Readability Analysis of Privacy Policies

Marcus Voss and Eric Lee Weißbrich

Humboldt Universität zu Berlin,
Institute of Information Systems
{marcus.voss,eric.weissbrich}@student.hu-berlin.de

Abstract. Generally the internet user is not aware about how his personal data is collected, used and disclosed by internet services he uses, because he doesn't read their privacy policies. Assuming that the readability of a policy influences this behavior, we develop tool Privacy Policy Analyzer, which is used to analyze privacy policies based on different measures for text difficulty. We provide an overview over the common classic readability measures as well as state of the art approaches and work for visualization. We show that the readability level of the privacy policies from the most popular web sites is at college level or above. In coherence with the average age of the internet user and their education, one essential statement is, that in general privacy policies are to difficult, when compared to the reading comprehension level of the average internet user. An improvement of readability is necessary to cover most internet users and give them a chance of understanding privacy policies easily. Using Google as an example we show that an improvement of readability can be achieved.

Keywords: Readability, Privacy Policies, NLTK

1 Introduction

1.1 Motivation

The control over personal information in the internet becomes an increasing need. Users disclose more and more information while using internet services on a daily basis without knowing how information is collected, used and disclosed by the service provider. Privacy policies do explain how the data provided by the user is processed.

Although most people do care about privacy and how their personal information is used on the internet, they do generally not read privacy policies - not even of the websites they are using most frequently [28]. By doing so they could improve awareness regarding their privacy in the internet.

One assumption could be that the fact that few people are actually reading privacy policies depends on their *readability*. In this paper we will therefore implement a tool to analyze and compare the readability levels required to comfortably read privacy policies of popular web sites. The precise research question will follow in section 1.2.

To cover the aspects of the research topic regarding privacy policies the paper is organized as follows: The theoretical part 2 first covers related work in this field. We will then give an overview over various established measures in section 2.2 and continue with recent work in the field of readability and the visualization of text difficulty. After these theoretical foundations our proposed practical approach is described in section 3 by starting with python and the frameworks that are used. We continue to describe how we transferred the theoretical foundations into a working program that delivers specific results with strong focus on the visualization of readability in section 3.3. As a result the detailed analysis of exemplary sites follows directly in section 4. Finally we deliver a conclusion, including topics fore future work and existing research limitations to our approach 5.

1.2 Research question

The goal of this paper is to analyze popular sites privacy policies by developing a tool based on different measures for text difficulty. The tool should facilitate the readability analysis a policy and provide for comparison between different company's privacy policies. It will further enable the evaluation of the discrepancy between the difficulty to read a policy and the reading comprehension level of the average internet user.

2 Theory

2.1 Related Work

In their paper "Are Online Privacy Policies Readable?" from 2010, Summeeth et al.[29] elaborate a similar research question. After explaining the requirements and restrictions of privacy policies in general they define specific requirements (e.g. structure, content, accessibility) that make these policies readable. They collect a selection of privacy policies from popular internet sites and analyze their readability with the aid of readability formulas. For the analysis Sumeeth et al. use various classic measures such as the Flesh Readability Score, Dale-Chall Index, Fog Index, Fry Graph and few others. After the general analysis they compare the given results with those of Hochhauser [15], Anton et al. [5] and Jensen and Potts [18] and come to the conclusion that privacy policies are getting shorter over time and that on average a reading comprehension at college level (grade levels of 12-14) is required to understand the policies. This is lower than the average education level of the internet user and U.S. population today. The general statement of the paper is that policies getting more and more readable, but that a certain awareness of the providers needs to be maintained to further keep and improve the accessibility and readability of the policies.

The unpublished paper from Huning, "Readability of Privacy Policies of German Websites - Are users able to understand what they agree on?" [16] has the main focus on a differentiation between German and English policies. It analyses various German privacy policies with the German Flesch Readability Score, the Lix and the Wiener Sachtextformel. The paper provides a basic overview over the readability, but the results are limited by the fact that the measures applied are not necessarily comparable among different languages.

