Coding the Efficient Privacy Policy: Striking a Balance between Lexical Density and Readability

MARTA ALIĆ
Department of Information Technology and Computing,
Zagreb University of Applied Sciences,
Vrbik 8,
CROATIA

Abstract: - Privacy policies play a crucial role in informing individuals about how their personal data is collected, used, and protected. However, the effectiveness of these policies can be hindered by their complexity and lack of readability. This paper aims to explore the relationship between two variables - lexical density and text readability to derive efficient privacy policies text. By striking a balance in document coding the rate of information entropy can be managed, as well as efficiency in transparency tools. The results of privacy policies of 146 healthcare institutions in the Republic of Croatia were analyzed for their lexical density and Flesch Reading Ease (FRE) score. The implications of findings on compared results can be useful in crafting optimal privacy policy, whether one should focus on text richness through improved density or on simplicity in text understanding.

Key-Words: - privacy policy, efficiency, transparency, lexical density, readability, coding, information theory, entropy.

Received: July 14, 2024. Revised: November 27, 2024. Accepted: December 11, 2024. Published: December 23, 2024.

1 Introduction

In information management and ethics, transparency is a concept related to information (a)symmetry, a state where one party has access to (better) information than the other party. A canonical example of this relates to the used car market, where sellers know whether their cars are in good condition or are so-called "lemons" (i.e., in poor condition), but buyers have no way of knowing this completely, [1]. Consequently, to make a decision buyers have to factor in the risk of buying a car in poor condition or forego the purchase altogether. To highlight the significance of information in the functioning of markets, a group of economists was awarded the Nobel Prize in 2001 for their analysis of how imperfect information can lead to market failures, emphasizing information asymmetry as a key prerequisite for transparency [2].

So, in an economic context, information asymmetry is crucial for competitive market models. But in today's digital economy environment, privacy has become an asset, a market value as the activities that were until recently private are now becoming a source for analyzing the interests, characteristics, beliefs, worldviews, and intentions of individuals for profit. Utilizing numerous internet services, users consciously and unconsciously share various data with various entities: among themselves, towards organizations, and public authorities. And, although the systems of the digital economy rely on data exchange for the benefit of all stakeholders and society as a whole, the possibilities of data misuse, such as discrimination, [3], [4] And manipulation, are alarming.

Decision-making about privacy is partly the result of a rational "calculation" of costs and benefits. [5], influenced by the perception of these costs and benefits, as well as by social norms, feelings, and heuristics. Today individuals are constantly engaging in privacy-related transactions, even when privacy compromises may be intangible or when the exchange of personal data may not be a visible or primary component of the transaction. For example, querying a search engine is equally valuable as selling personal data (preferences, interests) in exchange for a service (search results).

In such a market, users, or data subjects, are active participants. For them to make informed decisions and have control over their data, it is necessary to ensure mechanisms for the realization of their privacy rights - quality transparency tools.

2 Problem Formulation

2.1 Efficient Information Transparency

The concept of transparency implies two dimensions: visibility, i.e., the degree of completeness of information and the ability to find it, and infertility, i.e., the degree to which information can be used to make correct decisions. [6], effective tools should be aimed at satisfying high levels of both dimensions of transparency to ensure the reduction of information asymmetry.

However, the problem of choosing the right type of information to disclose requires a deep understanding of the characteristics of the entities to be disclosed, [7]. Information is a concept that implies dependence on the context in which data is interpreted. Therefore, disclosed information should consist of significant, truthful, understandable, accessible, and useful data. Such information is referred to as semantic [8], [9] and can be pragmatically linked to decision-making processes.

Semantic information carries the key attributes that provide information in the fundamental understanding of information as a set of data that serves the recipient in the communication process, to eliminate uncertainties or reduce uncertainty and to take certain actions. So, by presenting a communication form through which signals are transmitted between data controllers and individuals, information becomes one of the main attributes of efficiency in transparency tools. Increasing transparency requires eliminating or minimizing any "noise" or interference as a disruptive factor in the communication process. In the context of the mathematical model of the communication system [10], two fundamentally different ways of message transmission are distinguished: through discrete signals and through continuous signals. Discrete signals can represent only a finite number of different, recognizable states, while in continuous signals quantities can vary in an infinite set of values.

Furthermore, concerning the level of „noise", communication can occur in the presence or absence of it, wherein the context of transparency, the goal is to ensure communication that aims to eliminate noise so that recipients have the ability to reproduce messages in their original form, leading to a reduction in information asymmetry as a fundamental objective.

