HAW HAMBURG

MASTERTHESIS
Paul Schwarz

Readability and Vagueness in Privacy Policies: A Multi-Country Analysis

FAKULTÄT TECHNIK UND INFORMATIK
Department Informatik

Faculty of Computer Science and Engineering
Department Computer Science

HOCHSCHULE FÜR ANGEWANDTE
WISSENSCHAFTEN HAMBURG
Hamburg University of Applied Sciences

Paul Schwarz

Readability and Vagueness in Privacy Policies: A Multi-Country Analysis

Masterarbeit eingereicht im Rahmen der Masterprüfung
im Studiengang *Master of Science Informatik*
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Marina Tropmann-Frick
Zweitgutachter: Prof. Dr. Boštjan Brumen

Eingereicht am: 25. April 2022

Paul Schwarz

Thema der Arbeit

Readability and Vagueness in Privacy Policies: A Multi-Country Analysis

Stichworte

Datenschutz, Datenschutzerklärungen, Lesbarkeit, Vagheit, Gesundheitswesen, Mobile Anwendungen, Analyse von Rechtsdokumenten

Kurzzusammenfassung

Datenschutzrichtlinien dienen dazu Nutzer darüber zu informieren, wie ihre personenbezogenen Daten von Unternehmen erfasst und verarbeitet werden. Im Idealfall können dadurch Nutzer in voller Kenntnis der Sachlage entscheiden, ob sie einem Dienst im Internet vertrauen, ihn nutzen und der beschriebenen Datenverarbeitung zustimmen wollen. Andererseits verfolgen Unternehmen beim Verfassen von Datenschutzerklärungen noch ein weiteres Ziel, sie versuchen sich gegen mögliche Rechtsstreits abzusichern.Der dabei potentiell entstehende Interesssenskonflikt kann zu komplexen und auch vagen Formulierungen führen, worunter die Lesbarkeit und Nachvollziehbarkeit der Datenschutzerklärungen leiden. Die vorliegende Thesis untersucht englischsprachige Datenschutzerklärungen von Applikationen der Kategorien "Medizin" und "Gesundheit & Fitness" aus dem Google Play Store von 10 westlichen geprägten Ländern. Dabei wird unter anderem die Häufigkeit von vagen Formulierungen bestimmt und mithilfe gängiger Lesbarkeitsindizes für jede untersuchte Datenschutzerklärung das zum Verständnis benötigte Bildungsniveau ermittelt. Aus dem resultieren Datensatz lassen sich signifikante Unterschiede in Bezug auf Lesbarkeit und Vagheit zwischen den Datenschutzerklärungen verschiedener Länder feststellen. Diese Ergebnisse wurden darüber hinaus in Form eines Dashboards visuell aufbereitet und öffentlich verfügbar gemacht.

Paul Schwarz

Title of Thesis

Readability and Vagueness in Privacy Policies: A Multi-Country Analysis

Keywords

Privacy, Privacy Policies, Readability, Vague Language, Healthcare, Mobile Applications, Legal Document Analysis

Abstract

Privacy policies are legal documents used to inform users about how their personal data is collected and processed by companies. Ideally, users can then make an informed decision about whether or not to trust and use a service on the Internet, sharing their personal information in the way described by the privacy policy. On the other hand, companies pursue another goal when writing privacy policies: they try to protect themselves against possible lawsuits. The potential conflict of interest can lead to complex and vague formulations when writing privacy policies, which in turn affects the readability and comprehensibility of privacy policies. This thesis examines English privacy policies of applications in the categories 'Medical' and 'Health & Fitness' from the Google Play Store of 10 Western-influenced countries. In the process the frequency of vague lexical items and the grade level required to comprehend each privacy policy are determined using common readability indices. From the resulting dataset, significant differences in the readability and vagueness of privacy policies originating from different countries can be identified. These results are then visualized by implementing a dashboard which was made publicly available.

Contents

List of Figures

List of Tables

Acronyms

ANOVA Analysis of Variance.

ARI Automated Readability Index.

CNN Convolutional Neural Network.

GDPR General Data Protection Regulation.

HTTPS Hypertext Transfer Protocol Secure.

NLTK Natural Language Toolkit.

SMOG Simple Measure of Gobbledygook.

WEIRD western, educated, industrialized, rich, democratic.

XSS Cross-Site Scripting.

1 Introduction

Personal data is the gold of our time. Some of the largest corporations around the globe, like Google and Meta (formerly Facebook) have made the collection and analysis of personal data their absolute core competence and made a fortune with it. Among the personal data, medical data takes a very prominent position. On the one hand, with medical data you can learn a lot about diseases, like how they spread and who is susceptible to them. You could even try to predict who will likely get a certain disease if you collect enough data from an individual and compare it to a large pool of similar data collected from the society. Such tracking of medical data could be key to improving the lives of individual patients as well as society as a whole. [2] On the other hand, it could also be the key to a more dystopian future where people are disadvantaged because of potential future health problems.

To avoid any of this dystopian fiction even today, awareness has to be raised that not all personal data which can be collected should be collected. Secondly, users need to be more aware of what data they are giving up control over. Furthermore, personal data and medical data in particular should be anonymized wherever subsequent data processing permits. [2] To give an example of unnecessary tracked personal information, some of the most popular apps included in this thesis' dataset openly track private data like a person's sexual activities, even when the purpose of the app is to track a person's weight or to teach gymnastics exercises[1]. To prevent companies from arbitrarily collecting all kinds of data and handling it as they see fit, some countries around the world have set laws on how to respect the privacy of their citizens. One of the most known and most recent privacy laws made is the GDPR. According to the GDPR for example, data related to a person's sex life is considered as *sensitive personal data* and therefore to be treated extra carefully by businesses dealing with customers in countries where the GDPR applies. [14] And while the GDPR was a big step towards a more consistent jurisprudence regarding privacy and data protection, we are still a long way

[1] https://play.google.com/store/apps/details?id=splits.splitstraining.dothesplits.splitsin30days

from privacy being adequately respected everywhere in the world. Being aware that it is not guaranteed that privacy policies reflect reality, this thesis hopes to take a closer look towards the actual privacy situation in different places around the globe, by looking at said privacy policies instead of the local laws.

1.1 Approach

To investigate the readability and use of vagueness in the privacy policies of different countries, valid ways to measure readability and the occurrence of vagueness in texts was looked up in background literature and recent white papers and articles. This is because only if a variable is measurable, it can be compared between different samples, or in this case countries. To this end, five popular readability formulas were used, some of which had been established since the 1950s and are still in use today. For example, in several U.S. states legal documents, such as insurance policies must not exceed the educational level of a ninth grader, even since the 1980s. The purpose of this quality gate is to include the majority of the population. [36] The same procedure was used to determine the amount of vague language per privacy policy. Several studies were sighted that dealt with the measurement of vagueness in legal documents and some even dealt with privacy policies. Their approaches ranged from manual annotation, to automatic counting of vague lexical items, to the usage of machine learning. Those approaches were analyzed in terms of their success rate and their resulting datasets of vague language were slightly adapted for this thesis.

1.2 Research Objective

The motivation behind this work was to get a sense of the extent to which companies' intentions for handling their customers' data differ from country to country. During that, the main aim was to find out to what extent the communication of data practices between companies and customers differs from country to country. At this point, it should be mentioned that this thesis exclusively examines whether privacy policies are formulated more or less vaguely in some countries. The question of why this is the case can only be guessed at marginally based on the tests performed.

In pursuing this goal of the thesis, it is a balancing act not to write a guide for organizations on how to choose a hosting location where privacy and transparency are kept to a minimum. But since this thesis mainly includes so called western, educated, industrialized, rich, democratic (WEIRD) countries [22], the risk of creating such an abusable guide seems to be postponed until some future work extends this thesis by many more countries. Nevertheless, at this point it is trusted that the readers of this thesis will recognize the purpose of this work in identifying the places with potential for improvement and also to highlight positive examples that others can orient themselves on.

1.3 Chapter Preview

- **Chapter 2. Theoretical Background** introduces ambiguity and classifies vagueness among these forms. After looking at the reasons for companies to use vagueness in their privacy policies, this chapter explains the most common formulas to calculate the readability of text samples. Lastly, some of the more important statistical methods applied later in this thesis are briefly explained.

- **Chapter 3. Related Work** summarizes currently released articles and papers which for example analyzed rhetorical patterns in privacy policies as well as a taxonomy to split the sentences of any privacy policy into logically related segments. In addition, this chapter dives deeper into the various subcategories of vague terms and presents analyses that examined vagueness in legal documents and privacy policies. Finally, this chapter closes with notable approaches that tried to improve or even replace the current form of privacy policies.

- **Chapter 4. Methods** describes the general approach to data collection, data processing and data visualization. Among other things, it describes how the country of origin of the privacy policies was determined, which Python libraries were used and how the final result of this thesis, an interactive dashboard, was designed.

- **Chapter 5. Results** presents the results of the applied readability formulas and the calculated total percentage occurrence of vague language in the privacy policies investigated in this thesis. Furthermore, the differences between shorter and longer privacy policies are discussed, and deviations of individual results from readability formulas are presented.

- **Chapter 6. Discussion** sums up and interprets the results from the previous chapter and compares them to the results from similar studies presented in the Related Work chapter. At this point, unexpected results are also discussed, such as the finding that shorter privacy policies were not automatically more readable and precise than long privacy policies.
- **Chapter 7. Conclusion** wraps up the significance of the findings of this thesis. Subsequently, the limitations of this thesis are summarized and evaluated, such as small amount of privacy policies studied compared to the totality of privacy policies on the Internet. In the last section, possible further research goals for future continuations of this work are discussed.

2 Theoretical Background

2.1 Ambiguity in Natural Language

Before diving deeper into the vagueness and its forms we will have to take a short look at ambiguity, which is basically the super category of vagueness. According to Massey et al. [34] ambiguity emerges when a statement is missing relevant information, or when a phrase allows multiple interpretations and the reader has to guess which meaning the author intended. Linguists generally address vagueness as a form of ambiguity. This thesis will stick to the definitions of Massey et al., who created a taxonomy for ambiguous natural language. Note that their work is specifically written from the perspective of software engineers with the goal of simplifying the creation of software in compliance to legal texts. Massey et al. therefore defined six categories of ambiguity in natural language of which Table 2.1 provides a quick overview.

Lexical ambiguity occurs when a word or phrase has multiple valid meanings. In the example 'Alice walked to the bank.' one does not know whether Alice walked to a bench, a riverside, or a financial institute, without getting further context. [34]

Syntactic ambiguity occurs when sequences of words do have more than one valid grammatical parsing. In the example 'Quickly read and discuss this document.' its not certain if 'quickly' refers to only one verb or both. One does not certainly know if the task is to hurry up with both, reading and discussing, or if it is allowed to have a in-depth discussion. [34]

Semantic ambiguity occurs when a sentence has multiple possible interpretations based on the surrounding context. Each word of the sentence has a unique meaning and the sentence has a single parse tree, but the correct interpretation of the sentence,

Table 2.1: Categories of ambiguity reproduced from Massey et al. [34]

Category	Definition	Example
Lexical	A word or phrase with multiple valid meanings	Alice walked to the bank.
Syntactic	A sequence of words with multiple valid grammatical interpretations regardless of context	Quickly read and discuss this document.
Semantic	A sentence with more than one interpretation in its provided context	Alice and Bob are married.
Incompleteness	A grammatically correct sentence that provides too little detail to convey a specific or needed meaning	Combine flour, eggs, and salt to make fresh pasta.
Referential	A grammatically correct sentence with a reference that confuses the reader based on the context	There are many reasons why lawyers lie. Some are better than others.
Vagueness	A statement that admits borderline cases or relative interpretation	Alice is tall.

however, requires more context. Two example sentences would be 'Alice and Bob are married.' or 'Bob kissed his wife, and so did Carl.'. The reader requires more context to determine whether Alice and Bob are married to each other or separately. One also does not know if Bob has a reason to be upset. [34]

Incompleteness occurs when a statement does not provide sufficient information to allow a single clear interpretation. In the provided example sentence, 'Combine flour, eggs, and salt to make fresh pasta.' some necessary information is missing such as quantity of materials and working steps to be applied. [34]

Referential ambiguity occurs when a phrase does not have a clear reference. A simple example where referential ambiguity is happening in everyday language is when pronouns cannot be certainly matched to an antecedent. In the example sentences 'There are many reasons why lawyers lie. Some are better than others.' one cannot certainly tell if 'some' refers to the reasons or the lawyers. [34]

Vagueness occurs when a term or statement admits borderline cases or relative interpretation. The example sentence 'Alice is tall.' is not precise, because depending on the readers own height they might agree or disagree on this claim. Another example would be the claim 'Our service generates a lot of data.', depending on whether this sentence is dropped in a small company or for example in a Silicon Valley tech giant, 'a lot of data' could mean a few gigabytes to several petabytes of data. [34]

In general, it can be said that ambiguity is not always necessarily clearly ascertainable and is in the eye of the beholder. In addition, some forms of ambiguity seem easier to automatically detect than others. For example, it seems easier to scan a text for salient word combinations like 'a lot of' that suggest vagueness than to look for evidence of incompleteness. Some forms of ambiguity appear to be rather difficult to identify automatically without using machine learning.

2.2 Vagueness in Natural Language

One of the main reasons for companies to use vague language is the requirement that their policies must be *comprehensive* which can be achieved by keeping it short by merging multiple data practices into one statement using generalizations. On the other hand, an ideal privacy policy must be *accurate* for all data practices throughout the entire company and its systems. Many organizations also do not want to have to constantly update their privacy policy or in case a data practice changes. Additionally, permanently keeping track if a privacy policy is still compliant to the current data practices, will result in more work load the more detailed a privacy policy is formulated. In order to keep this potential workload at a minimum, one strategy for organizations can be using vagueness in their privacy policy statements to cover both, their current data practices and potential future data practices. This way the privacy policy does not have to be updated in case the potential future data practices become the actual current data practices. [1]

Figure 2.1 visualizes a simple example from Bhatia et al. how a generalization of current and future data practices might look like. In the first step the user information 'shipping address' and 'ZIP code' are combined into 'address information'. In the same step the purposes 'order fulfilment' and "marketing purposes" are generalized into a vague condition 'as needed' to account for both current practices. To cover even potential future targeted advertising practices, the vague modal verb 'may' is added to the policy statement, while 'address information' is subsumed by 'location information'. [1] In the end,

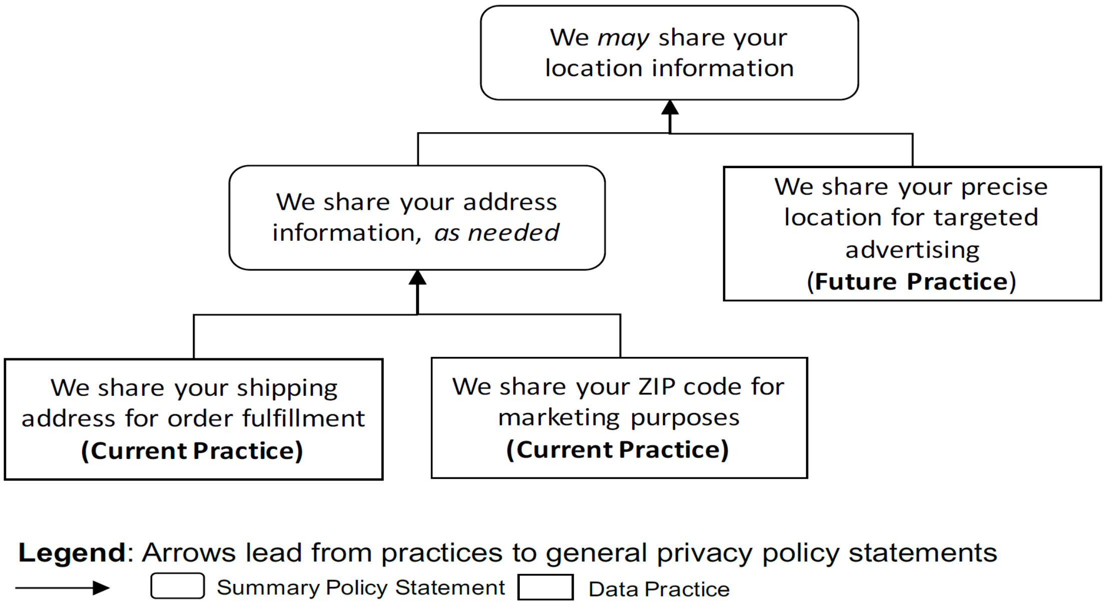

We *may* share your location information
We share your address information, *as needed*
We share your precise location for targeted advertising (**Future Practice**)
We share your shipping address for order fulfillment **(Current Practice)**
We share your ZIP code for marketing purposes **(Current Practice)**
Legend: Arrows lead from practices to general privacy policy statements
Summary Policy Statement
Data Practice

Figure 2.1: Generalizing data practices into privacy policy statements, reproduced from Bhatia et al.[1]

only one short statement remains, 'We may share your location information.' which is easier to read than the original three statements, but it is much harder for the reader to gain insight into actual data practices.