Another relating publication is "A Comparative Study of Online Privacy Policies and Formats" by McDonald et al. [23]. They compare three different approaches of privacy policy formats to improve their readability. The first approach is to derive a shortened and standardized form of arbitrary privacy policies. The second is a standardization of privacy policies in a bulleted format. The third and last evaluation is on the common non-standardized policy. With these three formats of notation they tested the readability with more than 700 participants and came to the result that all kinds of notations were disliked and were not easily readable for everyone.

An approach from a different point of view gives the article "How Users Read and Comprehend Privacy Policies" by Vu et al.[32]. Vu et al. analyze reading habits and understanding regarding privacy policies with Eye-gaze data. The participants in the survey have to answer various questions regarding policy contents. With the Eye-gaze data Vu et al. observe how the attendees search for information in the text. The result of the study is that most attendees find it very hard to extract the relevant information out of policies because of their bad structure and incomprehensible formulation.

Agrawal et al. do not focus solely on the readability, but rather focus on the comparison regarding the content of privacy policies in their paper "Ranking Privacy Policy" [4]. The goal is to rank different privacy policies against each other. They develop a new mathematical model to rank every policy based on content. The model analyzes the various components of the policies and provides an final score, on which one could decide how "good" or "bad" the given policy seems to be. The paper however only analyzes only a few exemplary sites.

2.2 Readability measures

The Literacy Dictionary (by Harris and Hodges 1995) defines *readability* as “the ease of comprehension because of style of writing” (as found in [13]). Dale and Chall give a more refined definition in 1949, by stating that readability is "the sum total (including interactions) of all those elements within a given piece of printed material that effect the success of a group of readers have with it. The success is the extend to which they understand it, read it at an optimal speed, and find it interesting" as found in [6]).

Readability formulas are an attempt to make assessing readability of texts objective - most are even so objective that they can be automated. A readability formula is an equation which combines different text features that best predict text difficulty [6].

The development of such an equation is done by studying the relationship between different text features and readability. The proof of success of a readability formulas is a correlation with the score obtained by a reading comprehension test applied to texts of different difficulties (so called criterion passages). Those comprehension tests can be a multiple choice tests assessing the readers understanding, also the so called Cloze Test (see below), or expert rating of the the difficulty. Also concurrent validity - that is the analysis of correlation with another reading measure - is a common method to verify a formula [13].

The Cloze Test was proposed by Taylor in 1953 [30] when he wanted to oppose the then widely used readability formulas. He reasoned that text difficulty was rather influenced by the relationship among words and not by the words on itself. So he proposed a deletion test, where words of a passage where randomly deleted and had to be filled in by the reader. The amount of words that can be deleted without influencing the readers understanding than influences the final Cloze Score. However, this test can by its nature not be used instead of the formulas, since it is relying on a large panel of testing readers. By giving a better alternative to the then widely used multiple choice tests it does still however have a big influence on readability measures as it is widely used for testing the correlation of readability formulas [6].

Most readability formulas are based on two different kinds of inputs that have been verified as the text features best predicting a text's readability by most of the research studies:

1. measures based on syntactic difficulty,
2. measures based on semantic difficulty.

The first refers to the grammatical complexity and is in most formulas measured by the sentence length (e.g. average words per sentence). This rather simple metric functions quite well as a predictor, even when compared to more complex syntactic measures [6]. The second refers to the meaning of the words. A common measure among many formulas is the word length (e.g. measured in syllables or number of letters in a word). Another approach is the word frequency - this can be realized by comparing the appearance of the words of the text to the frequency of the words in other pre-classified texts (see language models), or the fact that words of the text do or don't appear on a list of easy words (see Dale-Chall Readability index in 2.2). The Bermouth Formula even combines both approaches for semantic difficulty: word length and a word list (see [6], and [13]).

One of readability formula's most common usage is in education, to help students to learn to read better by finding the right book for for a student's current reading level. But besides in schools, readability has also been in interest by companies such as General Motors, as well as the U.S. Army and the U.S. Navy. In the United States readability is even required by law for insurance policies in some states. In 1984 the New York federal appellate court ruled in a case where a letter by the U.S. Department of Health was too difficult to read [13].