2.2 Transparency Tools and Information Theory

In the context of tools and technologies, privacy policies or notices are positioned as an *ex-ante* transparency tool to raise awareness among users and inform them about the processing practices of their personal data. It is a document that familiarizes data subjects with the procedures related to the collection, sharing, use, and storage of their personal data, illustrating the entire life cycle of personal data within the organization and it is a mandatory requirement of compliance with frameworks and regulations on data protection, [11], [12].

Although much research has focused on demonstrating the usability of this tool, [13], [14], [15] on various technologies and interfaces [16], [17]. They have often been shown to be ineffective [17] due to their incomprehensibility, [18], [19], [20].

Since privacy policies, as a communication form between privacy stakeholders, are expressed in limited letters of the alphabet a discrete communication system or channel is assumed in relation to the Shannon-Weaver mathematical model and can be related to the coding theory in terms of data compression and differencing where the data entropy is a variable represents an absolute mathematical limit on how effective data from the source can be losslessly compressed onto a presumably noiseless channel. [10].

So, entropy allows quantification of information rate in a language, considering how predictable the information is, or how much redundancy exists in information source. This relates to how efficiently the language can be compressed, or source can be coded. The fundamental concept of information theory postulates that the "informational value" of a communicated message is related to the level of surprise associated with its content. In instances where a highly probable event takes place, the message conveys minimal information. Conversely, when a highly improbable event unfolds, the message becomes significantly more informative.

Privacy policies in principle represent highly probable content with a defined structure where information surprise in terms of information value is not relevant as the qualitative factors of source coding.

2.3 Coding Values

2.3.1 Lexical Density and Informativeness

Lexical density can be used to determine „information rate“ in terms of meaningfulness. As a simple proportion of lexical words (lexemes) in relation to the total number of words (occurrences) in text it can be used as a measure of the effectiveness of source coding and can be used as an entropy metric. Higher-density texts are more descriptive and therefore contain more information as the value is closer to the 1 as an absolute limit of effectiveness.

Measuring lexical density is one of the methods employed to describe discourse, and it consequently relies on the language register and genre of the text. Expository texts, such as news, informative, and technical articles, typically exhibit a higher lexical density compared to fiction. A maximum threshold for non-fiction texts is set to 40%, [21]. Privacy statements represent a specific form of linguistic expression. On the one hand, as somewhat legal documents, they have the specifics of exhibiting texts that tend to a higher lexical density, above 40%, while on the other hand, due to parts that rely on enumeration in their form, they retain the specifics of colloquial language, whose results show lexical density below 40%, and are marked by a lower representation of lexical words, [22].

2.3.2 FRE and Readability

Since the 1920s, educators have been developing methods to benchmark the complexity of texts by analyzing vocabulary weight and sentence lengths. They distilled their insights into readability formulas as objective assessments of the text weight. [23], mathematical equations obtained by regression analysis to assess the difficulty or complexity of the text for reading, but also for understanding.

The 1950s marked a period of significant progress in the field of readability. Researchers developed numerous new formulas during this time, solidifying the role of these tools in evaluating text complexity. While more than 40 different readability formulas are commonly mentioned in contemporary literature, it is important to note that by the 1980s, over 200 different readability formulas had been published. Among the many, a few stand out due to their widespread use and validation. These include the Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level Index, SMOG, Fog Test, Fry Formula, and Dale–Chall Formula. Readability levels are typically expressed in one of two ways: as a numerical value representing the weight of the text itself, often situated on a descriptive scale ranging from ‘very easy’ to ‘very difficult’; or as a numerical indicator of the educational level required to comprehend the text.

The Flesch Reading Ease formula, in particular, remains one of the most commonly used tools for assessing readability, especially in English-speaking contexts. Named after its Austrian-American creator, the FRE was first introduced as a method to gauge the complexity of children's textbooks [24]. The formula has since been simplified and reduced to focus on two primary variables: the average word length and the average sentence length. The formula is expressed as:

$$FRE = 206.835 - 0.846 * wl - 1.015 * sl$$

where *wl* is the word length expressed by the number of syllables (word length), and *sl* is the length of the sentences expressed by the number of words (sentence length).

In addition to English, various readability formulas have been adapted and validated for use in other languages, such as German, French, Dutch, Danish, Chinese, Russian, Swedish, Vietnamese, Korean, Hindi, and Hebrew, among others. This adaptability highlights the universality of the need for readability assessments across different linguistic contexts.