2.3 Readability in Natural Language

Readability is often confused with legibility, which means how easy it is to distinguish one letter from another in a particular font. In other words, legibility is the clarity of a typeface. Readability instead makes some texts easier to comprehend than others. [9]

George Klare defines readability as *'the ease of understanding or comprehension due to style of writing'*. [27] This definition solely focuses on writing style, and ignores issues such as content, coherence, and organization of the text.

G. Harry McLaughlin, the creator of the SMOG readability formula, defines readability as *'the degree to which a class of people find certain reading matter compelling and comprehensible'*. [38] With this definition McLaughlin emphasizes that readability depends on the interaction between the text and its readers characteristics such as literacy, prior knowledge, and motivation. [9]

The definition of readability by Dale and Chall, creators of the Dale-Chall readability formula, is even more extensive: *'The sum total (including all the interactions) of all those elements within a given piece of printed material that affect the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting'.* [8]

2.4 Readability Formulas

Since the 1920s writers like George Klare, Rudolf Flesch, Edgar Dale and Jeanne Chall experimented with using vocabulary difficulty and sentence lengths to evaluate difficulty levels of texts. Their formulas got widely used in various field like journalism, healthcare, law and insurances. Often these formulas got altered to fit better to specific fields, resulting in more and more specialized readability formulas. By the 1980s there were already over 200 different formulas for estimating the readability of natural language texts. [9] In addition to this, many formulas got updated over the time. One reason for these updates were the releases of standardized reading tests by McCall and Crabbs [35] who released and modified their tests in 1926, 1950, 1961, and 1979. Those standardized reading tests contained short stories followed multiple choice questions to evaluate ones reading skills. Some readability formulas got refined over the years to correlate as good as possible with the current release of McCall and Crabbs' standardized reading tests. [9]

This large number of often similar readability metrics leads to redundancy. [15] Therefore, only a small selection of popular readability formulas is presented in this thesis. Since even these selected formulas appear with slightly different decimal numbers in related work [13], and in some popular implementations[1], the version of each formula presented here corresponds to its current implementation, if any, in *NLTK* which is the Python library mainly used for this thesis.

[1] https://github.com/shivam5992/textstat/blob/master/textstat/textstat.py and https://github.com/cdimascio/py-readability-metrics/tree/master/readability/scorers

Assumption 1 *It is only a matter of minimal deviations in the formulas, which have presumably arisen over the years, during which they have been adapted for different applications and in part also by different authors. The exact implementation of NLTK was chosen for reference because it is presumed to be one of the most used libraries in the sighted field of related work, that also used Python as their main programming language. Furthermore the implementation in NLTK coincides with the definitions of the readability formulas compiled in 'The Principles of Readability' by William H. DuBay [9] which is one of the most outstanding books in the field.*

2.4.1 Flesch-Kincaid Grade Level

Rudolf Flesch originally released his Reading Ease Score formula in 1948 [17], which was adjusted and mapped to grade levels by Kincaid et al. [26] in 1975. The formula produces U.S. reading grade levels and is defined as follows:

$$FKG = 0.39\left(\frac{number\ of\ words}{number\ of\ sentences}\right) + 11.8\left(\frac{number\ of\ \textbf{syllables}}{number\ of\ words}\right) - 15.59 \quad (2.1)$$

2.4.2 Automated Readability Index

The Automated Readability Index was developed by Smith and Senter [41] for the U.S. Army in 1967. The formula produces U.S. reading grade levels and is defined as follows:

$$ARI = 0.5\left(\frac{number\ of\ words}{number\ of\ sentences}\right) + 4.71\left(\frac{number\ of\ \textbf{letters}}{number\ of\ words}\right) - 21.43 \qquad (2.2)$$

2.4.3 Coleman-Liau Index

Coleman and Liau [6] created their readability formula in 1975. Together with the ARI it is one of the few formulas counting letters per word instead of syllables per word. This is because computer programs were estimating syllable counts by counting vowels at the time. This estimation of syllable counts was not very accurate and took more effort than just counting letters. The formula produces U.S. reading grade levels and is defined as follows:

$$CLI = 5.88\left(\frac{number\ of\ \textbf{letters}}{number\ of\ words}\right) - 30\left(\frac{number\ of\ sentences}{number\ of\ words}\right) - 15.8 \tag{2.3}$$

2.4.4 Gunning Fog Index

The Gunning fog index was created by Robert Gunning [20] in 1952 and is based on counting polysyllables, which means words consisting of 3 or more syllables, though proper nouns and compound words are excluded in that calculation. Also, common suffixes (such as -es, -ed, or -ing) do not count as a syllable in the context of the formula. The formula calculates how many years of education are necessary to comprehend a text and it is defined as:

$$GFI = 0.4\left(\frac{number\ of\ words}{number\ of\ sentences}\right) + 100\left(\frac{number\ of\ \textbf{polysyllables}}{number\ of\ words}\right) \tag{2.4}$$

2.4.5 SMOG Index

G. Harry McLaughlin's [38] 'Simple Measure of Gobbledygook' goes commonly by its acronym as the SMOG index. The SMOG index is also based on counting polysyllables, just like the Gunning fog index, to which McLaughlin wanted to provide a simpler alternative. Because McLaughlin normed his formula on 30-sentence samples, the results for shorter text samples are usually corrected by a SMOG-Conversion table created by Harold C. McGraw [24].

$$SMOG = 3 + \sqrt{number\ of\ \textbf{polysyllables} \times \frac{30}{number\ of\ sentences}} \tag{2.5}$$

The result of the formula estimates the years of education required to understand the text to which the formula is applied to. [13]

2.4.6 Dale-Chall Formula

The original Dale-Chall formula was published by Edgar Dale and Jeanne Chall [8] in 1948. It is based on a list of basic words, 80% of which were known to fourth graders at that time. These common words got updated and extended to a list of 3,000 words in 1995 [5], with the release of the New Dale-Chall (NDC) formula. The index is based upon counting difficult words that are not contained in this basic word list. The NDC formula is defined as:

$$NDC = 0.1579 \underbrace{\left(\frac{number\ of\ \textbf{difficult words}}{number\ of\ words} \times 100\right)}_{pdw}$$

$$+ 0.0496 \left(\frac{number\ of\ words}{number\ of\ sentences}\right) + x_{\text{adj}} \tag{2.6}$$

$$x_{\text{adj}} = \begin{cases} 0 & for \quad pdw \leq 5 \\ 3.6365 & for \quad pdw > 5 \end{cases}$$

The score resulting from the NDC formula has a weak scaling and needs to be corrected for higher grades. Figure 2.2 shows the mapping between the calculated NDC scores and U.S. reading grade levels. The scores are shown on the right side of the picture. In comparison with, for example, GFI, it can be seen that NDC is not linear and has fewer gradations at higher levels of education.

2.5 Statistical analysis

A comprehensive introduction to statistical methodology is beyond the scope of this thesis. Therefore, only the statistical methods applied in this thesis will be briefly explained here.

2.5.1 ANOVA

ANOVA, an abbreviation for 'analysis of variance', is a statistical procedure that can be used to test whether the means of a measured variables are the same in different measurement groups. One typical use case for ANOVA would be to test if students from schools A, B, and C do have equal mean IQ scores. To do this, one would make the null hypothesis that all students, regardless of school, have the same IQ on average.

A subsequent ANOVA is used to test whether the null hypothesis that there are no statistically significant differences between the tested groups is valid.

There are several requirements before one can perform an ANOVA [11]:

- **Independent observations:** Measurements are independent if the measured value of one group does not depend on or is not influenced by the measured value from another group. This is arguably the most important requirement.
- **Normality:** All measurements of each group should follow a normal distribution. The importance of this requirement decreases with the sample size of each group.
- **Homogeneity:** The variances within all tested groups should be equal.
- **No outliers:** Because ANOVA is a parametric statistical test, it is not very robust against outliers, i.e. values that are far away from the mass of other values. A single outlier can make an otherwise significant result seem non-significant.
- **Continuous scaling level:** the examined variable needs to be measured at an interval or ratio level, in other words the variable needs to be continuous. Example would be the measured time, weight, force, or like in the example above the measured IQ.
- **2 or more groups:** The amount of groups to be compared should consist of at least 2 independent groups. Usually for only 2 groups one could just use an independent-samples t-test, that is why ANOVA is recommended for 3 or more groups, like the 3 schools A, B, and C from the example above.

2.5.2 Kruskal Wallis

The Kruskal-Wallis test, is basically a non-parametric equivalent to the ANOVA. Instead of using the means of a measured value, the Kruskal-Wallis test uses the mean ranks. Due to the fact that all measured values are written into an ascending list and only the rank of each entry is used for further calculation, it is not necessary for the Kruskal-Wallis test that there are no outliers and that the data is normally distributed. Beside that, all other requirements from the ANOVA are also required for the Kruskal-Wallis test. Like the ANOVA is the extension to the t-test, the Kruskal-Wallis test is the extension to the Mann-Whitney U test when dealing with more than 2 groups.[11]

2.5.3 Post-hoc Analysis

Both the ANOVA and the Kruskal-Wallis test results only identify that there is a stochastically dominant group among the investigated groups, but they do not identify which one. Therefore, after a rejection of a null hypothesis, one needs to perform pairwise comparisons to identify the group pairs that cause this stochastically dominance. These pairwise comparisons are called post-hoc analyses. Typical post-hoc tests for the Kruskal-Wallis tests are the Dunn test, the Conover test, or a Mann-Whitney U test. On top of that, the resulting p-values of the pairwise comparisons are usually corrected by either a Bonferroni, Holm, Sidak, or Holm-Sidak corrections. These corrections of the p-values are done to counteract the so called multiple comparisons problem. This means that as the number of statistical comparisons increases, the noisiness of the data increases the probability that a random test will determine a significant result. A textbook example would be a pharmaceutical company testing a new drug to see if it has a positive effect on human health. The more health stats the company tests, the more likely it is that the company finds one random health stat that increased for enough persons participating in the study. Without a post-hoc analysis, this hypothetical study could suggest that the improved health stat is caused by the drug being tested, which is false. [11]

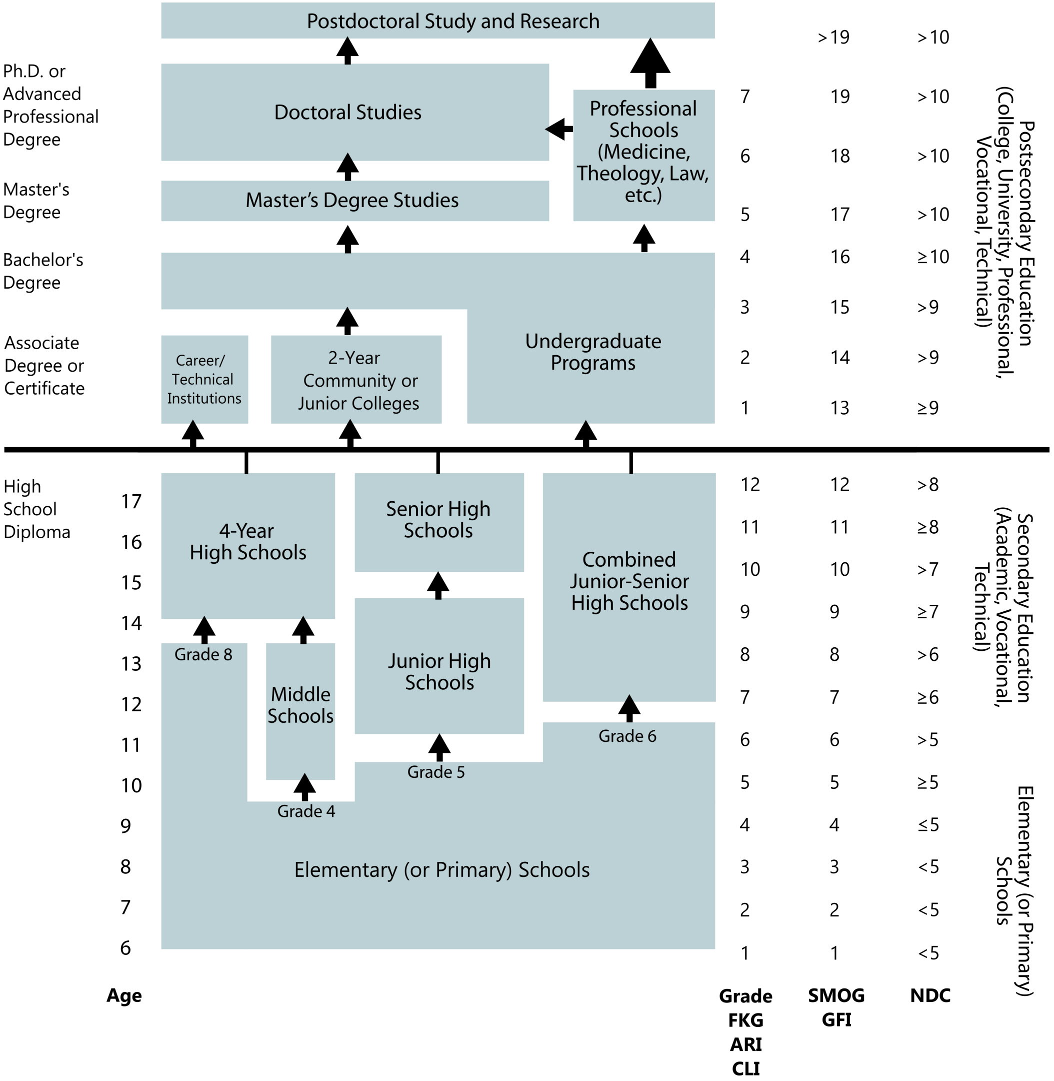

Postdoctoral Study and Research
Ph.D. or Advanced Professional Degree
Doctoral Studies
Professional Schools (Medicine, Theology, Law, etc.)
Master's Degree
Master’s Degree Studies
Bachelor's Degree
Associate Degree or Certificate
Career/ Technical Institutions
2-Year Community or Junior Colleges
Undergraduate Programs
High School Diploma
4-Year High Schools
Senior High Schools
Combined Junior-Senior High Schools
Junior High Schools
Middle Schools
Grade 8
Grade 6
Grade 5
Grade 4
Elementary (or Primary) Schools
17
16
15
14
13
12
11
10
9
8
7
6
Age
>19 >10
7 19 >10
6 18 >10
5 17 >10
4 16 ≥10
3 15 >9
2 14 >9
1 13 ≥9
12 12 >8
11 11 ≥8
10 10 >7
9 9 ≥7
8 8 >6
7 7 ≥6
6 6 >5
5 5 ≥5
4 4 ≤5
3 3 <5
2 2 <5
1 1 <5
Grade FKG ARI CLI
SMOG GFI
NDC
Postsecondary Education (College, University, Professional, Vocational, Technical)
Secondary Education (Academic, Vocational, Technical)
Elementary (or Primary) Schools

Figure 2.2: Overview of U.S. grade levels, reproduced from the U.S. Department of Education [12] and supplemented by a mapping to readability formula scores.