Below a selection of the most popular readability formulas is described regarding its history, included features and the calculation.

Flesch Reading Ease The Flesch Reading Ease is one of the most used and tested readability measures. It was developed by Rudolf Flesch in 1949 and is based on the number of words, sentences and syllables. The original result is a number from 0 to 100, with 0 as the most difficult and 100 the easiest readability score [10].

$$FRE = 206.835 - 1.015 \cdot \frac{\text{total words}}{\text{total sentences}} - 84.6 \cdot \frac{\text{total syllables}}{\text{total words}} \tag{1}$$

Flesch-Kincaid Grade Level The Flesch-Kincaid formula is a modified version of the original Flesch formula. The intention was a mapping of grade levels to the results of the Flesch reading ease. It was developed in 1976 as part of a study by the U.S. Navy [10]. The grade level can be considered as being the number of school years that someone needs to comfortably read a text (according to the American school system: e.g. 12 for 12th grade, or 13 for first year of college). This concept is used as discussed in the following by several measures.

$$FK - Grade = 0.39 \cdot \frac{\text{total words}}{\text{total sentences}} + 11.8 \cdot \frac{\text{total syllables}}{\text{total words}} - 15.59 \tag{2}$$

Rate Index (RIX) The Rate Index (RIX) is one of the more simple readability tests. It can not only be applied to English, but also on most of the western European languages [17].

$$RIX = \frac{\text{long words}}{sentences}, \tag{3}$$

$$\text{where long words} = \text{words where number of characters} > 6$$

The usual result of the formula is between 0 and 55+. Where 0 is very easy and 55 or more very difficult [17]. The RIX is based on the *LIX* formula which was developed for the Swedish language.

Coleman Liau Index The Coleman-Liau index (CLI) was introduced 1975. The intention of this formula is a simple mechanism to calculate the readability level without counting syllables. That means the automated generation of the CLI was quite simple and did not need any manual counting of text parts. The result from the Coleman Liau Index results into a grade level [17].

$$CLI = 5.89 \cdot \frac{characters}{words} - 30 \cdot \frac{sentences}{words} - 15.8 \tag{4}$$

Gunning Fog Index The Gunning Fog Index gives also results into a grade leven and also includes sentence length and word length as parameters [17].

$$GFI = 0.4 \cdot \frac{words}{sentence} + 100 \cdot \frac{\text{complex words}}{words}, \tag{5}$$
$$\text{where complex words} = \text{words with three or more syllables}$$

SMOG G. Harry McLaughlin published his SMOG formula in 1969. By counting the number of words of more than two syllables (polysyllable count) in 30 sentences, he provides this simple formula:

$$\text{SMOG} = 3 + \sqrt{\text{polysyllable count}} \tag{6}$$

The SMOG formula generally predicts scores at least two grades higher than the Dale-Chall formula [10].

ARI Smith and Senter created the Automated Readability Index in 1967 for the U.S. Army [10]. They used an modified electric typewriter to derive the necessary counters for words and sentences. The ARI formula produces reading grade levels:

$$ARI = 0.50 \cdot \text{words per sentence} + 4.71 \cdot \text{strokes per word–21.43} \tag{7}$$

New Dale Chall The Dale-Chall Readability index uses a word list of easy, ”familiar” words to express the semantic difficulty. For the New Dale-Call Readability Formula, Chall and Dale updated their list of 3000 easy words in 1995 from their original formula of 1948. The 3000 words are known to 80 percent of students in 4th grade and may hence be viewed as the most elemental words in the English language. This can basically be seen as a simple language model, as here the list of 3000 words is classified as ”easy” - hence a text that uses more of these words can also be seen as easy (and vice versa for ”hard” words). The new formula correlates .93 with the mean Cloze Scores, making it the most valid of the popular formulas when compared to other of the classic formulas [6] [11].

$$\text{Raw Score} = 0.1579 \cdot PDW + 0.0496 \cdot \text{Average Sentence Length}, \tag{8}$$

$$\text{where PDW} = \frac{\text{words not on Dale-Chall list}}{totalwords}$$

If PDW is greater than 5%, then:

$$\text{Adjusted Score} = \text{Raw Score} + 3.6365\text{, otherwise}$$

$$\text{Adjusted Score} = \text{Raw Score}$$

The according grade level can be obtained from table 1.