2.3.3 Values relations and Interplay

This paper aims to explore the relationship between two variables - lexical density and text readability - as derived values in text encoding, concerning the rate of information entropy as a measure of transparency efficiency. Transparency is maximized when information is conveyed clearly, completely, and without unnecessary complexity. A text that is both, dense and readable, achieves high transparency efficiency by ensuring that critical information is accessible and understandable, without overwhelming the reader.

Formulas for calculating variables are generally based on different methods for encoding information. While lexical density focuses on content, namely the frequency of units conveying a specific message, readability formulas such as Flesch Reading Ease are directed towards complexity, i.e., the length of units for information transmission, or meaning. Although both values aim for transparency efficiency, or reducing information asymmetry, towards a higher result on an ordinal scale relative to an absolute value of 1, they may be in contradictory relationships concerning noise reduction.

In information theory, noise refers to anything that interferes with the accurate transmission or interpretation of a message. In privacy policies, this could be excessive legal jargon, overly complex sentences, or too much information packed into a single paragraph. While reducing lexical density might lower entropy and make the text more predictable and easier to understand, it might also strip away some of the nuances that are important for fully informed decision-making. Conversely, increasing lexical density could introduce more noise in the form of complexity, potentially leading to misunderstandings or misinterpretations, and lower readability results, which in turn reduces transparency.

When it comes to decision-making based on privacy policies, the text should be clear, comprehensive, and informative—neither overly simplistic nor unnecessarily complex. This clarity is crucial for measuring the effectiveness of information transparency. Increasing the lexical density of a message typically leads to an increase in entropy, resulting in greater randomness or unpredictability, which can be a significant factor in the decision-making process for the data subject. The informational value of surprise is important, as users need to be alerted to any unusual or unexpected data processing practices. However, predictability is equally important because individuals should not be surprised by hidden clauses or unclear language; they need to fully understand what they are consenting to. This is where a careful balance between lexical density and readability becomes essential.

3 Problem Solution

The research material includes the examination of the results of two variables measured on the privacy policies of 146 healthcare institutions in the Republic of Croatia.

The text of the document was entirely copied into a blank Word document where the unit of further analysis was set: titles were removed, and email addresses and hyperlinks were replaced with X's to not influence the syllable count and lexical density results. Additionally, using the Wordcount option, the number of words and characters (excluding spaces) was recorded, and the number of sentences was manually counted by the author, a Croatian language and literature professor. Then, the text was copied into [25], chosen as the most reliable tool for syllable counting after comparing manual counting results and different syllable counting software. Furthermore, to assist in analyzing the number of full words, the text was copied into a text analysis tool. [26], to extract lexical words, or lexemes (nouns, verbs, pronouns, adjectives, numbers, adverbs), based on their frequency of occurrence in the text. Lexemes were extracted from the obtained results, and their final count was recorded for further analysis of the text's lexical density. Subsequently, based on the obtained results of syllable, word, and sentence counts, readability indices were calculated using the Flesch Reading Ease (FRE) formula adapted for the Croatian language. The analysis was conducted using the customized formula for the Croatian language in the Excel software.

The Croatian variant of the formula, developed by [27] based on a contrastive analysis of English and Croatian corpora consisting of 100,000 words from various types of texts and publications published after 1995., adjusts the index by 50: FRE = 206.835 - 0.846 wl - 1.015 sl + 50. Readability is expressed on a scale from 0 to 100, where each category is described descriptively. In the Croatian variant, the scale is as follows: 80 – 100 = easy; 60 – 80 = standard; 50 – 60 = fairly difficult; and 0 – 50 = very difficult.

The analysis of the obtained results was primarily conducted concerning the descriptive indicators of the sample as values on an ordinal scale. For both variables, the mean, minimum, and maximum were calculated (Table 1).

Table 1. Results of lexical density and FRE values

	Lexical density	Flesch Reading Ease (FRE)
Mean	45,66%	26,07
Median	45,45	22,87
Minimum	21,13%	-20,11
Maximum	77,08%	68,14

Comparing the results of both values in Table 1, it's possible to conclude that both the mean and median values of lexical density fall within a range higher than 40%, which is characteristic of expository texts, while the upper limit does not exceed 77%. This high lexical density indicates a very content-heavy text, likely filled with specialized or technical language, which could be harder for a general audience to understand. Adversely, a minimum of 21% suggest a tendency to defined structures of policy representations and minimal information value provided.