3 Related Work

3.1 Rhetorical Patterns in Privacy Policies

Privacy policies are legal documents generally used by providers of websites or internet applications to inform their users about how and which of their personal data is collected and processed while using their services. If a user accepts a privacy policy, the user gives permission to the service provider to manage their data exactly in the manner the privacy policy states. Therefore, to enable users to make informed decisions privacy policies should be as complete as possible. Unfortunately, this usually comes with increasing complexity and often less comprehensibility. [7]

A large share of internet users seem not to read the privacy policies of the websites and services they are using on the internet. There may be many reasons for this behavior, like the lack of alternative privacy friendly services and also the sheer amount of privacy policies that take way too much time to read. McDonald and Cranor [37] calculated in 2008 that the average American Internet user would need 201 hours per year to read all privacy policies of the services used in the same time period.

While there is only a small number of users that actually investigate the contents of privacy policies, organizations still do care about what they write into their privacy policies. Severe and unnecessary privacy invasions could be detected by one of these few readers and might lead to more social awareness. This is what organizations typically try to avoid in order to keep a positive image and keep the trust of their customers. Such trust into a company can also be harmed by privacy policy that are difficult to understand. Research indicates that such mistrust does not necessarily lead to users avoiding a website or service but to a reduction of users in general and furthermore leads users to falsify their personal data, use throwaway email accounts, and several other practices to protect their privacy. [39]

One explanation for the privacy policies being difficult to read and very complex while users would prefer short and clear statements is that website providers value their company's need to be legally protected from potential lawsuits against their data practices over the customers needs. [10] [39] Current research investigating this conflict of interests for writers of privacy policies gives interesting insights. For example, Irene Pollach's [39] vocabulary analysis of privacy policies showed that companies tend to sugar-coat their data handling practices by highlighting positive aspects and de-emphasizing privacy invasions. This sugar-coating is happening for example when customers are told that their data will only be shared with *carefully selected* organizations and that customers will only receive mail of '*great interest*' to them. Furthermore there is a trend to choose verbs that exclude companies from statements that are not in the user's interest. An example would be the usage of '*you receive*' when talking about spam e-mails instead of just saying '*we send them*'. Another option is to switch to passive structures like in the sentence '[...] for you to know which information <u>we</u> collect, how <u>we</u> use that information, and with whom <u>it</u> may be shared'. Pollach further states that one cannot safely say this de-emphasizing and omitting of self-references, when it comes to negative data practices, are caused by poor writing skills or whether they are caused by strategic usage of ambiguity to confuse the readers. Still Pollach points out that these rhetorical patterns are used but should be avoided if companies want to help users to comprehend their privacy policies. [39]

3.2 OPP-115: A Manually Annotated Website Privacy Policy Corpus

In their work 'The creation and analysis of a website privacy policy corpus', Wilson et al. [44] created a detailed set of 115 privacy policies which have been manually annotated by law school students. During this annotation process the authors grouped sentences which were related in their content into text segments. In the next step these segments were annotated with one of 10 high-level categories which you can see in Figure 3.1, illustrated as the upper shaded blocks.

The following list contains a more detailed description for each of the 10 high-level categories that can be assigned as a topic to the text segments of a privacy policy according to the taxonomy of Wilson et al.[44]:

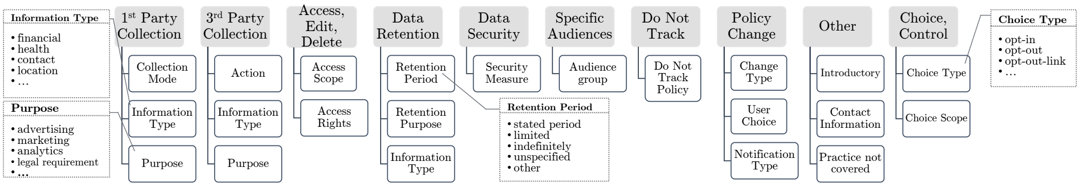


Figure 3.1: The privacy taxonomy of Wilson et al. [44]

1st Party Collection: how and why a service provider collects user information.
3rd Party Collection: how user data may be shared with or collected by 3rd parties.
Access, Edit, & Deletion: if and how users may access, edit, or delete their data.
Data Retention: how long user information is stored.
Data Security: how user information is protected.
Specific Audiences: practices that pertain only to a specific group of users (e.g., children, Europeans, or residents of a specific state).
Do Not Track: if and how Do Not Track signals e.g. for advertising are honored.
Policy Change: if and how users will be informed about privacy policy changes.
Other: additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.
User Choice/Control: choices and control options available to users.

Wilson et al. further define multiple attributes for each of the high-level categories, which basically results in a total of 122 subcategories. The white blocks in Figure 3.1 show a few of them, as well as some example values that could be assigned to these attributes. For example a text segment of a privacy policy that contains information on how long a service provider plans to keep user information collected during the usage of the service, may be tagged with the high-level category *Data Retention* but in addition it might be assigned with the attribute *Retention Period* if the text segment contains any detail about the exact duration until the user data gets finally deleted.

In Figure 3.2 one can see an example of the composition of privacy policies of five popular websites selected and annotated by Wilson et al. after their taxonomy. One can recognize that the individual content categories are often mixed up. Different websites place different emphasis on individual categories, some even do not write anything at all about individual categories. [44] This is another indicator that privacy policies have no longer a simple clarification purpose to educate the users in a straightforward way about how their data is handled, but that some privacy policies became a tangled patchwork

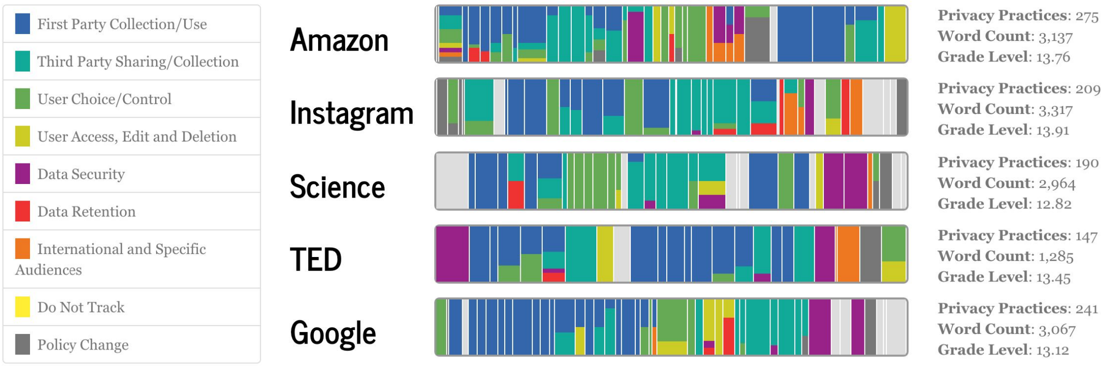


Figure 3.2: Comparison of 5 websites regarding their segments topics [44]

with the sole purpose of protecting the service provider from lawsuits. It seems like the priority is to cover all legal pitfalls for the companies rather than to create an easily readable document for the users.

3.3 Polisis: Automated Analysis of Privacy Policies Using Deep Learning

The OPP-115 dataset from Section 3.2, as well as its introduced taxonomy for assigning topics to segments of privacy policies, has since been used by many researchers as the basis for their work. One of these works is Polisis [21], which is a Convolutional Neural Network (CNN) trained with the OPP-115 annotations for automatic content segmentation of random privacy policies. It also includes a chatbot for user-friendly querying of the content of a privacy policy. Polisis absolves this segment classification in several stages. In the first step the high-level category of a segment is determined, afterwards Polisis also tries to add further attributes to the segments. These attributes are represented as white boxes in Figure 3.1. Trained on the OPP-115 dataset, Polisis is able to classify the segments of privacy policies with a precision, recall, and F-score of 66% each on average of all 10 high-level categories shown in Figure 3.1. Polisis can be used to analyze new privacy policies and enables users to quickly search for details in a privacy policy without having to read the whole policy. [21]

Table 3.1: Frequently occurring vague items in the JRC-Acquis counted by Li [31]

Semantic group	Vague items occurring 100+ times in the JRC-Acquis
quantity	some, or more, several, about, a period of, a number of, many, a list of, around, approximately, or less, (a) few, almost, a series of, a set of, a range of, a group of, an amount of, roughly, a proportion of, a volume of, a mass of, a weight of
time	(at/by) the end of, no(t) later than, at the latest, as soon as, often, at any time, sometimes, from time to time, occasionally
degree	necessary, appropriate, applicable, relevant, effective, significant, sufficient, good, adequate, suitable, reasonable, substantial, acceptable, considerable
category	such cases, such measures, such information, such (a) product(s), such data, such (a) decision(s), such (a) person(s), such (a) time(s), such materials, such notifications, such provisions, such (a) substance(s)

3.4 Counting Vague Lexical Items in Legislative Texts

Vague language is not limited to our everyday speech, but occurs in all areas where natural language is used, including legal texts where precise interpretation is arguably essential. In her papers Li [30] [31] is investigating the usage of vague language in legal documents. Therefore, Li uses the JRC-Acquis, a corpus of legal texts collected by Steinberger et al. [42]. The JRC-Acquis includes the law of all member states of the European Union. It consists of 20 different languages and contains over 8,000 documents per language, of which Li set her focus on the English version of each legal document.

Li defines four semantic groups for frequently occurring vague lexical items in the JRC-Acquis. Depending on whether a sequence of words is vague regarding ’quantity’, ’time’, ’degree’, or ’category’ Li maps these four groups which are listed in Table 3.1 together with further examples.

In addition, Li split all the documents included in the JRC-Acquis into four time periods (T1: 1958-1979, T2: 1980-1989, T3: 1990-1999, and T4: 2000-2006) to measure the occurrence of each vague item over the decades. With this work, Li shows the unexpected high frequency of vague language even in legislative texts, which became even more noticeable over the years. Furthermore, Li contributes a corpus of vague lexical items that can be applied to all kinds of natural language documents, such as privacy policies. [31]

3.5 Predicting Vague Language in Privacy Policies

A different approach of detecting vague language than Li's method of counting the occurrence of predefined vague terms was chosen by Liu et al. [33] and Lebanoff and Liu [29]. In both papers the authors state that while considering if a sequence of words is vague one needs to know the context, meaning one needs to pay attention to the sentence around the potential vague lexical items. For example, in "Users may post to our website" the word may indicates a permission but not a possibility, thus the sentence is not considered vague. [29]

To include the context of each potential vague term, Liu et al. [33] tokenized the individual privacy policy sentences using Word2Vec to receive vector representations of each word of the privacy policy. These vectors are fed to a deep neural network and do encode the semantic and syntactic aspects, as well as the potential vagueness of each word. Since their detection of vague terms is still not fully automatic and remains untested for larger datasets, their work is extended by Lebanoff and Liu [29]. This second study uses a corpus of 100 Privacy Policies which gets manually annotated by skilled native English-speaking readers. To keep the annotation workload at a minimum, candidates for vague sentences are being filtered from these 100 privacy policies using a list with cue words for vagueness collected by Bhatia et al. [1]. Each of the filtered 4.5K sentences was presented to five random annotators which rate the vagueness-level of each sentence. The authors chose to use a bipolar scale from 1 (extremely clear) to 5 (extremely vague) to let the users rate the already as probably vague identified sentences. Furthermore the annotators were asked to mark all words they considered vague in each sentence.

Figure 3.3 shows the results of the annotated sentences. The left diagram shows the percentage breakdown of the annotated sentences according to the number of vague words they contain. Note that if the annotators were able to select 0 vague words if they did not regard a sentence as vague. On the right diagram we see the percentage breakdown of the average vagueness scores the annotators mapped to the privacy policy sentences. The average sentence vagueness score is 2.4±0.9. [29]

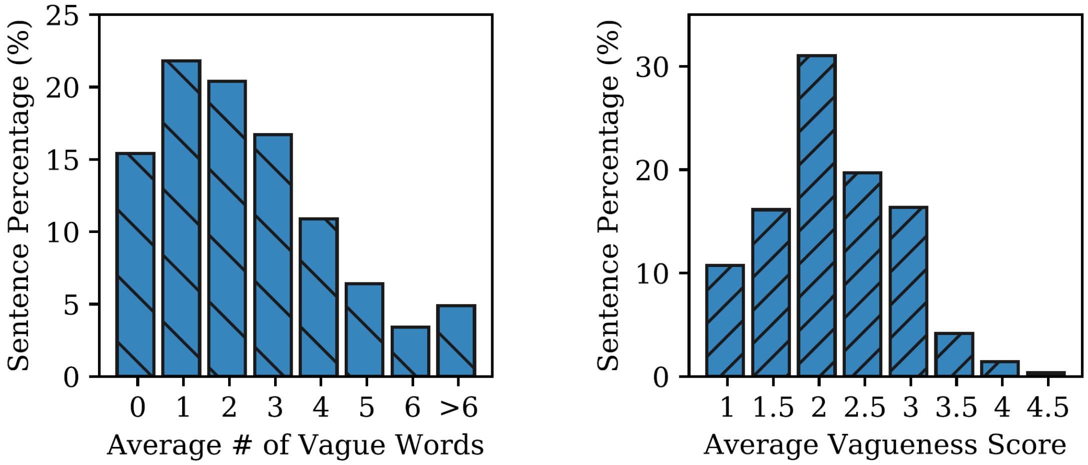


Figure 3.3: (Left) percentage of sentences containing different numbers of vague words. (Right) perc. of sentences with different levels of vagueness. From Lebanoff and Liu [29]

Lebanoff and Liu create several neural networks and receive interesing results. Especially the comparison of their created context-aware (68% precision, 54% recall, and an F-score of 60% while detecting vague words) versus context-agnostic classifier (11% precision, 78% recall, and an F-score of 20% while detecting vague words) suggests that context information is necessary for detecting vague words. Another lesson learned is that even with context information it seems quite difficult to accurately detect vague lexical items automatically. Lebanoff and Liu further note that the biggest source for false positives were adjectives with 37.2% and nouns with 35.2% of all false alarms. On the other hand, for the misses, 47.6% were nouns and 25.1% were adjectives. The reason why it is especially hard for adjectives and nouns to decide whether they are vague or not is probably because their vagueness heavily depends on their context. [29]

One remarkable result is that in 47.2% of the cases 3 or more annotators agreed with their vagueness score of a sentence and in 12.5% of the cases 4 or more annotators chose to the same score. This indicates that humans do have their difficulties determining vague language and its degree of vagueness. Another important thing to note is that around 15.5% of the sentences which contained at least one word of Bhatia et al.'s vague cue word list [1] are not considered vague anymore after being manually checked by the annotators. This gives one an idea about the rate of false positives one should expect using methods by Li [31], described in Section 3.4, to detect vague language.