Table 1. Dale-Chall adjusted scores [11].

ADJUSTED SCORE	GRADE LEVEL
4.9 and Below	Grade 4 and Below
5.0 to 5.9	Grades 5 - 6
6.0 to 6.9	Grades 7 - 8
7.0 to 7.9	Grades 9 - 10
8.0 to 8.9	Grades 11 - 12
9.0 to 9.9	Grades 13 - 15 (College)
10 and Above	Grades 16 and Above (College Graduate)

Fry Readability Graph The above measures all produce a numeric measure directly, often with the score expressed by a grade level. The Fry Readability Graph as published in 1977 plots the readability on a chart that has the length of words on the x-axis and the sentence length on the y-axis. The grade level is then determined graphically through the interception of the values. Since its use is seen as easy it is, too, one of the most widely used methods [13] [11]. Figure 1 shows the chart used. For examples with values see the results section 4.

2.3 State of the art approaches to assessing readability

Recent approaches are generally still using the ideas as mentioned above. For example Si et al. [26] and Collins-Thompson et al. [8] extend the idea of the word list for assessing the semantic difficulty (which can be viewed as simple language model) to complete unigram language models, which represent the frequency distribution of words in for given texts. They then both use a supervised classification algorithm to classify web sites that where pre-labeled based on expert judgment.

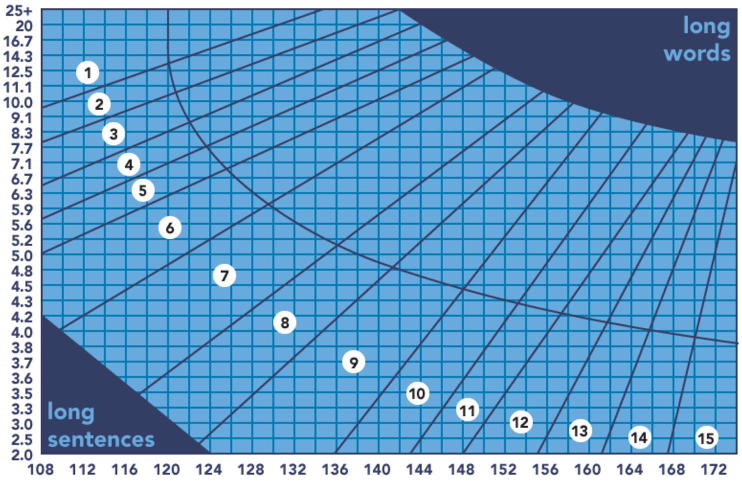


Fig. 1. An example of a Fry Readability Graph.

Schwarm and Ostendorf [25], Feng et al. [12] and Heilman et al. [14] do also use language models, but do as many of the classic formulas also include a measure for syntactic difficulty. They included in addition to the input selected by the word frequencies, also syntactic information obtained by Part-of-Speech tagging - this is the mapping of a labels such as "subject", "object" and "predicate" to the words of a sentence (POS-tagging). This can be represented in a so called parse tree (see figure 2).

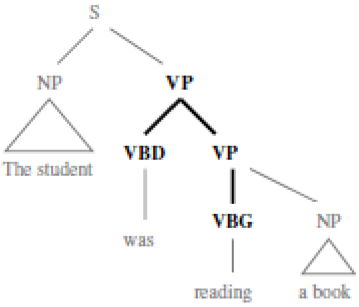


Fig. 2. An example parse tree (source: [14]).

Other than the above mentioned approaches, Kanungo and Orr [19] did not use a classifier, but a regression model, and focused on a more narrow application: short web summaries. As input for the model they use different values for the semantic difficulty (average characters per word, average syllables per word, percentage of complex words), different features specifically for this task (e.g. Capital Letter Fraction) and even other readability formulas (FOG, Flesch and Flesch-Kincaid).

