identify both texts and authors whom scholars have yet to investigate the topic of privacy in, see Table 6. We performed a final post-hoc inspection of frequent query words for word sense, and only discovered one correction to be made: "Private" appears a number of times in Crane's war novel, *The Red Badge of Courage*, to mean the military rank of private; nonetheless, even with this correction made, the novel still scored highly in Table 6.

5. Conclusion and Future Work

We report that our statistical method shows evidence of low- to low/medium strength of correlation between 3 of the categories of the *Privacy Dictionary* with observed scholarship on the topic of privacy in specific literary texts, and may assist our aim of identifying

Text	rel. F	Scholarship on privacy in text
Maria Edgeworth, *Castle Rackrent*	44.09	0
George Meredith, *The Egoist*	33.50	0
John Galt, *The Entail*	33.22	0
Stephen Crane, *The Red Badge of Courage*	32.12	0
George Gissing, *New Grub Street*	30.96	0
Samuel Butler, *Erewhon*	29.62	0
Samuel Butler, *The Way of All Flesh*	29.44	0
Norman Douglas, *South Wind*	28.63	0
Thomas de Quincey, *Confessions of an English Opium Eater*	28.54	0
Joseph Conrad, *The Secret Agent*	28.20	0
Joseph Conrad, *Under Western Eyes*	27.90	0

Table 6: Combined relative F (per 10,000 word tokens) of Intimacy, NormsRequisites, and PrivateSecret categories in modified *Privacy Dictionary*. Top results for texts where none of the author's novels were marked 1 above.

candidate texts which may be promising for research on the topic of privacy.

Did our method work? Discourse around privacy abounds in the earliest pages of Maria Edgeworth's *Castle Rackrent*, which scored highest by our ultimate method. The author's preface begins with a discussion of public vs. private lives: "We cannot judge either of the feelings or of the characters of men with perfect accuracy, from their actions or their appearance in public [...] it is only by a comparison of [people's] actual happiness or misery in the privacy of domestic life that we can form a just estimate of [them]" (Edgeworth 1801, 4). A later introduction to the novel from 1895, by Anne Thackeray Ritchie, identified privacy as a main thematic of the novel: "I don't think people could feel quite so strongly now about their own affairs as they did [in the past]; there are so many printed emotions, so many public events, that private details cannot seem quite as important" (Thackeray Ritchie 1895).

These are very promising signs that our method has identified a novel in which privacy plays a significant part, and perhaps the most surprising aspect is that *Castle Rackrent* is among the earliest in our long nineteenth century corpus, published in 1800, before the supposedly transformational discourses of privacy which took place in the ensuing century. The topic of privacy in *Castle Rackrent* and especially the other novels identified in Table 6 merits further scholarly investigation. Recalling that our project began with the desire to trace discourses relating to AI in literary texts, it is interesting that one of our methods' suggested texts is Samuel Butler's 1892 novel *Erewhon*, a satirical utopian fiction that has been considered a seminal text in the conceptual history of artificial intelligence (Brownsword 2017).

On a legal note, Vasalou et al.'s *Privacy Dictionary* is provided to researchers who request it under a Creative Commons No Derivatives license, which is how we obtained it, but an interesting and open legal question is that, as two of the authors of this article are based at a university in the European Union and now beneficiaries of the text and data mining laws ushered in by the 2019 *Directive on Copyright and Related Rights in the Digital Single Market* (popularly known as the *Text and Data Mining Directive*), EU researchers may no longer be bound by the No Derivatives restriction of the Creative Commons license when text and data mining CC-licensed material for scientific research. Article 3 of the *Text and Data Mining Directive* provides that "research organisations and cultural heritage institutions [may] in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access," and that "[a]ny contractual provision contrary to the exceptions provided for in Articles 3 [...] shall be unenforceable."[5] This could be interpreted that the No Derivatives restriction of the license (or any other CC restrictions) may not be enforceable in this research context, a position that the Creative Commons organization has commented on tangentially, in relation to Article 4 of the *Text and Data Mining Directive*, noting that "[b]ecause there are many different methods for conducting text and data mining, however, there may be some types of mining activities that will implicate the licensed rights" (Lazarova et al. 2021). Arguably, beneficiaries of the *Text and Data Mining Directive* would not be restricted by the No Derivatives restriction when performing text and data mining on material to which they have lawful access, such as a resource like the *Privacy Dictionary*, but it remains an open question as to whether the creation and publication of a derivative dictionary outside of the scientific publication would go past the act of "text and data mining." In our case, in the spirit of scholarly courtesy, we requested Vasalou's permission to modify the dictionary for our experiments, which she graciously agreed to.