3.6 Readability in Privacy Policies

There are many papers investigating readability in all kinds of publicly available documents throughout the Internet. The studies of Ermakova et al. [13] and Fabian et al. [15] focus explicitly on analyzing the readability on large scale privacy policy data sets. Both papers used commonly accepted readability indices like Gunning Fog, Coleman-Liau or the Flesch-Kincaid grade level. Ermakova et al. collected and analyzed 5,000 privacy policies from healthcare websites and e-commerce websites. In their comparison they conclude that policies of healthcare websites are significantly shorter and provide significantly better readability of their privacy policies than top e-commerce websites. Furthermore in their comparison Ermakova et al. conclude hat commercial and non-commercial healthcare websites have similarly long privacy policies, but commercial healthcare websites contain more readable privacy policies. [13]

Fabian et al.[15] created a corpus of 50,000 privacy policies but did not distinguish between the different purposes of the websites. Instead, they added another dimension to the analysis by calculating the readability indices for different top-level domains. This way, one can perform comparisons between different types of organisations like comparing the readability data of educational entities (.edu) with governmental entities (.gov). In the end, both papers come to similar conclusions: the average education level expected to comprehend privacy policies lies between 13 and 16 years of education, which is alarmingly high. [13] [15]

On the side of smaller scale studies, there are a few papers that already examined the readability in combination with vagueness in privacy policies. Cadogan [4] compared the privacy policies of 3 organizations and Krumay and Klar [28] compared the privacy policies of 15 organizations. For their analysis both used a software named Wordcount[1] which calculated common readability indices like Gunning fog index and the Flesch-Kincaid grade level, as well as percentage values for the usage of vague words out of the box. Both papers mention a threshold of 2% of vague words as recommended by Wordcount to avoid an impression of uncertainty and a lack of clarity [4]. There is however no explanatory note for this exact value available.

[1] https://number27.org/wordcount

Unfortunately the tool Wordcount was written in the early 2000s, using the discontinued Adobe Flash Player[2]. These compatibility issues make Wordcount unusable for the further research of this thesis.

3.7 Transparency Reports on the Adoption of Privacy and Security Enhancing Technologies on the Web

Felt et al. [16], a group of security experts from Google Chrome's and Mozilla Firefox's web browser teams, wrote a paper looking at privacy on the Internet in the opposite direction, focusing on the users instead of companies. Instead of looking at privacy policies, which represent the claims of an organization to value a user's privacy, Felt et al. examined the websites visited by users and reviewed the technical measures in place to protect the privacy of users of these websites. To check the websites security measurements the team used Mozilla HTTP Observatory[3], a tool for checking how sophisticated a website is securing the communication with its users from a technical point of view. Mozilla HTTP Observatory therefore applies several tests, e.g. for HTTPS-support or Cross-Site Scripting (XSS) vulnerabilities, to a requested website and grades the website depending on its performance in these tests.

Having access to huge amounts of anonymized user data from Chrome and Firefox, Felt et al. are able to compare the adoption of essential security standards like HTTPS around the globe, discovering regional disparities. Figure 3.4 is taken from Felt et al. and shows the median rate of HTTPS usage by country in February 2017 among Firefox users. It is noticeable that some smaller countries perform exceptionally well in this study. The 75th percentile of the measured HTTPS usage rate is above 90% in Libya, Syria, Venezuela, Ecuador, and Iraq. Felt et al. hypothesize that internet users in small countries are mainly surfing on advanced websites like Google and Facebook which support HTTPS by default, probably due to the lower amount of local alternative web content.

Table 3.2 shows an excerpt of the exact results measured by Felt et al. [16]. Looking at the larger countries, the best performing countries regarding the HTTPS usage are the United States of America, followed by Mexico and India. Among the lower end outliers

[2] https://www.adobe.com/products/flashplayer/end-of-life.html

[3] Mozilla's HTTP Observatory is written by April King, former Head of Web Security at Mozilla and one of Felt's co-authors [16]. It is publicly available under https://github.com/mozilla/http-observatory or https://observatory.mozilla.org/ if you want to use the web interface

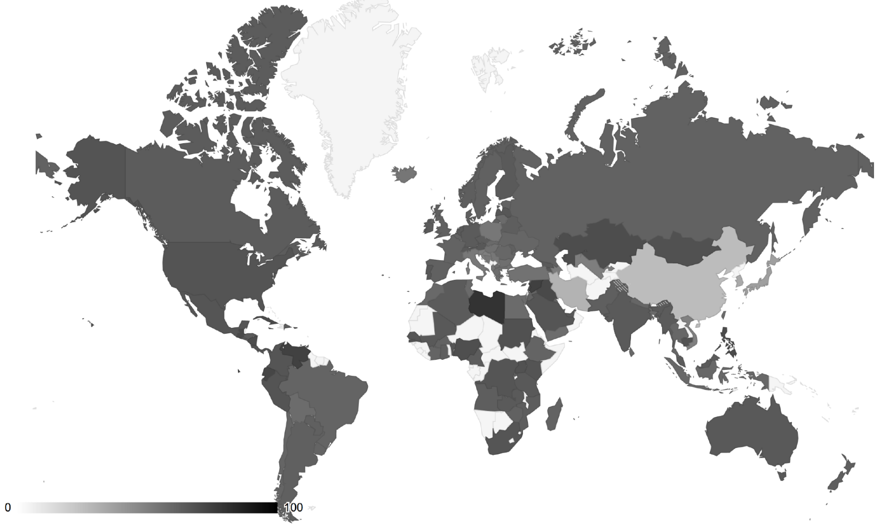


Figure 3.4: median rate of HTTPS usage by country, from Felt et al. [16]

are some East Asian countries, for example China, South Korea and Japan. In the case of China, the authors speculate that their low HTTPS adoption rate is due to the Great Firewall. The case of Japan and South Korea, however, requires further investigation to provide an educated guess for the cause of their poor performance. [16]

Focusing mainly on the 'consumer side' through measuring the user's browsing behavior, Felt et al. [16] also have a look at the 'producer side' by measuring the HTTPS availability in each country. Their results can be seen as well in the right columns of Table 3.2. The authors again realized regional disparities although just comparing a rather small number of websites, as they limited their research on the Top 100 most popular websites per country. For a global perspective the authors also investigated the worldwide top 1 million websites whose HTTPS support grew from 30% in February 2016 to 40% in February 2017. At the same time the worldwide amount of websites which use HTTPS per default grew from 5% to 10%. [16]

This latter and more comprehensive comparison on the server side provides a more detailed insight into the actual global data security situation, than just looking at the top 100 most popular websites. It would be very interesting to expand the amount of ana-

[4] Median HTTPS usage rate of Firefox users in February 2017 by country

[5] The percentage amount refers to the Alexa Top 100 websites in the respective country

Table 3.2: Global comparison of HTTPS availability and adoption by Felt et al. [16]

Country	Median HTTPS usage[4]	HTTPS support[5]	HTTPS default[5]
France	61%	67%	16%
Germany	64%	86%	27%
India	65%	68%	16%
Japan	37%	57%	19%
Mexico	66%	80%	19%
Russia	61%	80%	24%
South Korea	33%	75%	14%
United States	67%	81%	18%

lyzed websites per country to receive further insight into the data security development of the individual countries.

3.8 Notable Approaches to Enhance Privacy Policies

As mentioned in Section 2.2 one of the major excuses for the usage of generalizations and thus vague language in privacy policies is the reduced workload that an organization has when it is not forced to constantly update its privacy policy as soon as a data practice changes. However, using vague language in privacy policies to reduce an organization's workload has several downsides, both for its clients and for the organization itself. The downside for users is that they receive an inaccurate privacy policy and therefore cannot be sure what is happening with their private data, which ultimately increases users' overall perception of risk when reading these privacy policies [1], which in turn might damage an organization's public image and could lead to a decline in user numbers.

In the past, there have been several projects to address these downsides, though the following projects are either already discontinued or only merely developed, yet they still provide interesting approaches to design privacy policies using less vague language.

Using Privacy Icons to summarize data practices with one view is an idea proposed by many researchers [25] [23] and by the European Union in the context of GDPR. [19] The idea addresses the problem that the average internet user is rarely willed to invest time into reading privacy policies. Unfortunately in practice, internet users also tend to ignore cues that try to warn and keep the user secure while browsing the web as Whalen and Inkpen [43] analyzed. For their study Whalen and Inkpen used eye-tracking data to monitor the effectiveness of security risk cues like Chrome's and Firefox's padlock icon which indicates if a site uses HTTPS. While the icon was commonly recognized, users rarely interacted with the icon to receive further certificate information. The more security-savvy users were more likely to stop their current actions when confronted with warnings, while most users did not understand the cues or just ignored them once they logged into a website. [43] Another study by Friedman et al. [18] showed that Internet users are often unable to recognize secure connections when surfing. Of the various technologically educated groups, only participants from the high-technology community managed to recognize insecure connections reasonably accurately in the Friedman et al. study.

Although both studies, Friedman et al.'s and Whalen and Inkpen's, were conducted in the early 2000s, and the overall security awareness of Internet users may be better today, too large a proportion of users will likely always have difficulty interpreting the risks to Internet security and privacy, or simply ignore investing time in educating themselves about these risks on an ongoing basis, even if it could be done by looking at a few icons. Still, reducing the needed invested time to keep up a reasonable security and privacy on the web, will be a great benefit for the majority of Internet users. For example, nowadays security-savvy users are able to use browser plugins like HTTPS-Everywhere[6] which automatically redirects the users to the HTTPS-version of their visited websites. An identical optional feature later also got implemented into the web browsers Firefox (Version 83) and after that into Chrome (Version 94). Such default settings that improve privacy or security on the web without investing a lot of time are a great benefit for the Internet users, which is why in theory the following approach P3P was a promising solution from the user perspective.

P3P stands for Platform for Privacy Preferences Protocol[7] which is an outdated protocol developed by the World Wide Web Consortium (W3C) in the early 2000s. This

[6] https://github.com/EFForg/https-everywhere

[7] https://www.w3.org/TR/P3P11/

protocol allowed website providers to encode their intended use of the user's personal data into machine readable code. This way, users could easily recognize what was going to happen with their data, even when surfing on foreign language websites. In addition, the protocol allowed users to predefine default rules for what kind of processing of their various personal data they agreed to. The Idea was to set these rules once and ideally never have to look at a privacy policy again, which suits the majority of internet users who do not bother to waste time on reading privacy policies. [32] Supporting P3P meant a lot of extra work and barely any benefits for website providers, also P3P was never adopted by other browsers than Microsoft Internet Explorer and Microsoft Edge, as it was considered too complex for the average user.

AutoPPG is a tool created by Yu et al. [45] which performs automatic privacy policy generation from the source code for Android applications. AutoPPG first performs various static code analyses to understand the internal processing of personal data. Then, natural language processing is used to generate correct and understandable sentences to describe these behaviors. The authors of AutoPPG even validated their software comparing existing privacy policies of apps with their newly generated privacy policies. This comparison was done by a small group of human readers who ended up rating the automatically generated privacy policies on average as more readable and understandable. Although very promising, this approach also has its weaknesses that could prevent companies from implementing it. First of all, AutoPPG only generates privacy statements based on the apps source code. If an organization has data practices that range over several systems, sharing user data from one internal service to another, then AutoPPG is not able to keep track of that. Second, the automatic text generation, like all software, is not error prone and needs manual validation. And third, a privacy policy consists of more than just data practices, there will always be manual additions to the privacy policy text needed e.g., about the user's personal rights or how to contact the developers of an application. After all, AutoPPG is not a commercial product but yet a promising scientific approach. [45]

4 Methods

Since at the early stages of this thesis the scope and requirements for a possible visualization tool of the data were not exactly defined, the programming language Python was chosen for the implementation of the thesis to be as flexible as possible later on. Most of the tasks could be performed using Python, but in some cases other more suitable software was used, e.g. SPSS[1] for validating statistical significance results in the early stages.

The following sections provide an insight into the procedures and tools used for data selection, data collection and data processing. Figure 4.1 provides a brief overview of the selection process. Of the 3,992 privacy policies scraped, 713 made it into the final dataframe used for data visualization.