The readability results indicate a high degree of difficulty in understanding the text, with a median value of approximately 23, a mean value of 26, and the lowest value being negative in 6 institutions, indicating extremely complex texts. This could be due to very long sentences and/or very complex vocabulary, making the text nearly incomprehensible for the average reader.

Since the examined sample consists of institutions that can be divided into two groups based on the measured values, it's valid to conduct a comparative analysis of institutions regarding the selected values.

The tendency of institutions to code text regarding lexical density lower than the mean is shown by 76 institutions or 52%. Also, a significant number of 70 institutions (48%) show results in the range of 40-50%.

When it comes to readability, the frequency results of the relevant scale show more concerning outcomes shown in Table 2.

Table 2. FRE category frequency

FRE value	Weight category	No. of institutions
0-50	very difficult	133
50-60	fairly difficult	11
60-80	standard	2
80-100	easy	0

So, a great majority (91%) of the examined privacy policies are very difficult to read and understand, and not a single text can be considered easy to comprehend.

Segmenting further the sample of 133 privacy policies that are very difficult concerning lexical density, the result of the mean value slightly decreased from 45.66% to 45.28% (while the median value remained the same), indicating consistency in the values of lexical density of privacy policies.

However, concerning the privacy policies that achieved readability scores above 50, the median and mean values are slightly higher, at 49%, with a minimum value of 39% and a maximum of 61%, suggesting a proportional increase in both values throughout the text.

4 Conclusion

The analysis reveals an interesting interrelation between lexical density and readability in the examined privacy policies. Despite lexical density being relatively consistent across the texts, indicating a uniformity in the frequency of terms conveying specific messages, the readability scores suggest a significant challenge in comprehending the content. This disparity between lexical density and readability implies that while the texts may contain a consistent density of terms, the complexity and structure of these terms contribute to the overall difficulty in understanding the content. So, while the information may be rich, leading to an increase in unpredictability and higher entropy values, it is not effectively communicated or easily understood by the target audience, indicating substantial "noise" in the communication channel. That is, higher entropy may indicate a wealth of information, but if it cannot be effectively communicated and understood by the audience, its value may be diminished.

However, the challenge lies in balancing information richness with noiseless coding praxis. Therefore, there is a need to optimize the encoding of information to maintain a balance between these dichotomous values, ultimately reducing entropy without compromising the comprehensibility of the content.

In the context of privacy policies, where clarity and comprehension are paramount for informed decision-making prioritizing readability is shown to be essential. By using clear and concise, yet limited wording, and logically organizing information, policy writers can make the content more understandable to a wider audience, reducing the risk of confusion or misinterpretation, thus elevating the transparency and efficiency in the process.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the author used ChatGPT service in order to translate content from Croatian, authors' native language. After using this tool/service, the author reviewed and edited the content as needed and take full responsibility for the content of the publication.

References:

[1] G. Akerlof, "The Market For 'Lemons': Quality Uncertainty And The Market Mechanism," *Q J Econ*, vol. 4, no. 3, pp. 488–500, 1970, doi: 10.2307/1879431.

[2] J. B. Rosser Jr., "A Nobel Prize for Asymmetric Information: The economic contributions of George Akerlof, Michael Spence and Joseph Stiglitz," *Review of Political Economy*, vol. 15, no. 1, pp. 3–21, Jan. 2003, doi: 10.1080/09538250308445.

[3] J. Kas, R. Corten, and A. van de Rijt, "The role of reputation systems in digital discrimination," *Socioecon Rev*, vol. 20, no. 4, pp. 1905–1932, Oct. 2022, doi: 10.1093/ser/mwab012.

[4] P. S. Attri and H. Bapuji, "Digital Discrimination in Sharing Economy at the Base of the Pyramid," in *Sharing Economy at the Base of the Pyramid: Opportunities and Challenges*, B. and S. D. M. Qureshi Israr and Bhatt, Ed., Singapore: Springer Nature Singapore, 2021, pp. 221–247. doi: 10.1007/978-981-16-2414-8_10.

[5] A. F. Westin, *Privacy and Freedom*, no. 1. New York: Atheneum, 1967. doi: 10.2307/3479272.

[6] G. Michener and K. Bersch, "Identifying transparency," *Information Polity*, vol. 18, no. 3, pp. 233–242, 2013, doi: 10.3233/IP-130299.