Future work can expand our experiments through larger corpora. As noted in the section 1, now that we have obtained results for a dictionary method that correlates with scholarly attention to privacy in specific texts, we could apply our method to "the great unread" of the long nineteenth century and identify non-canonical candidate texts which may reveal heretofore unknown commentary or even contributions to the rich conceptual history of the evolution of privacy.

One of many limitations to our classification method is that papers on "Novel X and Privacy" may simply be more likely in authors with a very wide scholarly reception. As "privacy" as a primary topic for literary scholarly discourse is far from a very common or obvious one, it may be possible that the authors/texts with the greatest amount of scholarly commentary in general, such as Austen, might contain a paper on "Austen and privacy," simply because Austen studies has produced so many "Austen and Topic X" papers. This could be investigated by additionally classifying our canonical texts as "more canonical" (e.g. Austen, Dickens, Conrad, Thackeray) and "less canonical" (e.g. Abraham Cahan, Miles Franklin, Sarah Orne Jewett).

While there has been sophisticated computational work on interpreting the semantic coherence of automatically-generated topic models (e.g. Lau et al. 2014), we did not explore such methods in this paper, as the *Privacy Dictionary* was not automatically generated, but created by researchers following considerable theoretical and empirical methods. However, there may be promising paths in evaluating the *Privacy Dictionary* and other bespoke, human-made dictionaries through the methods applied to topic model evaluation.

Finally, now that we have created a classification of "scholarly attention to privacy in Text X" and "not" (Data Availability), future work could explore the distinctions between these sub-corpora through more sophisticated methods than dictionary query, e.g. stylometric signals and machine learning, which could replace our method.

[5] See `https://eur-lex.europa.eu/eli/dir/2019/790/oj`.

6. Data Availability

A .csv file of the canonical English novels with the citations to scholarship we used in our classification, as well as .png files of visualizations of our dictionary queries may be found at https://github.com/erikannotations/JCLS_Privacy.

7. Acknowledgements

The authors would like to thank the editors and peer reviewers for their insightful comments.

8. Author Contributions

Erik Ketzan: Conceptualization, Writing, Statistical Analysis

Jennifer Edmond: Writing

Carl Vogel: Statistical Analysis

References

Ackerman, Alan (1997). "The Right to Privacy: William Dean Howells and the Rise of Dramatic Realism". In: *American Literary Realism, 1870-1910* 30.1, 1–19. ISSN: 00029823. http://www.jstor.org/stable/27746711 (visited on 12/20/2023).

Alawad, Mohammed, Hong-Jun Yoon, Shang Gao, Brent Mumphrey, Xiao-Cheng Wu, Eric B Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Linda Coyle, et al. (2020). "Privacy-Preserving Deep Learning NLP Models for Cancer Registries". In: *IEEE Transactions on Emerging Topics in Computing* 9 (3), 1219–1230. 10.1109/TETC.2020.2983404.

Bennett, Colin J. (2010). *The Privacy Advocates: Resisting the Spread of Surveillance*. MIT Press.

Berk, Ronald A. (1978). "Empirical Evaluation of Formulae for Correction of Item-Total Point-Biserial Correlations". In: *Educational and Psychological Measurement* 38 (3), 647–652. 10.1177/001316447803800305.

Blanke, Tobias, Michael Bryant, and Mark Hedges (2020). "Understanding Memories of the Holocaust – A New Approach to Neural Networks in the Digital Humanities". In: *Digital Scholarship in the Humanities* 35 (1), 17–33. 10.1093/llc/fqy082.

Bratman, Ben (2001). "Brandeis and Warren's The Right to Privacy and the Birth of the Right to Privacy". In: *Tenn. L. Rev.* 69, 623.