4.1 Data Collection

For the scope of this thesis, privacy policies of applications from the Google Play Store[2] were targeted. Furthermore, this thesis limits itself through focusing only on privacy policies of applications from the Google Play Store categories 'Medical' and 'Health & Fitness'. For both of these application genres, the top 200 most downloaded apps from the Google Play Store of 10 countries were gathered, including the United States of America, Canada, the United Kingdom, Australia, and few western European countries. This resulted in around four thousand privacy policies, as seen in Figure 4.1. The focus was set intentionally upon English-speaking countries to make the results more comparable. To retrieve the initial dataset, the Node.js application *google-play-scraper*[3] was used. This application manages to easily filter Google Play Store applications and scrape their metadata. Among this metadata collected from the Google Play Store are the apps

[1] https://www.ibm.com/analytics/spss-statistics-software
[2] https://play.google.com/store
[3] https://github.com/facundoolano/google-play-scraper

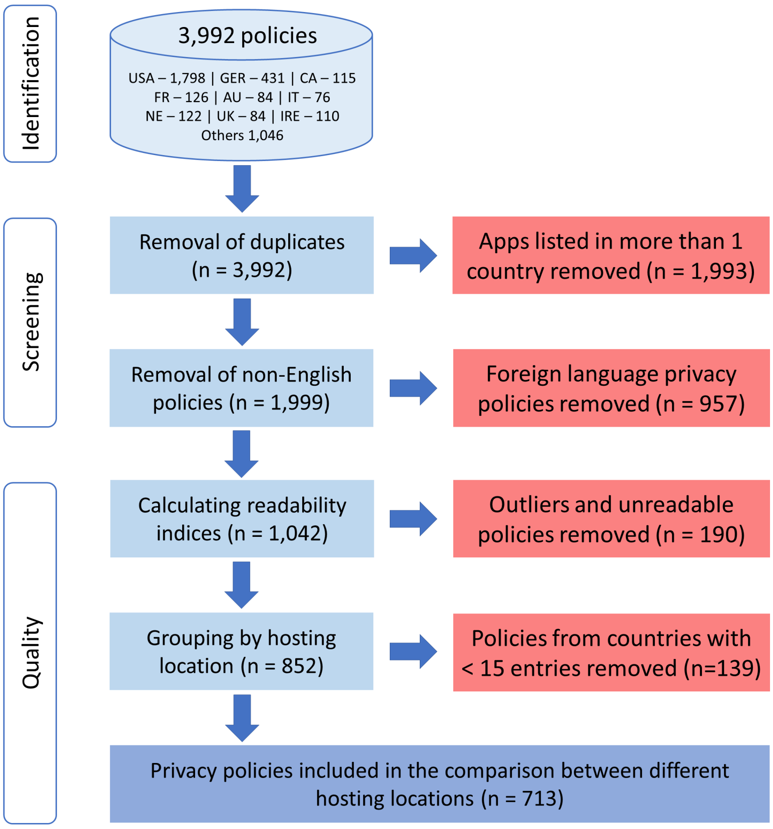

Identification
3,992 policies
USA – 1,798 | GER – 431 | CA – 115
FR – 126 | AU – 84 | IT – 76
NE – 122 | UK – 84 | IRE – 110
Others 1,046
Screening
Removal of duplicates (n = 3,992)
Apps listed in more than 1 country removed (n = 1,993)
Removal of non-English policies (n = 1,999)
Foreign language privacy policies removed (n = 957)
Quality
Calculating readability indices (n = 1,042)
Outliers and unreadable policies removed (n = 190)
Grouping by hosting location (n = 852)
Policies from countries with < 15 entries removed (n=139)
Privacy policies included in the comparison between different hosting locations (n = 713)

Figure 4.1: Selection process of the privacy policy data

reviews, its rating, age recommendation, and several other information. On the other hand, the possible filters include the app category, language, and pricing. Unfortunately, the Google Play Store does not contain the privacy policy itself, but a URL to the website that usually should contain the privacy policy. Though the received metadata contains a field for the 'country', this field just indicates the Play Store instance where the application is available from, and is not to be confused with the country where the service is hosted. Therefore, the currently scraped metadata must still be supplemented with the hosting location and the actual full text of the privacy policies.

In order to enrich the dataset with the privacy policy texts, *Beautiful Soup*[4] was used to scrape the textual content of the websites behind the already gathered privacy policy URLs. Using *Beautiful Soup* is a very easy way to gather large amounts of privacy policies. Unfortunately, it does not only scrape the text belonging to the privacy policy,

[4] https://pypi.org/project/beautifulsoup4/

but all text found on the website. Often this can be the terms of service or just the same privacy policies written in another language. Overall, this textual noise is caused by the tradeoff of using Beautiful Soup to automatically parse HTML pages, rather than investing more time in manual scraping or using machine learning to train an agent to recognize irrelevant text which does not belong to the actual privacy policy.

Assumption 2 *In order not to exceed the time frame of the thesis, it is assumed at this point that the textual noise is similar for most of the scraped privacy policies and thus influences the following natural language processing operations equally heavy for all scraped texts. Additionally, a method to detect heavy outliers among the raw scraped texts e.g., by counting words comparing readability score results, was added later on.*

For the other missing information, the hosting location, a script was written to lookup the hosting location for each privacy policy by using the service *ipwhois.io*[5], a global library that keeps track of the geographic location behind IP addresses. In addition, a brief validation of this hosting location checker was performed by manually reading the texts of the scraped privacy policies until 20 entries were found that explicitly referred to the location of their servers in their policies. Of these 20 policies, 15 matched exactly the result provided by the script using *ipwhois.io*, 3 listed different countries, and 2 privacy policy listed the calculated country with the addition of 'and other countries'. The method used to determine the location may not be totally accurate, but most of the scraped privacy policies omit this piece of information in their natural language texts and the Google Play Store also does not provide this metadata.

Assumption 3 *Note that, while the used location checking method is usually consistent with the majority of the hosting locations described in natural language text, one cannot be absolutely certain that this is also the case for privacy policies that omit this information. At this point, it is assumed that using ipwhois.io is a proper way to automatically get a suitable indicator of where the vast majority of services are located.*

[5] https://ipwhois.io/documentation

4.2 Data Processing

4.2.1 Selection of Readability Formulas

Due to the wide range of readability formulas, as described in Section 2.4, this work will only include the following indices:

- Gunning fog index (GFI)
- Automated Readability Index (ARI)
- Flesch-Kincaid Grade Level (FKG)
- SMOG index (SMOG)
- Coleman-Liau index (CLI)

The Dale-Chall index (NDC) is intentionally not included, because it is based on a list of 3.000 words common to 4$^{\text{th}}$ graders in the year 1995. This thesis is dealing with privacy policies, which mostly emerged after the year 1995 and therefore may contain many words that would distort the predicted readability score. Other authors working with privacy policies, like Krumay and Klar [28], also mentioned that the Dale-Chall formula seems to be biased in this research field.

Combination of Readability Formulas

All selected readability indices have in common that they return a floating-point value as a result for the grade level or the years of education needed to comprehend the underlying text. Therefore, this thesis will summarize all calculated readability indices into one value which will be referred to as *Mean Readability Grade (MRG)* in the following sections. This is done reduce the amount of redundant information and having a single consistent comparative value for readability.

$$MRG = \frac{ARI + CLI + GFI + FKG + SMOG}{5} \tag{4.1}$$

4.2.2 Processing Text Statistics

To gather relevant text statistics like the word counts, syllable counts, sentence counts, the Python libraries *NLTK*[6] and *textstat*[7] were used. Both libraries also include methods to calculate the readability indices selected in Section 4.2.1. During this application of the selected readability formulas, several outliers were manually verified or deleted. For example, privacy policies that were too short to perform a proper calculation of the SMOG formula (2.5) were manually checked. Usually, these checked privacy policies just turned out to be very short and straight forward, but some scraped website texts consisted only of error codes and were therefore deleted, as seen in Figure 4.1.

In order to measure the occurrence of vague language in the scraped privacy policy texts, another method was implemented to count the occurrence of vague lexical items. To do this, the vague language corpus provided Li [31], which is shown in Table 3.1, was used. While measuring the occurrence of vague language the same distinction between the four semantic categories of vagueness ('quantity', 'time', 'degree', and 'category') used by Li, was applied. The results for each vague category were divided by the total word count of each privacy policy text to get a comparable percentage occurrence of vague lexical items for each privacy policy.

Lastly, a calculation for the mean reading time for privacy policies for each hosting location was done. As a baseline of reading speed, the findings by Brysbaert [3] were used. For native English speakers, Brysbaert's meta-analysis measured an average reading speed of 238 words per minute, 300 words per minute for higher-skilled readers, and 139 for non-native speakers.

4.3 Data Visualization

To visually represent the results that were previously only accessible by running the written Python scripts as a *Jupyter Notebook*, another goal of this thesis was to create a browser accessible dashboard that allows interactive exploration of the collected dataset.

[6] https://github.com/nltk/nltk

[7] https://github.com/shivam5992/textstat

4.3.1 Dashboard Implementation

For the implementation of the dashboard, the framework Plotly Dash[8] was chosen. Dash is one of the major frameworks supporting the creation of dashboards in Python and is built on top of the frameworks Flask, Plotly.js and React.js. It allows developers to create dashboards using only Python code, without having to touch any code of the frameworks it is built on. Also, Dash is a cross-platform solution as it renders directly in the browser.

4.3.2 Dashboard Deployment

To make the dashboard more accessible, the Python code was provided via a Github repository[9]. Initially, this repository just contained a *Docker Compose* file to enable anyone who has access to the codebase to create a *Docker Image* and run the dashboard application as a *Docker Container*. In the later stages of the dashboard development process, an automatic deployment pipeline was created using *Heroku* to make the dashboard easier accessible through a web browser[10]. This pipeline was kept very basic and simply is designed to deploy a new version of the dashboard whenever a feature branch is merged into the main branch of the Github repository.

[8] https://plotly.com/dash/

[9] https://github.com/pcschwarz/analyzing-privacy-policies

[10] https://analyzing-privacy-policies.herokuapp.com/

5 Results

One hypothesis that this thesis is trying to validate, is that the country of origin of a privacy policy should have no influence on the precision of its wording. Likewise, the country of origin should not affect the readability of a privacy policy, at least not for English speaking countries. The subsequent result sections will follow the same pattern and answer the now introduced checklist (A) to (D):

(A) A Shapiro-Wilk test to check if the investigated variable is normally distributed.

(B) A Mann-Whitney U test to check if there are differences between the scraped Google Play Store app genres Medial (MED) and Health & Fitness (H&F).

(C) A Kruskal-Wallis test to check if the distribution of the investigated variable is the same across all tested hosting locations. The Kruskal-Wallis test was chosen because step (A) usually indicated that the underlying data was skewed.

(D) Optionally, in case the Kruskal-Wallis test suggested to reject the null hypothesis, a pairwise comparison of the hosting locations was performed to identify the causing countries.

To make the results of steps (A) to (D) easier to interpret Table 5.1 gives an overview ranking of all results of the conducted linguistic tests for each pair of hosting location and app genre. Furthermore, the table shows the amount of privacy policies processed for each hosting location. Note that in the table only the mean ranks for each test are shown, the exact numbers will be presented in the following dedicated sections of this chapter.

[1] Mean reading time required by a native English speaker to read a privacy policy

[2] Percentage occurrence of vague lexical items in a privacy policy

Table 5.1: Mean ranks of hosting locations (total N=713, lower ranks are better)

Hosting Location	Genre	N	MRG	FKG	GFI	CLI	ARI	SMOG	Reading Time[1]	Vague Items[2]
Australia	MED	40	470.9	450.3	477.8	456.0	472.2	491.8	338.6	487.3
Australia	H&F	20	496.6	480.3	497.0	436.7	493.1	463.5	323.6	379.8
Canada	MED	34	385.1	383.5	380.0	443.1	367.3	407.2	259.9	289.6
Germany	MED	70	332.6	334.3	338.9	302.1	334.8	332.8	364.0	353.4
Germany	H&F	29	332.7	341.6	327.2	321.7	333.0	319.6	414.3	285.0
Ireland	MED	20	336.6	378.2	339.7	244.6	325.2	303.8	385.3	337.4
Ireland	H&F	25	295.4	300.2	314.8	272.4	308.5	294.1	377.3	389.2
Netherlands	H&F	16	332.7	330.4	334.3	293.8	334.0	350.3	318.8	306.0
UK	MED	26	384.8	379.3	407.4	273.3	396.2	382.5	375.1	451.8
UK	H&F	16	341.3	334.6	355.1	248.0	351.5	320.4	427.0	355.4
USA	MED	217	358.9	358.0	348.4	405.5	355.8	359.1	344.4	338.8
USA	H&F	200	334.8	336.5	337.3	332.8	337.6	335.2	373.2	361.4

5.1 Readability Grade Levels

In order not to overload the diagrams in this section, figures will primarily focus on the MRG results as the only indicator for the required reading skill levels. Naturally, all provided figures are also included in the finished dashboard[3] which can be used to recreate the diagrams using a readability formula of your choice.

Figure 5.1 shows the boxplots for the MRG across the hosting locations. Looking at these boxplots one can notice that the median Irish privacy policy requires a grade level of 18.7 while the median Australian privacy policy requires its readers to have 23.5 years of education to be understood. The underlying dataset indicates that Australian privacy policies require 4.8 more years of education than Irish and 4.3 more years of education than US American privacy policies to be understood.

[3] https://analyzing-privacy-policies.herokuapp.com/

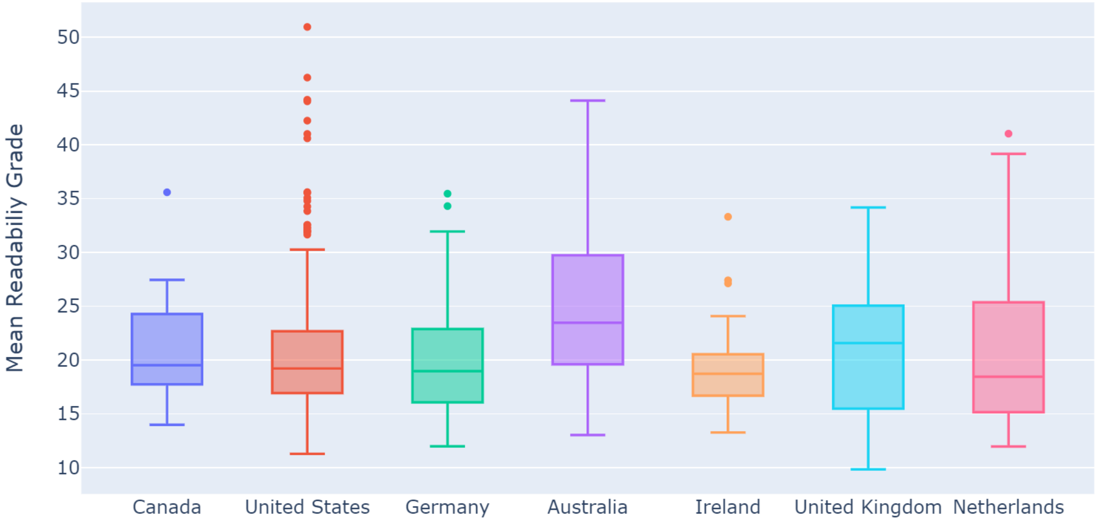


Figure 5.1: Readability Grade Levels

Keep in mind that in the data selection process combinations of app genres and countries that occurred less than 15 times were filtered out. This is the reason why the Netherlands and Canada are only shown once in Table 5.1.

Furthermore, note that all noticeable outliers have been re-checked manually, usually outliers were caused by wrong scrapes where the companies stuffed multiple pages like frequent asked questions, terms of service and multiple privacy policies for different regions into one web page. These outliers were correctly rescraped. This means that the outliers one can see in the boxplots are caused by extremely long and difficult to read privacy policies. Therefore these outliers were not removed from the dataset as they are based on correct measurements.

(A) The Mean Readability Grade is not approximately normally distributed as assessed by a Shapiro-Wilk test ($W = 0.894\,;\ p < 0.001$).

(B) According to a Mann-Whitney U test ($p = 0.096$) the distribution of the Mean Readability Grade is the same across the genres MED and H&F.

(C) According to a Kruskal-Wallis test ($\chi^2(6) = 26.475$, $p < 0.001$) the distribution of the Mean Readability Grade **is not** the same across all tested hosting locations. Hence a pairwise comparison between the hosting locations was performed in the next step.

(D) The pairwise comparison, summarized in Table 5.2, shows that there are statistical significant differences between Ireland and Australia. The mean rank of MRG for Australian privacy policies lies 165.7 ranks behind the Irish ones. Also, it is 146.8 ranks behind Germany and 132.1 ranks behind the United States. All of these cases are statistically significant with a Bonferroni adjusted p-value of 0.001 or lower.

Table 5.2: Short comparison of hosting locations regarding Mean Readability Grades (full version available in the Appendix as Table A.1).