Kate et al. [22] use a variety of syntactic, semantic (from different language models: e.g. genre-specific and not genre-specific) and lexical features also using a regression model, and show that the combination of these different kinds of features result in a higher correlation than any of the features on their own.

Xing et al. [34] further improve the idea of a language model. Instead of using simply a unigram model, they use a bigram model reasoning that the order of words influences the information contained by a sentence and hence influencing readability - e.g. if you reverse a sentence a classical measure such as the Flesch score will not change while you can certainly assume that the sentence is then less readable.

Petersen and Ostendorf [24] also combine a variety of features. They extend the language model to n-gram language models, also include POS-parsing information and even classical measures in their model (e.g. Flesch-Kincaid score, average sentence length, average number of syllables per word).

As all the approaches introduced in this section need a big sample of already labeled training data (similar to the criterion passages as discussed with the classical measures), these approaches are out of scope of this paper. This paper will therefore focus on the classical measures, and on the visualization of readability measures. We will provide an approach to enhance the Dale-Chall Measure, as one of the more sophisticated and most accurate classic approaches, and validated it against the other measures (see section 3.5).

2.4 Visualization of readability

Besides measures where the actual method consists of a graphical visualization such as the Fry Readability Graph, research was e.g. done by Karmakar and Zhu [21] [20] in further visualizing readability. Karmakar and Zhu [20] propose two different visualizations for numeric readability formulas (they use the Flesch-Kincaid Reading Ease index, the Gunning Fog index, the Coleman-Liau index and the SMOG index), focusing on the comparison of multiple paragraphs of a text to quickly identify sections with low readability:

- color coded representation,
- and Chernoff faces.

For the first they categorized the readability levels into five categories ranging from red (the hardest), to yellow (neutral), to green (the easiest). The readability is then displayed as a colored ring or as a colored string each representing a measure. The Chernoff faces [7] are cartoon-like faces that can visualize multiple variables with different parts of the face. Here they mapped each readability score to a part of the face (e.g. Flesch-Kincaid mapped to the shape of the faces (circular to oval), ARI is mapped to the orientation of the mouth (smiling to sad), etc.).

Karmakar and Zhu [21] also focus on simplifying the quick visual assessment of the reading difficulty of a text, where differences within the text can easily be spotted. They propose three different visualizations. They base on the classical reading formula components as mentioned above: word complexity (here word length in syllables and letters, and words familiarity as in Dale-Chall) and sentence complexity. Figure 3 shows an example representation proposed by them. For sentence complexity they do not only use the common metric sentence length, but try to visualize the structural complexity by visualizing the division of clauses in a sentences. The other representation is the visualization of the Dale-Chall Words in a sentence (red not included, grey is included and black is numeric), and also the sentence length. This representation is also implemented in the tool proposed. For an example see section 4.

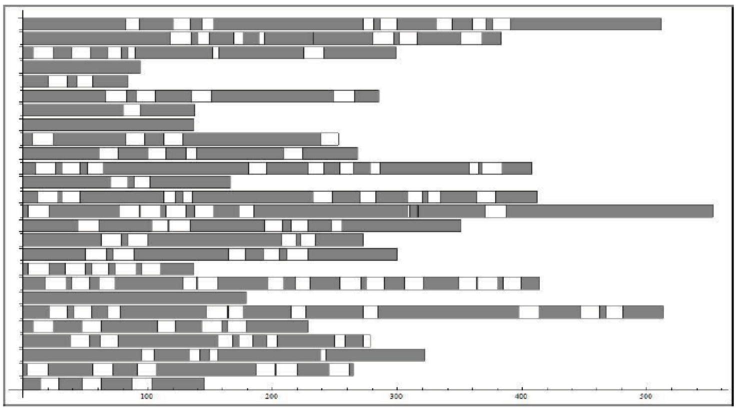


Fig. 3. Visual representation of text complexity (source: [21]).

3 Method

3.1 Python, Libraries and Frameworks

As the programming language we choose Python because of the existing frameworks, libraries and the ease of use that made the development process modest. As the development environment we used the Enthought Python Distribution (EPD)[1] which includes Python 2.7 and over 100 common libraries which are useful for scientific computing.

