MIPRO 2023, May 22 - 26, 2023, Opatija, Croatia

Privacy Notice Informativeness: in a Search for Benchmark

Marta Alić, Ph.D.
Zagreb University of Applied Sciences, Zagreb, Croatia
marta.alic@tzv.hr

Abstract - **The term transparency in the disciplines of information management, business ethics and information ethics is commonly used for forms of information visibility and access to information, with the aim of reducing information asymmetry among stakeholders. The principle of transparency is also one of the main mechanisms in data protection and regarding regulations, such as General Data Protection Regulation. But requirements for assesing it's functionality are somewhat ambiguus and hard to evaluate. One of proposed metrics can be the amount of information the transparency mechanism is trying to convey. The basic (ex ante) tool of transparency is the publication of privacy notices, with the purpose of informing the individuals about the procedures related to the collection, sharing, use and storage of their personal data. By using lexical density as a metric to evaluate how much information there is in privacy statements of two major IT companies, Google and Microsoft, across five languages (Croatian, English, German, French, Italian) and comparison within set and generic sections of privacy notice, the aim of this paper is to identify reference scale for this, specific legal documents, between descriptive and explanatory narrative.**

Keywords - privacy notice, lexical density, informativeness, transparency, Google, Microsoft

I. Introduction

In modern times of information abundance, when the rights of the individual are strengthened, transparency is set as an integral part, a mechanism, in building a trust and a system in which the individuals take an increasing place as a stakeholders in decision making. Transparency is also the principle built-in the legislation of various areas of human activity. The individuals' right to their own choice is becoming stronger in the field of privacy protection, as a fundamental human right, and in today's world of digital transformation, it is increasingly focused on the protection of personal data in digital environments. In this context, privacy is considered to be the right and ability of persons to communicate through digital channels, while controlling the dissemination of information and being informed about who and how participates in the interaction, what information is exchanged and what the information is used for. Furthermore, individuals have broad rights to information transparency of the data collected and processed about them. The basic (ex ante) tool of transparency is the publication of privacy notice, a set of data aimed at familiarizing data subjects with the actions of the system or organization regarding the collection, sharing, use and storage of their personal data. Theoretically, an organization's privacy notice contains all the information that users need to be aware of in order to make informed decisions and control the processing of their personal data. The concept itself is defined as The Individual Participation Principle of the fair information practices. The principle refers to the right of individual „to have data relating to him communicated to him, within a reasonable time, at a charge, if any, that is not excessive; in a reasonable manner, and in a form that is readily intelligible to him" [1]. Today, reporting on data practices is an important aspect of data protection frameworks and regulation, such as the General Data Protection Regulation (GDPR) [2].

II. Research context

Although a lot of research is focused on examining the content of privacy policy notifications in relation to the regulatory requirements of individual data protection standards, there is a lack of studies that focus on the quality of the mechanism by which these contents are communicated to the target public.

Natural language processing and machine learning methods have also been used to examine the corpus of privacy notices from the standpoint of linguistics and philology. [3][4] through machine learning algorithms categorize different sections of priva5te policy statements, while [5][6][7] are aimed at examining the readability, ie. intelligibility of content. [8] suggests supervised learning in determining the categories of data processing covered by the privacy notice, and other approaches are directed towards comparing the parts, section of the text [9][10] of the privacy statement.

A. Lexical density

Lexical density determines how meaningful the text is, that is, how much information it conveys. Higher density texts are more descriptive and therefore contain more information. Essentially, communication in written form is denser in lexical meaning than that in the colloquial form [11] [12].

Lexical density, expressed by the proportion of lexical words (lexemes) in relation to the total number of words (occurrences), therefore affects the degree of ease with which the recipient understands the message in a particular communication process and can affect memory and sentence retention [13] .

Measuring lexical density is one of the methods used to describe discourse, and therefore depends on the language register and genre of the text. The higher the number of lexical words, the denser the lexical text, so in more complex texts this value is closer to the value 1 [14]. Expository texts, such as news, informative and technical articles, have a higher lexical density than fiction. According to research by the website Analyze My Writing [15], articles on Wikipedia have a density between 55% and 58%, while a sample of articles from the BBC News and New York Times have a density between 56% and 58%. The same page also researched various forms of short fictional prose, the results of which indicate a lexical density between 48% and 51%. For comparison, interview transcripts were taken as a sample of colloquial language whose average lexical density was close to 45%.