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Ireland-Australia	165.739	40.618	4.080	.000	.001
Germany-Australia	146.814	33.698	4.357	.000	.000
Netherlands-Australia	146.763	57.953	2.532	.011	.238
United States-Australia	132.102	28.439	4.645	.000	.000
United Kingdom-Australia	111.212	41.438	2.684	.007	.153
Canada-Australia	94.362	44.213	2.134	.033	.689

5.1.1 Deviating Results for the Coleman-Liau Index

The observations on the MRG results were largely congruent with those of the other readability indices FKG, GFI, SMOG, and ARI. Only the values of CLI showed some deviations which can be seen in Table 5.3. Looking at the CLI one can recognize more statistical significances between the investigated countries than by just looking at the MRG. Just like for the MRG, the results for the CLI show statistical significances between Australia and Germany, the United States of America, and Ireland. In addition to that, Table 5.3 shows that the better results of the United Kingdom over Australia, Canada and also the United States of America are statistically significant. In the same way the additional better results of Irish privacy policies over U.S. American and Canadian privacy policies are no coincidence according to the Bonferroni adjusted p-values. Same goes for the German and Canadian privacy policies.

[3] Adjusted by Bonferoni Correction

Table 5.3: Short comparison of hosting locations regarding the Coleman-Liau index (full version available in the Appendix as Table A.2).

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Ireland-United States	-110.560	32.318	-3.421	.001	.013
Ireland-Canada	183.033	46.802	3.911	.000	.002
Ireland-Australia	189.494	40.617	4.665	.000	.000
United Kingdom-United States	-106.960	33.343	-3.208	.001	.028
United Kingdom-Canada	179.433	47.516	3.776	.000	.003
United Kingdom-Australia	185.895	41.437	4.486	.000	.000
Netherlands-Canada	149.244	62.442	2.390	.017	.354
Netherlands-Australia	155.706	57.951	2.687	.007	.151
Germany-United States	-62.746	23.027	-2.725	.006	.135
Germany-Canada	135.220	40.941	3.303	.001	.020
Germany-Australia	141.681	33.698	4.205	.000	.001
United States-Australia	78.935	28.439	2.776	.006	.116

5.2 Mean Reading Times

For all privacy policies collected, it takes a native English speaker an average of 12 minutes to read a privacy policy. Figure 5.2 shows the results of the time required to read privacy policies from different hosting locations. The application genres are mixed up in this diagram because a precalculated Mann-Whitney U test showed that there is no statistical difference between the Medical and Health & Fitness app genre when it comes to reading times.

(A) The Mean Reading Time (Native) is not approximately normally distributed as assessed by a Shapiro-Wilk test ($W = 0.871\,;\ p < 0.001$).

(B) According to a Mann-Whitney U test ($p = 0.054$) the distribution of the Mean Reading Time (Native) is the same across the genres MED and H&F.

(C) According to a Kruskal-Wallis test ($\chi^2(6) = 12.015$, $p = 0.062$) the distribution of the Mean Reading Time (Native) is the same across all tested hosting locations.

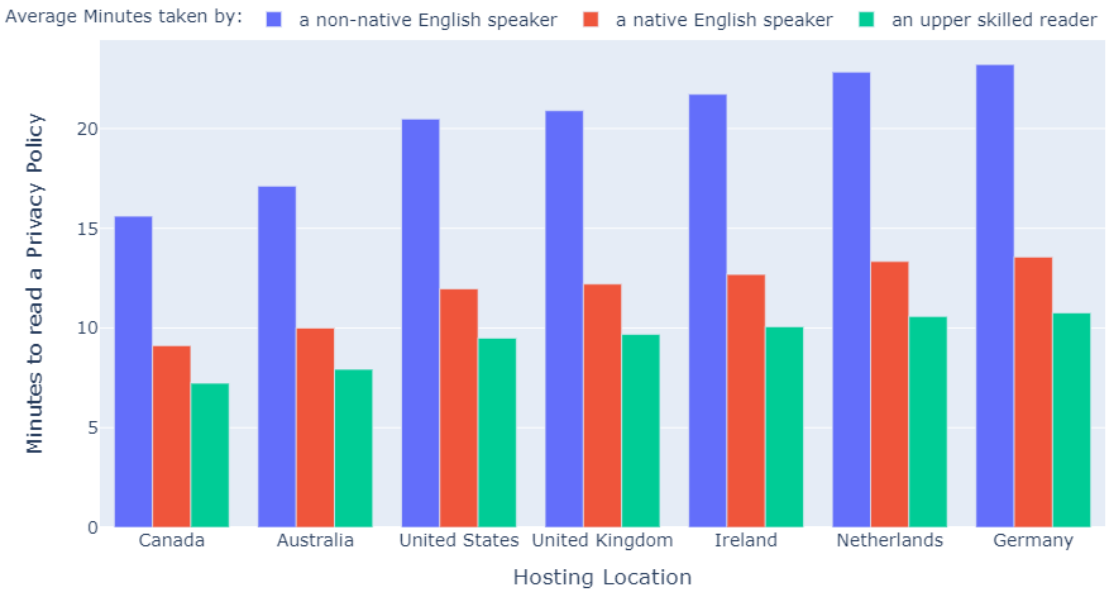


Figure 5.2: Mean reading times per privacy policy by hosting location

The largest difference in the processed data in Figure 5.2 lies between German privacy policies which take around 13.6 minutes to read and Canadian privacy policies which take on average 9.1 minutes to read for a native English speaker. Though this is a difference of 49% in required reading time, it's not statistical significant as the Kruskal-Wallis test performed in step (C) shows.

5.3 Occurrence of Vagueness

This section presents the results for the percentage occurrence of vague lexical items in the scraped privacy policies. Like the MRG focused plots in Section 5.1 this current section will only include plots for the total occurrence of all vague lexical items. Readers interested in the specific occurrence of items from individual semantic groups of vagueness, such as *quantity*, *time*, *degree*, or *category*, which were presented in Table 3.1, will find these detailed breakdowns in the final dashboard.

(A) The percentage occurrence of vague lexical items over all countries seemed normally distributed as assessed by a Shapiro-Wilk test ($W = 0.996\,;\ p = 0.110$). But for individual countries like Ireland it was not normally distributed according to a Shapiro-Wilk test ($W = 0.943\,;\ p = 0.027$). Therefore the Kruskal-Wallis test was retained in step (C).

(B) According to a Mann-Whitney U test (p > 0.054) the distribution of the percentage occurrence of vague lexical items is the same across the genres MED and H&F.

(C) According to a Kruskal-Wallis test ($\chi^2(6) = 22.517$, $p < 0.001$) the distribution of the occurrence of vague lexical items **is not** the same across all tested hosting locations. Hence a pairwise comparison between the hosting locations was performed in the next step.

(D) The pairwise comparison, summarized in Table 5.4, shows that there are statistical significant differences between Australia and 3 other countries, namely Canada, Germany, and the Unites States of America. The mean rank of the percentage occurrence of vague lexical items for Australian privacy policies lies 165.1 ranks behind the Canadian ones ($p = 0.005$). In addition, Australian privacy policies rank 118.1 ranks behind German ($p = 0.010$) and 101.8 ranks behind the U.S. American privacy policies ($p = 0.005$).

Table 5.4: Short comparison of hosting locations regarding the occurrence of vagueness (full version available in the Appendix as Table A.3).

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Canada-United Kingdom	-125.486	47.514	-2.641	.008	.174
Canada-Australia	161.910	44.211	3.662	.000	.005
Netherlands-Australia	145.452	57.950	2.510	.012	.254
Germany-United Kingdom	-81.711	37.927	-2.154	.031	.655
Germany-Australia	118.135	33.697	3.506	.000	.010
United States-Australia	101.849	28.438	3.581	.000	.007
Ireland-Australia	85.294	40.616	2.100	.036	.750

5.4 Differences between Long and Short Privacy Policies

Working with the dataset of this thesis, one got the impression that there are differences between very long and very short privacy policies. Typically, the very short privacy policies came from small development teams and often contained only a few sentences in which the developers stated that they would not collect any or only limited data from their users. The very long privacy policies, on the other hand, often came from large companies that seemed to want to protect themselves against a lot of legal concerns.

To verify whether the assumption is tenable, it was decided to separately investigate both groups of long and short privacy policies. As a separator a length of 30 sentences was chosen, which is also the text length that the SMOG formula was normed on. This means in the following sections short privacy policies are those that contain less than 30 sentences.

A Mann-Whitney U test was calculated to determine if there were differences regarding the Mean Readability Grade (MRG) of long privacy policies and privacy policies shorter than 30 sentences. The distributions differed between both groups, according to a Shapiro-Wilk test ($W = 0.894\,;\ p < 0.001$). There was a statistically significant difference in the MRG between long ($\mathrm{M}_{\mathrm{Rank}} = 336.79$) and short privacy policies ($\mathrm{M}_{\mathrm{Rank}} = 419.13$), U $= 57,948.50$, Z $= 4.594$, $p < .001$.

At the same time, another Mann-Whitney U test was calculated to determine if there were differences regarding the percentage occurrence of vague lexical items in long privacy policies and privacy policies shorter than 30 sentences. The data was approximately normally distributed, according to a Shapiro-Wilk test ($W = 0.996\,;\ p = 0.110$). There was a statistically significant difference in occurrence of vagueness in long ($\mathrm{M}_{\mathrm{Rank}} = 379.02$) and short privacy policies ($\mathrm{M}_{\mathrm{Rank}} = 289.31$), U $= 35,229.00$, Z $= -5.005$, $p < .001$.

In summary this means, that the longer privacy policies of the underlying dataset of this thesis were significantly easier to read than the short privacy policies. But in the same time the longer privacy policies contained significantly more vague language than the short ones.

5.5 Miscellaneous Results

When looking at other groups besides hosting locations, such as comparing non-native English-speaking countries with native English-speaking countries, no significant differences were found. Neither for readability, average reading time, nor for the occurrence of vague lexical items.

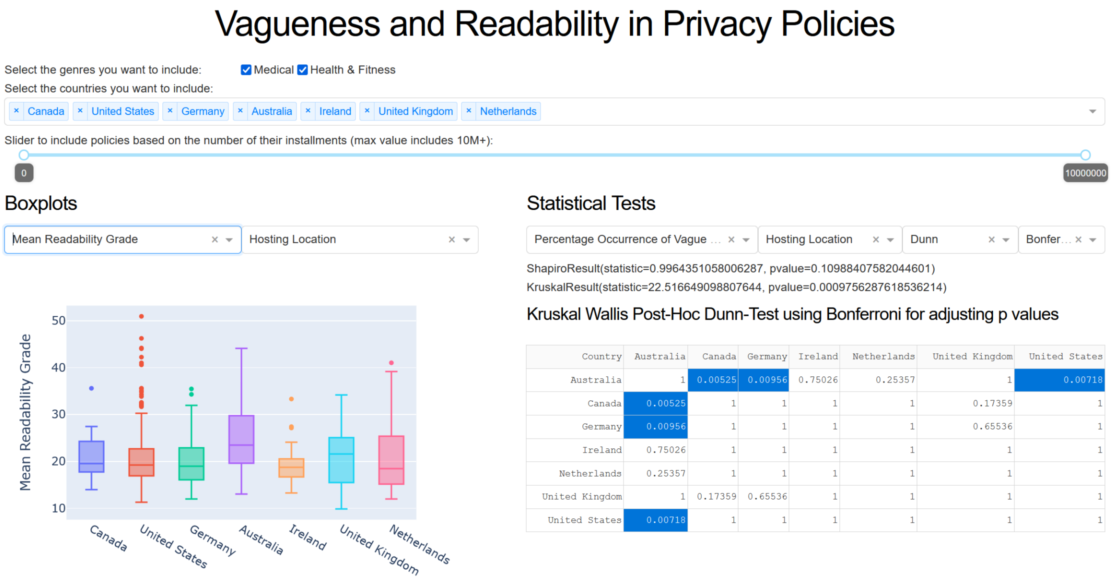
The same was the case for a subdivision of the privacy policies according to the download count of their associated apps. The following subdivision was made for this purpose:

- Very Low (less than 50,000 downloads)
- Low (between 50,000 and 250,000 downloads)
- Medium (between 250,000 and 1,000,000 downloads)
- High (between 1,000,000 and 5,000,000 downloads)
- Very High (more than 5,000,000 downloads)

No Significant differences in readability, reading time, or occurrence of vague lexical items were found between the 5 different download groups in these studies.

5.6 Dashboard Implementation Results

As mentioned in Section 4.3, a dashboard was implemented to visualize the results of this thesis. The resulting dashboard can be seen in Figure 5.3. In the final version it gives the user the ability to in- and exclude privacy policies from the dataset based upon their applications genre, their hosting location, as well as their number of downloads. This manipulation of the included privacy policies is done by a slider to vary the download range, a checklist to select the application genres to include and a drop-down list to select the hosting locations. Furthermore, the users can decide which of the measured values they want a boxplot to be drawn for and also for which values they want to perform the statistical significance tests presented in Section 2.5. The dashboard is responsive and automatically redraws its figures whenever the user interacts with the input fields.



Vagueness and Readability in Privacy Policies

Select the genres you want to include: ☑ Medical ☑ Health & Fitness

Select the countries you want to include:

× Canada × United States × Germany × Australia × Ireland × United Kingdom × Netherlands

Slider to include policies based on the number of their installments (max value includes 10M+):

0 10000000

Boxplots

Mean Readability Grade × | Hosting Location ×

Statistical Tests

Percentage Occurrence of Vague ... × | Hosting Location × | Dunn × | Bonfer... ×

ShapiroResult(statistic=0.9964351058006287, pvalue=0.10988407582044601)

KruskalResult(statistic=22.516649098807644, pvalue=0.0009756287618536214)

Kruskal Wallis Post-Hoc Dunn-Test using Bonferroni for adjusting p values

Country	Australia	Canada	Germany	Ireland	Netherlands	United Kingdom	United States
Australia	1	0.00525	0.00956	0.75026	0.25357	1	0.00718
Canada	0.00525	1	1	1	1	0.17359	1
Germany	0.00956	1	1	1	1	0.65536	1
Ireland	0.75026	1	1	1	1	1	1
Netherlands	0.25357	1	1	1	1	1	1
United Kingdom	1	0.17359	0.65536	1	1	1	1
United States	0.00718	1	1	1	1	1	1

Figure 5.3: Final implementation of the dashboard

The codebase of the dashboard can be found at GitHub[4] and a deployed version of the dashboard is available at Heroku[5].

The linked deployment uses the free plan from Heroku[5]. It **takes a few seconds to reload** after being in idle mode and it will remain online and publicly available as long as there is no unplanned high traffic on the site. Should Heroku[5] change its subscription model and the deployment is **unavailable for a longer time**, please follow the read.me instructions at the linked GitHub[4] repository to run the dashboard on your machine.