Brezina, Vaclav (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Brownsword, Roger (2017). "From Erewhon to AlphaGo: for the Sake of Human Dignity, Should We Destroy the Machines?" In: *Law, Innovation and Technology* 9 (1), 117–153. 10.1080/17579961.2017.1303927.

Canfora, Gerardo, Andrea Di Sorbo, Enrico Emanuele, Sara Forootani, and Corrado A. Visaggio (2018). "A NLP-Based Solution to Prevent from Privacy Leaks in Social Network Posts". In: *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 1–6. 10.1145/3230833.3230845.

Carnell, Rachel K. (1998). "Feminism and the Public Sphere in Anne Brontë's *The Tenant of Wildfell Hall*". In: *Nineteenth-Century Literature* 53 (1), 1–24. 10.2307/2902968.

Casillo, Francesco, Vincenzo Deufemia, and Carmine Gravino (2022). "Detecting Privacy Requirements from User Stories with NLP Transfer Learning Models". In: *Information and Software Technology* 146, 106853. 10.1016/j.infsof.2022.106853.

Clark, Anna (1996). "Contested Space: The Public and Private Spheres in Nineteenth-Century Britain". In: *Journal of British Studies* 35 (2), 269–276.

Cohen, Margaret (2002). *The Sentimental Education of the Novel*. Princeton University Press.

Cureton, Edward E. (1956). "Rank-Biserial Correlation". In: *Psychometrika* 21 (3), 287–290.

D'Acunto, David, Serena Volo, and Raffaele Filieri (2021). ""Most Americans Like Their Privacy." Exploring Privacy Concerns through US Guests' Reviews". In: *International Journal of Contemporary Hospitality Management* 33 (8), 2773–2798.

Edgeworth, Maria (1801). *Castle Rackrent; an Hibernian tale*. 3rd. J. Johnson.

Eve, Martin Paul (2019). *Close Reading with Computers: Textual Scholarship, Computational Formalism, and David Mitchell's Cloud Atlas*. Stanford University Press.

Fish, Stanley (1980). *Is there a Text in this Class?: The Authority of Interpretive Communities*. Harvard University Press.

Glass, Gene V. (1966). "Note on Rank Biserial Correlation". In: *Educational and Psychological Measurement* 26 (3), 623–631.

Green, Clarence (2017). "Introducing the *Corpus of the Canon of Western Literature*: A Corpus for Culturomics and Stylistics". In: *Language and Literature* 26 (4), 282–299. 10.1177/0963947017718996.

Hogenraad, Robert (2018). "Smoke and Mirrors: Tracing Ambiguity in Texts". In: *Digital Scholarship in the Humanities* 33 (2), 297–315. 10.1093/llc/fqx044.

Islam, Aylin Caliskan, Jonathan Walsh, and Rachel Greenstadt (2014). "Privacy Detective: Detecting Private Information and Collective Privacy Behavior in a Large Social Network". In: *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, 35–46. 10.1145/2665943.2665958.

James, Henry (1900). "The Letters of Robert Louis Stevenson". In: *The North American Review* 170.518, 61–77. ISSN: 00292397. http://www.jstor.org/stable/25104937 (visited on 12/20/2023).

Koehler, Karin (2016). *Thomas Hardy and Victorian Communication: Letters, Telegrams and Postal Systems*. Springer.

Lau, Jey Han, David Newman, and Timothy Baldwin (2014). "Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 530–539. 10.3115/v1/E14-1056.

Lazarova, Ana, Thomas Margoni, Ariadna Matas, Sarah Pearson, Julia Reda, Brigitte Vézina, Kat Walsh, and Stephen Wyber (2021). "Creative Commons Statement on the Opt-Out Exception Regime/Rights Reservation Regime for Text and Data Mining under Article 4 of the EU Directive on Copyright in the Digital Single Market". In:

https://creativecommons.org/wp-content/uploads/2021/12/CC-Statement-on-the-TDM-Exception-Art-4-DSM-Final.pdf (visited on 11/29/2023).