The higher lexical density of non-fiction texts (40-65%) in relation to fiction texts (40-54%) is also pointed out by [16], while the upper limit of 40% is set for non-fiction texts [11]. Furthermore, [17] reveals how lexical density can vary drastically within a single text, which will be examined in the research conducted.

But lexical density can also be affected by the relationship between communication participants [11]. According to [16] and [18] specific guidelines, law enforcement orders and literature that authors expect to be available to the reader for re-reading tend to maximize lexical density. In principle, privacy policy statements can be added to this group as legal texts with some instructions for action, available for reading "on request". In studies of the lexical density of spoken and written material in different European countries and at different age groups [19], the results showed that the lexical density of the one population group is similar and depends on the morphological structure of the mother tongue and the sampled age group, given that lexical density for an individual grows with age and education, depending also on communication style, circumstances and health condition [20]. Therefore, the lexical density in European countries is also the highest for adults, while the variations assessed as lexical diversity were higher for teenagers as the same age group (13-year-olds, 17-year-olds).

III. Methodology

The research was conducted on the privacy policies of two technology companies, which have a large number of digital services in their portfolio through which they collect a large amount of data, including personal ones - Microsoft and Google.

Privacy notices are divided into sections taking into account the principles of Fair Information Practices – information / awareness, choice / consent, access / participation, integrity / security and implementation / legal protection, accepted by the Organization for Economic Co-operation and Development (OECD) [21] and in the works of [3] and [4]:

- collection: the section explains what information and how it is collected from data subjects;

- purpose: the section explains the purpose of data collection and its use;

- selection / control: the section explains choices and control options available to data subjects;

- sharing - the section states with whom (third party) and under what circumstances the information is shared;

- security - the section discusses standard practices for the data protection;

- retention - the section explains the procedure and deadline for retention of data;

- specific - the section lists practices that apply only to a specific group of users (eg. children, Europeans or residents of California);

- policy change - the section explains how users will be informed about changes in privacy practices.

In the first part of the research, the privacy notices were analyzed regarding the sections in 5 languages: Croatian, English, German, Italian and French. When determining the lexical density, the formula was used:

number of different words / number of total words * 100

and the amounts are shown as a percentage.

The analysis was performed with the Text Analyzer tool [22], and the texts were purified from the peripheral text, titles, bullets, domains and other elements that could affect the results.

In the second part of the research, notices in the Croatian language were analyzed in more detail concerning the relationship of full-meaning words, or lexems, (nouns, verbs, pronouns, adjectives, numbers, adverbs) and grammatical words (prepositions, conjunctions, exclamations, particles). The analysis was performed manually by the author, a professor of Croatian language and literature, using the formula:

lexical words / total number of words * 100

in order to obtain a percentage representation of the ratio.

IV. Results analysis

A. *First part*

The first part of the research, related to the calculation of the ratio of the number of different words and the total number of words in the text of the privacy notice, shows comparative results in 5 languages (Croatian, English, German, French, Italian) according to the given sections.

The results of Google's privacy notice [23] in the five selected languages show a range of 43% to 57%, which is in line with the results for non-fiction texts. The highest average lexical density of privacy notice is in Croatian, while the lowest is in French. A comparative analysis of the lexical density of individual sections shows the highest lexical density (Table 1) of the section related to data security settings, except for German language. With the next section on higher density range, relating to policy change conditions, it is consisted of two short sections and their discourse implies an explanatory model of narrative, in contrast to the remaining parts of notices that focus mainly on specification or the listing of individual cases related to the topic and purpose of the section.

The lowest average lexical density (Table 1) in all languages has a section related to the collection, which in its content contains repetitive forms when listing all the necessary ways of collecting data.

The Microsoft Privacy Statement [24] in the five selected languages shows a range of 40% to 55% and the highest lexical density (Table 2) is shown in the Croatian language version, as in the case of Google's privacy notice.