In the early stages of this thesis, when it was unclear how many privacy policies and which countries would be included in the final dataset, there was the idea of creating a world map. This world map was supposed to contain data about the situation of readability and vagueness in as much countries as possible. As the data selection process and data processing progressed, it was clear that the final dataset would not include

[4] https://github.com/pcschwarz/analyzing-privacy-policies

[5] https://analyzing-privacy-policies.herokuapp.com/ please note that this website uses JavaScript and takes a few seconds to build up when first accessed, as it switches to idle after 30 minutes of no traffic to minimize resource consumption.

sufficient data to create an approximately complete world map. The reason why less data remained than expected was that through the data selection process, shown in Figure 4.1, duplicate or non-English privacy policies, for example, had to be removed from the dataset. From the starting 3,992 policies only 713 were kept for the final version of the dashboard. Therefore, the idea of adding a world map to the dashboard was discarded. Another issue was that some websites included different texts like their terms of use, their imprint, or general terms and conditions on the same page as the text of the their privacy policy. It is difficult to determine which scraped privacy policy texts are affected, but these mislabeled texts might skew the results. However, this bias presumably affects all countries in the same way. Therefore, comparisons between hosting locations in terms of linguistic test results should not favor one country or another because of these misscraped texts. In addition, much of the skewed data was corrected during a manual review of the outliers. For this reason, privacy policies were manually reviewed that were outliers in terms of very high word and sentence counts, or that simply contained very unusual numbers for the amount of vague language or readability scores calculated.

6 Discussion

6.1 Interpretation of the Results

Looking at the overall results for the Mean Readability Grade, it is quite frightening that the results indicate that the median privacy policy, of a healthcare related Android app, requires its reader to be nothing less than a university graduate. This result is equal for the SMOG Index, Gunning Fog, Flesch-Kincaid, and Automated Readability Index. Only the Coleman-Liau Index results were more optimistic, indicating a required education around the level of college graduates. However, apps are used throughout all education levels. Naturally, a privacy policy text cannot be written to cover all of them, but privacy policy authors should strive to write their texts in a way that a vast majority of society is able to comprehend them. Many other fields already adopted common practices to ensure this. For example, in several U.S. states it's a common requirement for legal documents such as insurance policies to be written at no higher education level than ninth grade, even since the 1980s [36].

Regarding the required reading times of the different hosting locations privacy policies, there were no statistical differences between any pair of countries. In other words, the length of the privacy policies were about the same throughout all investigated countries.

Similar observations were made when grouping privacy policies by the number of downloads their associated app has. The popularity of an app had no influence on its privacy policies readability, required reading time, or occurrence of vague language. Likewise, these variables were not affected by whether the privacy policy originated from a country where English is the native language or not.

However, the results from Section 5.3 showed in the pairwise comparisons that, for example, Australian privacy policies are significantly more difficult to read than Irish, German and U.S. American privacy policies. Likewise, the percentage occurrence of vague lexical items in Australian privacy policies was significantly higher than in Canadian, German and U.S. American privacy policies. Unfortunately, the data collected in this thesis does not provide any scientifically supported conclusions about the reasons for the different performance of individual hosting locations. This thesis only answers the question if there are differences between the hosting locations and only for a small amount of WEIRD countries. At the same time, this international comparison shows that the readability and comprehensibility of privacy policies is poor, even in the more advanced countries. If the situation in WEIRD countries is already bad, this is no reason for optimism that it will be any better in the rest of the world.

At last, it is noteworthy that especially in the Coleman-Liau Index Irish and British privacy policies performed significantly better than Australian, Canadian and U.S. American privacy policies.

6.1.1 Unexpected Results

The results of Section 5.4, which compared longer and shorter privacy policies, were a bit unexpected. When dividing between long and short privacy policies, one might expect short privacy policies to be simpler written and therefore easier to read than long privacy policies due to their conciseness. In contrast to this, the short privacy policies analyzed during this thesis were significantly harder to read than the privacy policies containing 30 or more sentences. At the same time the short privacy policies contained significantly less vague language.

At first these results looked irritating, but one possible explanation for this might be, that shorter policies are usually written by developers who are specialized on coding instead of writing. One example for this would be the app 'Heart Rate Monitor' from the Google Play Store genre 'Health & Fitness'. The app was installed over five million times and the following text represents their full privacy policy, copied as is.[1]

Accurate Heart Rate Monitor - Privacy Policy
May 02, 2017

- *We do NOT collect any personal information. 'Personal Information' is information that identifies you or another person, which may be collected when you use this application.*

[1] Taken from the developers blog https://repsiventure.blogspot.com linked to the Play Store Page https://play.google.com/store/apps/details?id=com.repsi.heartrate by the 29th of March 2022

- *All images received by the phone camera is discarded except your estimated heart rate which is stored locally on your phone and can be deleted anytime. None of this information sends to us or a third party.*
- *We use 'google analytics' to gather anonymously general usage of the application, which includes demographic information such as country and gender. This is in line with the google analytics privacy policy.*

This provided example is one of the shortest scraped privacy policies in this thesis. Its sentences are kept rather simple, but it is quite likely not written by a professional English writer. These very short privacy policies often contain a lot of bullet point lists or they combine headlines into their texts in a way that sometimes confuses common implementations of readability formulas like *NLTK*'s. On the other hand, longer privacy policies seem to be related to bigger companies or organizations usually. These bigger players are more likely to have professional writers check their legal documents before they get released. Such quality checks could be the reason why long privacy policies performed better across the used readability formulas in this thesis. Especially short policies which consist of a few rather long sentences, like the next example sentence[2] are also negatively hitting on the readability scores of short privacy policies.

> '[...] *If you have concern about your personal identifiable information being misused, or if you want further information about our privacy policy and what it means, please feel free to email us at collageteam.feedback@gmail.com, we will endeavor to provide clear answers to your questions in a timely manner.*'

However, looking at the differences in the occurrence of vague language between shorter and longer privacy policies, shorter privacy policies contained significantly less vague lexical items percentwise. This could be an indicator that smaller companies and developer teams tend to sugar-coat their actual data practices less than bigger companies. The larger a company gets the more it might be concerned about negative publicity of a privacy lawsuit. To avoid this, some organizations could write their privacy policy vaguer to make it harder to be legally nailed down to it. Please be aware that this attempted explanation is still pure speculation. The available data only show a difference between shorter and longer privacy policies, but how that difference actually is caused is a combination of a number of factors that cannot be taken from the texts of privacy policies

[2] Taken from `https://inshotapp.com/website/collage/policy.html` linked to the App 'Body Editor' `https://play.google.com/store/apps/details?id=breastenlarger.bodyeditor.photoeditor` with 30+ Million downloads by the 29th of March 2022

alone. However, noting this difference leads to an attempt to include the exact size of the companies in a future comparison of privacy policies.

6.2 The Results in Comparison to Previous Studies

There have been several studies testing the readability of privacy policies in general without distinguishing between hosting locations or app genres. Their results usually ranged from lower college level up to post graduate levels. One quite similar study like the work of Ermakova et al. [13] compared readability of privacy policies of 1,166 e-commerce websites vs. 5,431 generic healthcare websites in 2015. Their results often indicated 3-4 years of education less necessary than the result of this thesis, which is a lot. This might be caused by the fact that this thesis kept a lot of outliers with very bad readability results since they were manually verifiable. Even when comparing the medians of the results for the readability in healthcare related websites with the results of this thesis there were big differences. Ermakova et al. measured a median Gunning Fog Index of 16.08 whereas this thesis measured a median Gunning Fog Index of 18.67. Other indices differed even more like the SMOG Index for which Ermakova et al. measured a median of 13.86 and this thesis measured a median of 18.7.

This was rather irritating at first, since Ermakova et al. were also focusing healthcare related privacy policies. But looking further in the comparison of the of the two studies other differences showed up. For example, the median healthcare privacy policy of Ermakova et al. contained only 762 words while the median privacy policy coming from the genres Health & Fitness or Health from the Google Play Store contained 2,409 words. Assuming the scraping process of this thesis was flawed, all potential outliers, resembling 187 of the 713 privacy policies, were manually checked and re-scraped again. This re-scraping process did not lower the word count nor the results of the readability formulas, which indicates that healthcare related privacy policies related to Apps from the Google Play Store are more complex than general healthcare related privacy policies. That theory assumes that the scraping process works similarly like, for example both handled collapsible content in privacy policies the same. This thesis tried to scrape all accordion elements and collapsible text throughout the data collection process. Another influencing factor could be that Ermakova et al. collected their data before 2015. The GDPR however released in April of 2016, which could also be a factor for many websites to update their privacy policies and maybe increase their word count by this.

Looking at the occurrence of vagueness in privacy policies the results of this thesis are a little higher than the results of Cadogan [4] as well as Krumay and Klar [28]. Krumay and Klar measured a percentage occurrence of vague lexical items between 0.13% and 1.65% across the 15 privacy policies they analyzed. On average their analyzed privacy policies contained 0.73% vague language. The 713 privacy policies of this thesis contained on average 1.19% vague language. The discrepancy in the results might be caused by sample size and, as the tool Krumay and Klar used for their measurements is no longer available, one cannot tell how their method to identify vague language differs from the method used in this thesis.

7 Conclusion

7.1 Summary of the Thesis

During this master thesis, privacy policies of apps from 10 WEIRD countries in the categories 'Medical' and 'Health & Fitness' from the Google Play Store were scraped. Starting with around 3,992 policies, after data selection and data processing, 713 English-language privacy policies remained for which the readability grade level and the percentage of vague language were calculated. Significant regional differences in readability and the frequency of vagueness were found. It also turned out that the short privacy policies were less readable, but at the same time less vague than the longer privacy policies examined in this thesis. To visualize the results, and to provide a way to work interactively with the results, a dashboard was implemented and publicly deployed.

7.2 Limitations

The insights gained from the processed privacy policy dataset of this thesis are limited. First, this thesis just did include privacy policies originating from apps of the Google Play Store. Furthermore, the data collection process was limited on the English version of Google Play Store of only 10 WEIRD countries. Privacy policies of apps from the Apple App Store or from generic websites or web services across the internet were not included.

The hosting location check performed in this thesis by looking up IP addresses via ipwhois.io is not perfectly accurate but should be one of the better indicators, as most scraped privacy policies did omit this information and the Google Play Store also does not provide this metadata.

Regarding the calculated amount of vague language contained in the privacy policies in this thesis, one needs to note that the vague lexical items are automatically counted.

This means there will be some sentences detected as vague by the python script used in this thesis which would turn out to be not vague when read by a human. As mentioned in Section 3.5, one should expect around 15.5% false positives. This was the amount of vague lexical items which turned out to be precise after human verification in the work of Lebanoff and Liu [29]. In their work Lebanoff and Liu used a largely consistent dataset of vague lexical items as the one used in this thesis (see Table 3.1). It is assumed that this effect hits every hosting location and genre in the same way, but one should be aware that the actual amount of vague language is probably less than this thesis measured.

Lastly one of the major limitations of this work is that it is only looking at privacy policies. Privacy policies are just documents in the end. Even if the text of a privacy policy is written perfectly precise, only the developers of a web service or app can certainly tell if the written text agrees with the actual internal data practices of an organization. It's a matter of trust, users of apps need to trust in the correctness of privacy policies, which in a certain they can because companies want to keep their users trust. It just might be, that in some cases the users trust is not the number one priority of an organization. The findings of this thesis however are based upon the assumption that privacy policies can be trusted.

7.3 Significance of this Study

All in all, this thesis just covers a tiny portion of privacy policies from the Internet. Apart from that, related studies all used different methods to identify vague language and sometimes also different formulas, as mentioned in Section 2.4, and therefore occasionally came to different absolute values for vagueness and readability indices. However, this work showed that there are significant differences between some hosting locations, even among WEIRD countries between which one would not necessarily expect such differences. It further gives a hint how much privacy is valued in different countries. On the other hand this thesis results could be misused by people looking for a hosting location in an area where privacy concerns have lesser priority. To be fair, for this unwanted use-case an extended covering more than just 10 countries would be necessary. Lastly it must be said, this thesis provides only indications and no proven explanations for the observations. To scientifically explain the poor performance of, for example, Australian privacy policies from the dataset of this thesis, a more in-depth linguistic analysis is needed.

7.4 Future Directions

The easiest next step to extend the research of this thesis is to include more genres and countries from the Google Play Store. This would increase the pool of countries that remain in the dataset after processing the privacy policies and one could, for example, draw world maps showing the differences in readability and vagueness in privacy policies. Including Apples App Store or privacy policies from all kinds of websites would even increase the informative value of the underlying dataset. As this thesis mainly covers the situation in WEIRD countries, it would be exciting to investigate how good or bad the situation is in other technically less developed countries.

It would also be interesting to split each privacy policy into the segments defined by Wilson et al. [44] shown in Figure 3.1 and afterwards test the readability and occurrence of vagueness on each of the segments of each privacy policy. This way one could investigate for example if the segment *3rd Party Collection*, a segment with most likely unpleasant content for the user, is more precise or less precise than the segment *Data Security*, which is a segment where organizations may shine by emphasizing their efforts to protect one's personal information. While comparing the different segments of privacy policies one could also look at the occurrence of the rhetorical patterns mentioned in Section 3.1. It would be interesting to see if some rhetorical patterns like switching to passive sentences or omitting self-references occur in different frequencies across the segments of each privacy policy. Especially segments like the *3rd Party Collection* are candidates to look for a deviating amount of sugar-coated language.

Another interesting approach would be not only to track the hosting location of application and its privacy policy but also to track which industry branch it belongs to. This way one could gain some insights on how different industries' privacy policies perform in readability and vagueness. Maybe there are differences between certain industries too.