Lescinski, Joan (1990). "Fierce Privacy in *The Wings of the Dove*". In: *Literature and Medicine* 9 (1), 125–133. 10.1353/lm.2011.0179.

Longfellow, Erica (2006). "Public, Private, and the Household in Early Seventeenth-Century England". In: *Journal of British Studies* 45 (2), 313–334. 10.1086/499790.

Macias, Steven J. (2010). "The Huck Finn Syndrome in History and Theory: The Origins of Family Priuvacy". In: *Journal of Law and Family Studies* 12 (1). https://epubs.utah.edu/index.php/jlfs/article/view/284 (visited on 11/29/2023).

Mark, Mark Algee-Hewitt and Mark McGurl (2015). *Between Canon and Corpus: Six Perspectives on 20th-Century Novels*. Universitätsbibliothek Johann Christian Senckenberg.

Milne, George R., Begum Kaplan, Kristen L. Walker, and Larry Zacharias (2021). "Connecting with the Future: The Role of Science Fiction Movies in Helping Consumers Understand Privacy-Technology Trade-Offs". In: *Journal of Consumer Affairs* 55 (3), 737–762. 10.1111/joca.12366.

Moretti, Franco (2000). "Conjectures on World Literature". In: *New Left Review* 1, 54–68. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature.pdf (visited on 12/06/2023).

Rayson, Paul Edward (2003). *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. Lancaster University (United Kingdom). https://ucrel.lancs.ac.uk/people/paul/publications/phd2003.pdf (visited on 12/06/2023).

Rheingold, Howard (2000). *Tools for Thought: The History and Future of Mind-Expanding Technology*. MIT press.

Schmidt, Thomas, Johanna Dangel, and Christian Wolff (2021). "SentText: A Tool for Lexicon-based Sentiment Analysis in Digital Humanities". In: *Information between Data and Knowledge*. 74. Werner Hülsbusch, 156–172. https://epub.uni-regensburg.de/44943/ (visited on 11/29/2023).

Silva, Paulo, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado (2020). "Using NLP and Machine Learning to Detect Data Privacy Violations". In: *IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 972–977. 10.1109/INFOCOMWKSHPS50562.2020.9162683.

Tavani, Herman T. (2007). "Philosophical Theories of Privacy: Implications for an Adequate Online Privacy Policy". In: *Metaphilosophy* 38 (1), 1–22. 10.1111/j.1467-9973.2006.00474.x.

Thackeray Ritchie, Anne (1895). "Introduction". In: *Castle Rackrent*. Macmillan and Co.

Underwood, Ted (2017). "A Genealogy of Distant Reading". In: *DHQ: Digital Humanities Quarterly* 11 (2).

Vakeel, Khadija Ali, Saini Das, Godwin J. Udo, and Kallol Bagchi (2017). "Do Security and Privacy Policies in B2B and B2C E-commerce Differ? A Comparative Study Using Content Analysis". In: *Behaviour & Information Technology* 36 (4), 390–403. 10.1080/0144929X.2016.1236837.

Vasalou, Asimina, Alastair J. Gill, Fadhila Mazanderani, Chrysanthi Papoutsi, and Adam Joinson (2011). "Privacy Dictionary: A New Resource for the Automated Content

Analysis of Privacy". In: *Journal of the American Society for Information Science and Technology* 62 (11), 2095–2105. 0.1002/asi.21610.

Wheeler, David (1998). "The British Postal Service, Privacy, and Jane Austen's *Emma*". In: *South Atlantic Review* 63 (4), 34–47. 10.2307/3201271.

Wynne, Martin (2006). "Stylistics : Corpus Approaches". In: *Encyclopedia of Language & Linguistics*. Ed. by Keith Brown. Elsevier, 223–226. 10.1016/B0-08-044854-2/00553-8.

Yeazell, Ruth Bernard (2001). "Sexuality, Shame, and Privacy in the English Novel". In: *Social Research* 68 (1), 119–144. http://www.jstor.org/stable/40971442 (visited on 11/29/2023).

Zuboff, Shoshana (2015). "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization". In: *Journal of Information Technology* 30 (1), 75–89. 10.1057/jit.2015.5.