Table 1 Google privacy policy lexical density (%)

Section / Language	Croatian	English	German	French	Italian	Average
Collection	45,80	31,81	36,64	31,78	38,07	36,82
Purpose	49,51	34,34	39,53	35,30	39,61	39,66
Select/Ctrl	49,32	35,40	41,61	35,22	40,15	40,34
Sharing	58,78	46,21	50,09	43,12	47,57	49,15
Security	73,57	61,25	57,94	56,96	65,92	63,13
Retention	62,50	51,12	60,71	51,60	58,22	56,83
Specific	55,88	41,54	44,26	41,83	45,59	45,82
Policy Change	62,57	57,05	59,03	51,67	55,20	57,10
Average	57,24	44,84	48,73	43,43	48,79	

The lowest lexical density (Table 2) is recorded in the English version of the Statement. Here, too, on average, the highest lexical density is in the security section, with the exception of the English version, which is also "followed" by the section on policy change conditions.

Table 2 Microsoft privacy notice lexical density (%)

Section / Language	Croatian	English	German	French	Italian	Average
Collection	42,25	28,73	34,59	28,66	31,36	33,12
Purpose	43,65	29,34	36,10	28,02	30,20	33,46
Select/Ctrl	39,42	25,27	31,99	26,55	29,92	30,64
Sharing	58,16	42,09	49,59	42,15	46,20	47,64
Security	85,1	59,23	79,48	66,02	72,63	72,51
Retention	64,00	47,47	53,25	49,28	47,67	52,33
Specific	34,20	25,20	33,25	25,25	26,13	28,81
Policy Change	75,00	63,76	71,54	58,89	69,33	67,70
Average	55,23	40,14	48,72	40,60	44,18	

The section on specific practices has the lowest lexical density in the Microsoft Statement, which is focused on aimed at determining the requirements under the California Consumer Privacy Act (CCPA), and they contain a distinct form of enumeration.

When comparing the average lexical density by sections in both organizations (Table 3), the results show low values (below 40%) precisely in the sections that are highly "enumerative", ie focus on specification rather than descriptive explanation, while the highest lexical density is recorded in sections with more explanatory expressions.

Table 3 Comparison of lexical density results of Google and Microsoft privacy notices (%)

Section	Google	Microsoft
Collection	36,822	33,120
Purpose	39,661	33,467
Select/Ctrl	40,344	30,640
Sharing	49,158	47,645
Security	63,134	72,510
Retention	56,833	52,336
Specific	45,822	26,139
Policy Change	57,109	67,708
Average	48,61038	45,44563

B. Second part

The results of a more detailed analysis of the lexical density of Google's privacy notice, conducted in the Croatian language version, shown in Table 4, confirm the highest density of the security section, followed by the section on policy change conditions, which is mainly descriptive. The lowest density was confirmed in the part related to the listing of data collection methods, followed by sections with the stated purposes and methods of selection, ie control, which focus on specifications, listing individual practices within the organization.

The average lexical density of the entire corpus (ratio of lexemes to total words) is 33.431%.

Table 4 Results of a detailed analysis of Google Privacy notice in the Croatian language

Section	Number of words	Different words	Grammatical words	Lexems	Density
Collection	786	360	159	201	25,57%
Purpose	721	357	149	208	28,84%
Select/Ctrl	590	291	125	166	28,13%
Sharing	495	291	101	190	38,38%

Security	193	142	33	109	56,47%
Retention	208	130	44	86	41,34%
Specific	569	318	109	209	36,73%
Policy Change	171	107	28	79	46,19%
Total	3733	1996	748	1248	

In the case of the Microsoft Privacy Statement, the results (Table 5) show the highest lexical density in the section thematically related to data security. Furthermore, the lowest lexical density is confirmed in the part that has the most enumerations (listings) - the conditions and settings of privacy under the California Consumer Privacy Act.

The average lexical density of the entire corpus (ratio of lexemes to total words) is 24.013%.

Table 5 Results of a detailed analysis of Microsoft Privacy Statement in the Croatian language

Section	Number of words	Different words	Grammatical words	Lexems	Density
Collection	1640	693	368	325	19,81%
Purpose	1560	681	332	349	22,37%
Select/Ctrl	1258	496	230	266	21,14%
Sharing	501	291	111	180	35,92%
Security	81	69	11	58	71,60%
Retention	374	239	81	158	42,24%
Specific	842	288	162	126	14,96%
Policy Change	132	99	27	72	54,54%
Total	6388	2856	1322	1534	

Comparison of both documents in Croatian language (Table 6), shows the difference for Google document by 10.898%, and for Microsoft 10.117%. The mutual difference between the results for the two statements is 3.164% for the analysis based on the formula for the ratio of the number of different words and the number of total words, and 2.383 based on the analysis of lexemes in relation to the total number of words, which shows consistent results that can serve as a reference values.