Bibliography

[1] Bhatia, Jaspreet ; Breaux, Travis D. ; Reidenberg, Joel R. ; Norton, Thomas B.: A Theory of Vagueness and Privacy Risk Perception. In: *2016 IEEE 24th International Requirements Engineering Conference (RE)* (2016), S. 26–35

[2] Brumen, Bostjan ; Heričko, Marjan ; Sevčnikar, Andrej ; Završnik, Jernej ; Hölbl, Marko: Outsourcing Medical Data Analyses: Can Technology Overcome Legal, Privacy, and Confidentiality Issues? In: *Journal of Medical Internet Research* 15 (2013), Dezember, Nr. 12, S. e283. – URL https://doi.org/10.2196/jmir.2471

[3] Brysbaert, Marc: How many words do we read per minute? A review and meta-analysis of reading rate. In: *Journal of Memory and Language* 109 (2019), Dezember, S. 104047. – URL https://doi.org/10.1016/j.jml.2019.104047

[4] Cadogan, Rochelle A.: An Imbalance Of Power: The Readability Of Internet Privacy Policies. In: *Journal of Business & Economics Research (JBER)* 2 (2004), Februar, Nr. 3. – URL https://doi.org/10.19030/jber.v2i3.2864

[5] Chall, Jeanne S. ; Dale, Edgar: *Readability Revisited - The New Dale-Chall Readability Formula*. Brookline Books, 1995. – ISBN 978-1-571-29008-3

[6] Coleman, Meri ; Liau, T. L.: A computer readability formula designed for machine scoring. In: *Journal of Applied Psychology* 60 (1975), Nr. 2, S. 283–284. – URL https://doi.org/10.1037/h0076540

[7] Costante, Elisa ; Sun, Yuanhao ; Petković, Milan ; Hartog, Jerry den: A machine learning solution to assess privacy policy completeness. In: *Proceedings of the 2012 ACM workshop on Privacy in the electronic society - WPES '12*, ACM Press, 2012, S. 91—-96. – URL https://doi.org/10.1145/2381966.2381979

[8] Dale, Edgar ; Chall, Jeanne S.: A formula for predicting readability. In: *Educational Research Bulletin* 27 (1948), S. 11–20

[9] Dubay, William: The Principles of Readability. In: *CA* 92627949 (2004), 01, S. 631–3309

[10] Earp, J.B. ; Anton, A.I. ; Aiman-Smith, L. ; Stufflebeam, W.H.: Examining Internet Privacy Policies Within the Context of User Privacy Values. In: *IEEE Transactions on Engineering Management* 52 (2005), Mai, Nr. 2, S. 227–237. – URL https://doi.org/10.1109/tem.2005.844927

[11] Eckstein, Peter P.: *Angewandte Statistik mit SPSS*. Springer Fachmedien Wiesbaden, 2016. – URL https://doi.org/10.1007/978-3-658-10918-9

[12] Education, National Center for Education S. U.S. Department of: *Digest of Education Statistics*. https://nces.ed.gov/programs/digest/d12/figures/fig_01.asp. 2012. – [Online; accessed 18-February-2022]

[13] Ermakova, Tatiana ; Fabian, Benjamin ; Babina, Eleonora: Readability of Privacy Policies of Healthcare Websites. In: *Wirtschaftsinformatik*, 03 2015, S. 1085–1099

[14] European Commission: *Regulation (EU) 2016/679 of the European Parliament and of the Council 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) (Text with EEA relevance)*. 2016. – URL https://eur-lex.europa.eu/eli/reg/2016/679/oj

[15] Fabian, Benjamin ; Ermakova, Tatiana ; Lentz, Tino: Large-scale readability analysis of privacy policies. In: *Proceedings of the International Conference on Web Intelligence*, ACM, August 2017, S. 18–25. – URL https://doi.org/10.1145/3106426.3106427

[16] Felt, Adrienne P. ; Barnes, Richard ; King, April ; Palmer, Chris ; Bentzel, Chris ; Tabriz, Parisa: Measuring HTTPS Adoption on the Web. In: *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC : USENIX Association, August 2017, S. 1323–1338. – URL https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt. – ISBN 978-1-931971-40-9

[17] Flesch, Rudolph: A new readability yardstick. In: *Journal of Applied Psychology* 32 (1948), Nr. 3, S. 221–233. – URL https://doi.org/10.1037/h0057532

[18] Friedman, Batya ; Hurley, David ; Howe, David C. ; Felten, Edward ; Nissenbaum, Helen: Users' conceptions of web security. In: *CHI '02 extended abstracts on Human factors in computing systems - CHI '02*, ACM Press, 2002, S. 746–747. – URL https://doi.org/10.1145/506443.506577

[19] GDPR: *Regulation (EU) 2016/679 of the European Parliament and of the Council 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) (Text with EEA relevance)*. May 2016. – URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[20] Gunning, Robert: *The Technique of Clear Writing*. New York : McGraw-Hill, 1952. – ISBN 978-7-000-01419-0

[21] Harkous, Hamza ; Fawaz, Kassem ; Lebret, Rémi ; Schaub, Florian ; Shin, Kang G. ; Aberer, Karl: *Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning*. 2018

[22] Henrich, Joseph ; Heine, Steven J. ; Norenzayan, Ara: Beyond WEIRD: Towards a broad-based behavioral science. In: *Behavioral and Brain Sciences* 33 (2010), Juni, Nr. 2-3, S. 111–135. – URL https://doi.org/10.1017/s0140525x10000725

[23] Holtz, Leif-Erik ; Nocun, Katharina ; Hansen, Marit: Towards Displaying Privacy Information with Icons. In: *IFIP Advances in Information and Communication Technology*. Springer Berlin Heidelberg, 2011, S. 338–348. – URL https://doi.org/10.1007/978-3-642-20769-3_27

[24] Hübner, U. ; Sax, U. ; Prokosch, H.-U.: *German Medical Data Sciences: A Learning Healthcare System - Proceedings of the 63rd Annual Meeting of the German Association of Medical Informatics, Biometry and Epidemiology (gmds e.V.) 2018 in Osnabrück, Germany – GMDS 2018*. München : IOS Press, 2018. – ISBN 978-1-614-99896-9

[25] Iachello, Giovanni ; Hong, Jason: End-User Privacy in Human-Computer Interaction. In: *Foundations and Trends® in Human-Computer Interaction* 1 (2007), Nr. 1, S. 1–137. – URL https://doi.org/10.1561/1100000004

[26] Kincaid, J. P. ; Fishburne, Robert P. ; Rogers, R L. ; Chissom, Brad S.: *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.* 1975

[27] Klare, George R.: *The Measurement of Readability.* Iowa State University Press, 1963. – ISBN 978-0-608-12978-5

[28] Krumay, Barbara ; Klar, Jennifer: Readability of Privacy Policies. In: *Data and Applications Security and Privacy XXXIV.* Springer International Publishing, 2020, S. 388–399. – URL https://doi.org/10.1007/978-3-030-49669-2_22

[29] Lebanoff, Logan ; Liu, Fei: *Automatic Detection of Vague Words and Sentences in Privacy Policies.* 2018

[30] Li, Shuangling: A corpus-based study of vague language in legislative texts: Strategic use of vague terms. In: *English for Specific Purposes* 45 (2017), Januar, S. 98–109. – URL https://doi.org/10.1016/j.esp.2016.10.001

[31] Li, Shuangling: Communicative significance of vague language: A diachronic corpus-based study of legislative texts. In: *English for Specific Purposes* 53 (2019), Januar, S. 104–117. – URL https://doi.org/10.1016/j.esp.2018.11.001

[32] Lindskog, Helena ; Trotzenfeldt, Helena ; Lindskog, Stefan: *Web Site Privacy with P3P.* New York : Wiley, 2003. – ISBN 978-0-471-21677-3

[33] Liu, Fei ; Fella, Nicole L. ; Liao, Kexin: *Modeling Language Vagueness in Privacy Policies using Deep Neural Networks.* 2018

[34] Massey, Aaron K. ; Rutledge, Richard L. ; Antón, Annie I. ; Swire, Peter P.: Identifying and classifying ambiguity for regulatory requirements. In: *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 2014, S. 83–92

[35] McCall, William A. ; Schroeder, Lelah C.: *McCall-Crabbs Standard Test Lessons in Reading, Book A -.* New York : Teachers College Press, 1979. – ISBN 978-0-807-75540-2

[36] McClure, Glenda M.: Readability formulas: Useful or useless? In: *IEEE Transactions on Professional Communication* PC-30 (1987), Nr. 1, S. 12–15. – URL https://doi.org/10.1109/tpc.1987.6449109

[37] McDonald, Aleecia M. ; Cranor, Lorrie F.: The Cost of Reading Privacy Policies. In: *Journal of Law and Policy for the Information Society* (2009)

[38] MCLAUGHLIN, G. H.: SMOG Grading - A New Readability Formula. In: *The Journal of Reading* (1969)

[39] POLLACH, Irene: What's wrong with online privacy policies? In: *Communications of the ACM* 50 (2007), September, Nr. 9, S. 103–108. – URL https://doi.org/10.1145/1284621.1284627

[40] RESCORLA, Eric: *HTTP Over TLS.* RFC 2818. Mai 2000. – URL https://www.rfc-editor.org/info/rfc2818

[41] SMITH, E A. ; SENTER, R.: Automated readability index. In: *AMRL-TR. Aerospace Medical Research Laboratories* (1967), S. 1–14

[42] STEINBERGER, Ralf ; POULIQUEN, Bruno ; WIDIGER, Anna ; IGNAT, Camelia ; ERJAVEC, Tomaz ; TUFIS, Dan ; VARGA, Daniel: *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.* 2006

[43] WHALEN, Tara ; INKPEN, Kori: Gathering evidence: use of visual security cues in web browsers. In: *Proceedings of Graphics Interface 2005.* School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada : Canadian Human-Computer Communications Society, 2005 (GI 2005), S. 137–144. – ISBN 1-56881-265-5

[44] WILSON, Shomir ; SCHAUB, Florian ; DARA, Aswarth A. ; LIU, Frederick ; CHERIVIRALA, Sushain ; GIOVANNI LEON, Pedro ; SCHAARUP ANDERSEN, Mads ; ZIMMECK, Sebastian ; SATHYENDRA, Kanthashree M. ; RUSSELL, N. C. ; NORTON, Thomas B. ; HOVY, Eduard ; REIDENBERG, Joel ; SADEH, Norman: The Creation and Analysis of a Website Privacy Policy Corpus. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Berlin, Germany : Association for Computational Linguistics, August 2016, S. 1330–1340. – URL https://aclanthology.org/P16-1126

[45] YU, Le ; ZHANG, Tao ; LUO, Xiapu ; XUE, Lei: AutoPPG. In: *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, ACM, Oktober 2015, S. 39–50. – URL https://doi.org/10.1145/2808117.2808125

A Appendix

Table A.1: Full comparison of hosting locations regarding Mean Readability Grades

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Ireland-Germany	18.925	37.031	.511	.609	1.000
Ireland-Netherlands	-18.976	59.952	-.317	.752	1.000
Ireland-United States	-33.637	32.318	-1.041	.298	1.000
Ireland-United Kingdom	-54.527	44.191	-1.234	.217	1.000
Ireland-Canada	71.377	46.803	1.525	.127	1.000
Ireland-Australia	165.739	40.618	4.080	.000	.001
Germany-Netherlands	-.051	55.498	-.001	.999	1.000
Germany-United States	-14.711	23.027	-.639	.523	1.000
Germany-United Kingdom	-35.602	37.929	-.939	.348	1.000
Germany-Canada	52.452	40.942	1.281	.200	1.000
Germany-Australia	146.814	33.698	4.357	.000	.000
Netherlands-United States	-14.660	52.471	-.279	.780	1.000
Netherlands-United Kingdom	-35.551	60.511	-.588	.557	1.000
Netherlands-Canada	52.401	62.444	.839	.401	1.000
Netherlands-Australia	146.763	57.953	2.532	.011	.238
United States-United Kingdom	20.890	33.344	.627	.531	1.000
United States-Canada	37.741	36.735	1.027	.304	1.000
United States-Australia	132.102	28.439	4.645	.000	.000
United Kingdom-Canada	16.850	47.517	.355	.723	1.000
United Kingdom-Australia	111.212	41.438	2.684	.007	.153
Canada-Australia	94.362	44.213	2.134	.033	.689

Table A.2: Full comparison of hosting locations regarding the Coleman-Liau index

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Ireland-United Kingdom	-3.599	44.190	-.081	.935	1.000
Ireland-Netherlands	-33.788	59.950	-.564	.573	1.000
Ireland-Germany	47.813	37.030	1.291	.197	1.000
Ireland-United States	-110.560	32.318	-3.421	.001	.013
Ireland-Canada	183.033	46.802	3.911	.000	.002
Ireland-Australia	189.494	40.617	4.665	.000	.000
United Kingdom-Netherlands	30.189	60.509	.499	.618	1.000
United Kingdom-Germany	44.214	37.928	1.166	.244	1.000
United Kingdom-United States	-106.960	33.343	-3.208	.001	.028
United Kingdom-Canada	179.433	47.516	3.776	.000	.003
United Kingdom-Australia	185.895	41.437	4.486	.000	.000
Netherlands-Germany	14.025	55.496	.253	.800	1.000
Netherlands-United States	-76.771	52.470	-1.463	.143	1.000
Netherlands-Canada	149.244	62.442	2.390	.017	.354
Netherlands-Australia	155.706	57.951	2.687	.007	.151
Germany-United States	-62.746	23.027	-2.725	.006	.135
Germany-Canada	135.220	40.941	3.303	.001	.020
Germany-Australia	141.681	33.698	4.205	.000	.001
United States-Canada	72.473	36.734	1.973	.049	1.000
United States-Australia	78.935	28.439	2.776	.006	.116
Canada-Australia	6.462	44.212	.146	.884	1.000

Table A.3: Full comparison of hosting locations regarding the occurrence of vagueness

Location Pair	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Bonferroni Adj. Sig.
Canada-Netherlands	-16.458	62.441	-.264	.792	1.000
Canada-Germany	-43.775	40.940	-1.069	.285	1.000
Canada-United States	-60.061	36.734	-1.635	.102	1.000
Canada-Ireland	-76.615	46.801	-1.637	.102	1.000
Canada-United Kingdom	-125.486	47.514	-2.641	.008	.174
Canada-Australia	161.910	44.211	3.662	.000	.005
Netherlands-Germany	27.317	55.495	.492	.623	1.000
Netherlands-United States	-43.603	52.469	-.831	.406	1.000
Netherlands-Ireland	60.158	59.949	1.003	.316	1.000
Netherlands-United Kingdom	-109.028	60.508	-1.802	.072	1.000
Netherlands-Australia	145.452	57.950	2.510	.012	.254
Germany-United States	-16.286	23.026	-.707	.479	1.000
Germany-Ireland	-32.840	37.029	-.887	.375	1.000
Germany-United Kingdom	-81.711	37.927	-2.154	.031	.655
Germany-Australia	118.135	33.697	3.506	.000	.010
United States-Ireland	16.555	32.317	.512	.608	1.000
United States-United Kingdom	65.425	33.342	1.962	.050	1.000
United States-Australia	101.849	28.438	3.581	.000	.007
Ireland-United Kingdom	-48.871	44.189	-1.106	.269	1.000
Ireland-Australia	85.294	40.616	2.100	.036	.750
United Kingdom-Australia	36.424	41.437	.879	.379	1.000

Glossary

Docker Docker is a software that enables container virtualization of applications. Applications can be packed into a Docker image. A Docker container is a running instance of a Docker image, effectively a running application that has all dependencies with it and can be executed by any Docker host, as long as the hardware requirements are fulfilled.

Jupyter Notebook Jupyter Notebook and JupyterLab provide nowadays an easy way to exchange scientific work for example in the field of designing algorithms. The so called Notebooks consist of an alternation of code cells and text cells which usually are used to provide background information to understand the code cells.

NLTK The Natural Language Toolkit is a Python library for processing human language data.

GDPR The General Data Protection Regulation (GDPR) is the European Union's current legal framework that defines how personal data, belonging to its citizens, may be collected and processed.

Great Firewall The Golden Shield Project, which is often just called the Great Firewall, is a project managed by the Chinese Ministry of Public Security to monitor and limit the internet access from inside China.

HTTPS The Hypertext Transfer Protocol Secure (HTTPS) is an encrypted version of the HTTP protocol which secures communication on the Internet. This protected connection allows clients to securely exchange sensitive data with a server, for example, banking activities. For more detailed information have a look at the RFC-Standard 2818 [40].

JRC-Acquis A corpus consisting of legal texts from all member states of the European Union. See Section 3.4 for more detailed information.

OPP-115 A detailed dataset of 115 privacy policies, which phrases got manually annotated by lawyers and segmented by topic. See Section 3.2 for more detailed information..

Polisis A CNN trained with the OPP-115 annotations for automatic content segmentation of random privacy policies.

Erklärung zur selbstständigen Bearbeitung einer Abschlussarbeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe. Wörtlich oder dem Sinn nach aus anderen Werken entnommene Stellen sind unter Angabe der Quellen kenntlich gemacht.

Ort Datum Unterschrift im Original