[7] J. Feng and Y. Wang, "'No Representation without Information Flow'-Measuring Efficacy and Efficiency of Representation: An Information Theoretic Approach," *WSEAS Transactions on Computers*, vol. 8, no. 3, pp. 494–505, 2009.

[8] S. Sequoiah-Grayson, "The metaphilosophy of information," *Minds Mach (Dordr)*, vol. 17, no. 3, pp. 331–344, 2007, doi: 10.1007/s11023-007-9072-4.

[9] M. Turilli and L. Floridi, "The ethics of information transparency," *Ethics Inf Technol*, vol. 11, no. 2, pp. 105–112, 2009, doi: 10.1007/s10676-009-9187-9.

[10] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[11] G. Greenleaf, "Sheherezade and the 101 data privacy laws : Origins, significance and global trajectories," *Journal of Law, Information and Science, Special Edition: Privacy in the Social Networking World*, UNSW Law Research Paper No. 2013-40, 2014, doi: 10.2139/ssrn.2280877.

[12] L. Luić and M. Alić, "Computing the intelligent privacy-engineered organization: a metamodel of effective information transparency enhancing tools/technologies," in *Human-Assisted Intelligent Computing*, in 2053-2563. IOP Publishing, 2023, Bristol pp. 6–1 to 6–15. doi: 10.1088/978-0-7503-4801-0ch6.

[13] P. Murmann and S. Fischer-Hübner, "Tools for Achieving Usable Ex Post Transparency: A Survey," *IEEE Access*, vol. 5, pp. 22965–22991, 2017, doi: 10.1109/ACCESS.2017.2765539.

[14] F. Schaub, R. Balebako, and L. F. Cranor, "Designing Effective Privacy Notices and Controls," *IEEE Internet Comput*, vol. 21, pp. 70–77, 2017, doi: 10.1109/MIC.2017.75.

[15] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal, "How Short Is Too Short ? Implications of Length and Framing on the Effectiveness of Privacy Notices," in *Symposium on Usable Privacy and Security (SOUPS)*, Denver, 2016.

[16] S. Wachter, "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR," *Computer Law and Security Review*, vol. 34, no. 3, pp. 436–449, 2018, doi: 10.1016/j.clsr.2018.02.002.

[17] A. Soumelidou and A. Tsohou, "Effects of privacy policy visualization on users' information privacy awareness level," *Information Technology & People*, vol. 33, no. 2, pp. 502–534, Jan. 2020, doi: 10.1108/ITP-08-2017-0241.

[18] H. Habib and L. F. Cranor, "Evaluating the Usability of Privacy Choice Mechanisms," in *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, Boston: USENIX Association, Aug. 2022, pp. 273–289.

[19] W. Brunotte, L. Chazette, L. Kohler, J. Klunder, and K. Schneider, "What about My Privacy? Helping Users Understand Online Privacy Policies," in *Proceedings of the International Conference on Software and System Processes and International Conference on Global Software Engineering*, in ICSSP'22. New York: Association for Computing Machinery, 2022, pp. 56–65. doi: 10.1145/3529320.3529327.

[20] J. R. Reidenberg, T. Breaux, L. F. Cranor, B. French, A. Grannis, J. Graves, F. Liu, A. McDonald, T. Norton, R. Ramanath, N. C. Russell, N. Sadeh and F. Schaub, "Disagreeable Privacy Policies: Mismatches between Meaning and Users Understanding," *Berkeley Technology Law Journal*, vol. 30, January, 2015, doi: 10.2139/ssrn.2418297.

[21] J. Ure, "Lexical density and register differentiation," *Applications of Linguistics*, pp. 443–452, 1971.

[22] M. Alić, "Privacy Notice Informativeness: in a Search for Benchmark," in *2023 46th MIPRO ICT and Electronics Convention (MIPRO)*, 2023. Opatija pp. 1496–1500. doi: 10.23919/MIPRO57284.2023.10159806.

[23] W. Dubay, *The Principles of Readability*, ERIC Clearinghouse, 2014.

[24] R. Flesch, "A Readability Formula In Practice," *Elementary English*, vol. 25, no. 6, pp. 344–351, 1948

[25] "Syllable Counter", [Online]. https://syllablecounter.net/ (Accessed Date: July 22, 2024).

[26] "Text Analyser," Online-Utility.org, [Online]. https://www.online-utility.org/text/analyzer.jsp (Accessed Date: June 25, 2024).

[27] S. Brangan, *"Developing readability formulas for healthcare communication in Croatian language,"* PhD thesis, Zagreb, 2011.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The author have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)