Table 6 Comparison of lexical density results of Google and Microsoft privacy notices in Croatian (%)

	1st part analysis		2nd part analysis	
Section	G	M	G	M
Collection	36,82	33,12	25,57	19,81
Purpose	39,66	33,46	28,84	22,37
Select/Ctrl	40,34	30,64	28,13	21,14
Sharing	49,15	47,64	38,38	35,92
Security	63,13	72,51	56,47	71,60
Retention	56,83	52,33	41,34	42,24
Specific	45,82	26,14	36,73	14,96
Policy Change	57,10	67,70	46,19	54,54
Average	48,61	45,44	37,71	35,327

G=Google; M=Microsoft

V. Discussion

In the context of privacy protection, the requirement of inferability [25], the ability of respondents to act based on the information obtained, serves as a global practice in the protection of the rights of individuals based on the concept of consent.

Transparency, which ensures this requirement, is a complex concept that goes beyond simply ensuring the visibility and presentation of this information, and it also includes the quality of its mechanisms [26]. Furthermore, as a functional requirement of information systems, it is linked to the attributes of trust [5] [27], responsibility [28] [29], but also informed decision-making by stakeholders in a certain system [30] [31].

Privacy notices, therefore, should provide information in a concise, effective, and concise manner to avoid fatigue due to excessive information [32]. Furthermore, the recommendations of the Article 29 Data Protection Working Party in its Transparency Guidelines under Regulation 2016/679 refer to best practices for clear writing [33], which include providing information in the simplest possible way, while avoiding complex sentences and language structures.

Also, the information should be concrete and clear and should not be formulated in an abstract or ambiguous way. Also, when designing information in a digital environment, the Working Party calls for the application of a layered approach, which assumes a design and graphic arrangement that supports enumeration.

Therefore, privacy statements represent a specific form of linguistic expression in perceiving the results obtained. 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.

And with regard to this duality, the difference in the discourse, it is necessary to consider the results of this research.

VI. CONCLUSION

Results of this research show that when it comes to benchmarking informativeness of privacy notices there is a significant dependency to the characteristics of language used when shaping the content of these documents, as analysis on documents in Croatian language show higher lexical density than in four other compared languages. Respecitvely, there is a notable discrepancy between document sections within all inspected languages, depending on the set target of information communicated within them. By determining the factors related to the characteristics of language when shaping the content of privacy notice and placing them in relation to the properties of transparency mechanisms, the research in question provides an insight into the importance of these aspects when shaping the strategy of information transparency, not only in the field of data protection and legal sciences, but the results can also be applied as a basis for the creation of guidelines for ensuring effective transparency tools, contributingt to the current literature in the field of privacy requirements engineering.

References

[1] IAPP - International Association of Privacy Professionals, “Fair information practices,” 2023. https://iapp.org/resources/article/fair-information-practices/

[2] European Parliament and of the Council, “Regulation (EU) 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation),” Off. J. Eur. Communities, vol. OJ L 119/1, pp. 1–88, 2016, [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[3] N. Guntamukkala, R. Dara, and G. Grewal, “A machine-learning based approach for measuring the completeness of online privacy policies,” Proc. - 2015 IEEE 14th Int. Conf. Mach. Learn. Appl. ICMLA 2015, pp. 289–294, 2016, doi: 10.1109/ICMLA.2015.143.

[4] S. Wilson et al., “The Creation and Analysis of a Website Privacy Policy Corpus,” in Proceedings ofthe 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1330–1340.

[5] T. Ermakova, H. Krasnova, and B. Fabian, “Exploring the impact of readability of privacy policies on users’ trust,” 24th Eur. Conf. Inf. Syst. ECIS 2016, no. April, 2016.

[6] A. K. Massey, J. Eisenstein, A. I. Anton, and P. P. Swire, “Automated text mining for requirements analysis of policy documents,” 2013 21st IEEE Int. Requir. Eng. Conf. RE 2013 - Proc., pp. 4–13, 2013, doi: 10.1109/RE.2013.6636700.

[7] G. Meiselwitz, “Readability assessment of policies and procedures of social networking sites,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8029 LNCS, pp. 67–75, 2013, doi: 10.1007/978-3-642-39371-6-8.

[8] E. Costante, Y. Sun, M. Petkovic, and J. Den Hartog, “A machine learning solution to assess privacy policy completeness (short paper),” Proc. ACM Conf. Comput. Commun. Secur., no. November, pp. 91–96, 2012, doi: 10.1145/2381966.2381979.

[9] F. Liu, S. Wilson, F. Schaub, and N. Sadeh, “Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies,” AAAI Fall Symp. - Tech. Rep., vol. FS-16-01-, pp. 264–269, 2016.

[10] R. Ramanath, F. Liu, N. Sadeh, and N. A. Smith, “Unsupervised alignment of privacy policies using hidden Markov models,” 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf., vol. 2, no. June, pp. 605–610, 2014, doi: 10.3115/v1/p14-2099.

[11] J. Ure, “Lexical density and register differentiation,” Appl. Linguist., vol. 443–452, 1971.

[12] M. A. K. Halliday, “Spoken and written language,” 1989.

[13] C. A. Perfetti, “Lexical density and phrase structure depth as variables in sentence retention,” J. Verbal Learning Verbal Behav., vol. 8, no. 6, pp. 719–724, 1969, doi: 10.1016/S0022-5371(69)80035-6.

[14] V. Johansson, “Lexical diversity and lexical density in speech and writing: a developmental perspective,” Work. Pap. Linguist., vol. 53, no. 0, pp. 61–79, 2009.

[15] “Analyze My Writing.” https://www.analyzemywriting.com/

[16] M. Stubbs, “Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture,” 1996.

[17] T. Berber-Sardinha, “Comparing corpora with WordSmith Tools: How large must the reference corpus be ?,” WCC ’00 Proc. Work. Comp. corpora, pp. 7–13, 2000.

[18] D. Biber, Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure. John Benjamins Publishing, 2007.

[19] V. Johansson, “Toward a cross-linguistic comparison of lexical quanta in speech and writing,” Writ. Lang. LiteracyWritten Lang. Lit., vol. 5, no. 1, pp. 45–67, 2002, doi: 10.1075/wll.5.1.03str.

[20] K. D. Bopp and P. Mirenda, “Prelinguistic predictors of language development in children with autism spectrum disorders over four-five years,” J. Child Lang., vol. 38, no. 3, pp. 485–503, 2011, doi: 10.1017/S0305000910000140.

[21] OECD, The OECD Privacy Guidelines. 2013. doi: 10.1787/5kgf09z90c31-en.

[22] “Text Analyser,” Online-Utility.org. https://www.online-utility.org/text/analyzer.jsp

[23] Google, “Privacy Policy,” 15.10.2019, 2018, [Online]. Available: https://policies.google.com/privacy

[24] Microsoft, “Privacy Statement.” https://privacy.microsoft.com/en-us/privacystatement

[25] G. Michener and K. Bersch, “Identifying transparency,” Inf. Polity, vol. 18, no. 3, pp. 233–242, 2013, doi: 10.3233/IP-130299.

[26] M. Alić, “Model vrjednovanja informacijske transparentnosti politika privatnosti.” [Online]. Available: https://dr.nsk.hr/en/islandora/object/ffzg%3A7203

[27] B. Zieni, “Software Requirements Engineering for Transparency,” University of Leicester, 2021. [Online]. Available: https://s3-eu-west-1.amazonaws.com/pstorage-leicester-213265548798/31619780/2021ZIENIBPhD.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAILQQVTTAWSUNAFFA/20220423/eu-west-1/s3/aws4_request&X-Amz-Date=20220423T143746Z&X-Amz-Expires=10&X-Amz-Sign

[28] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, “Information accountability,” Commun. ACM, vol. 51, no. 6, pp. 82–87, 2008, doi: 10.1145/1349026.1349043.

[29] D. Hess, “Social Reporting and New Governance Regulation,” Bus. Ethics Q., vol. 17, no. 3, pp. 453–476, 2007, doi: 10.5840/beq200717348.

[30] C. Ball, “What Is Transparency?,” Public Integr., vol. 11, no. 4, pp. 293–308, 2009, doi: 10.2753/PIN1099-9922110400.

[31] 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.

[32] European Commission, “Guidelines on transparency under GDPR,” pp. 1–40, 2018, [Online]. Available: http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227

[33] Europska komisija, “Pišimo jasno,” pp. 1–16, 2011.