We used the following libraries for our Privacy Policy Analyzer:

[1] http://enthought.com/products/epd.php

- For the visualization respectively the generation of the various charts and diagrams Matplotlib is used.[2]
- NumPy and SciPy provide fundamental functions for scientific computing (e.g. statistics), and provide the array data type which is needed for Matplotlib plots.[3]
- We decided to generate our reports with the result of the analytics as simple html files, with a pdf print option. For the template parsing and automated generation of these reports we deployed the python template engine "Cheetah".[4]

Most important is the Natural Language Toolkit (NLTK)[5] which provides core features for the analyzer. The NLTK is a framework with capabilities for detailed text analysis and natural language processing. For the project we used existing code from the NLTK contribution package, where several readability measures were implemented by students of the University of Agder.[6] This existing code already implements a lot of basic measures and syllable, sentence and word counting as well.

The existence of the NLTK with the readability contribution and the simple handling of lists, string manipulations and general simplicity were the reasons why we decided to build our project on the python programming language.

3.2 Implementation

Figure 4 shows a diagram depicting the architecture of the relevant parts of the program "Privacy Policy Analyzer" in an UML-like notation. The parts within the boundaries of the box are all complete novel and created by the authors. Additionally we used, as described above, parts of the Python NLTK contribution package (outside the box), and the NLTK for the enhancement of Dale-Chall. To the code of the contribution package only a few functions where added and minor improvements where done to the code. Table 2 shows the added functions.

The class *readabilityanalyzer* is the main class used to integrate all parts of the program, and can be used by the user to interact with all functionalities. Alternatively the user can use most functionalities also by using a simple GUI provided by the *input* class. The classes included in the *database* package provide for all functionalities needed to interact with the database storing the texts obtained from the websites. The *presentation* package provides several graphs based on the Matplotlib library for the data provided. The packages *output*

[2] `http://matplotlib.sourceforge.net/`

[3] `http://numpy.scipy.org/`

[4] `http://www.cheetahtemplate.org/`

[5] `http://www.nltk.org/`

[6] `http://code.google.com/p/nltk/source/browse/trunk/nltk_contrib/nltk_contrib/#nltk_contrib\%2Freadability`

Table 2. Functions added to the existing NTLK contribution source code.

In textanalyzer.py the following functions where added:	
def countNewDaleChallWordsNotInList	returns the Number of Words not in the original Dale-Chall-List (see section 3.5)
def countNewDaleChallWordsNotInList_enhanced	returns the Number of Words not in the enhanced Dale-Chall-List (see section 2.2)
In readabilitytests.py the following functions where added:	

def NewDaleChall
def NewDaleChall_grade
	return the original Dale-Chall score and grade using the above functions

def NewDaleChall_enhanced
def NewDaleChall_grade_enhanced
	return the enhanced Dale-Chall score and grade using the above functions

get_test_scores	returns the all test scores as a dictionary (before the results where only printed on the console)

and *output_comparison* include all functionalities needed to produce the HTML reports.

3.3 Visualization

The reports produced with this work focus much on visualizing readability, to more easily compare the readability among different privacy policies. This is done for one thing with a graphical readability measure (Fry Readability Graph, see section 2.2), a visualization of a classical numeric readability measure (see section 3.3 as in [21]), and different graphs that simply each visualize different numeric measures for better comparison.

The heat map for example makes it easy to quickly display different numeric measures, and hence compare either different measures for a text, or even see how a measure differs among different texts. It uses a color coding to indicate difficulty, with red being the hardest blending to orange and yellow for medium difficulty and then blending to green as the easiest (see section 7 for an example).

Other plots where implemented, though they are not used in the analysis of this paper. The radar chart is another plot to compare texts, although it is only useful for a smaller amount of texts. A larger area indicates higher scores and therefore a harder readability level. Additional a few statistical plots allow to analyze how the scores for the different measures are distributed among a larger number of texts.

3.4 Reports

The tool "Privacy Policy Analyzer" can generate two different reports depending on the objective. In general the reports are automatic generated HTML files which every common browser can display. The report on the left in figure 5

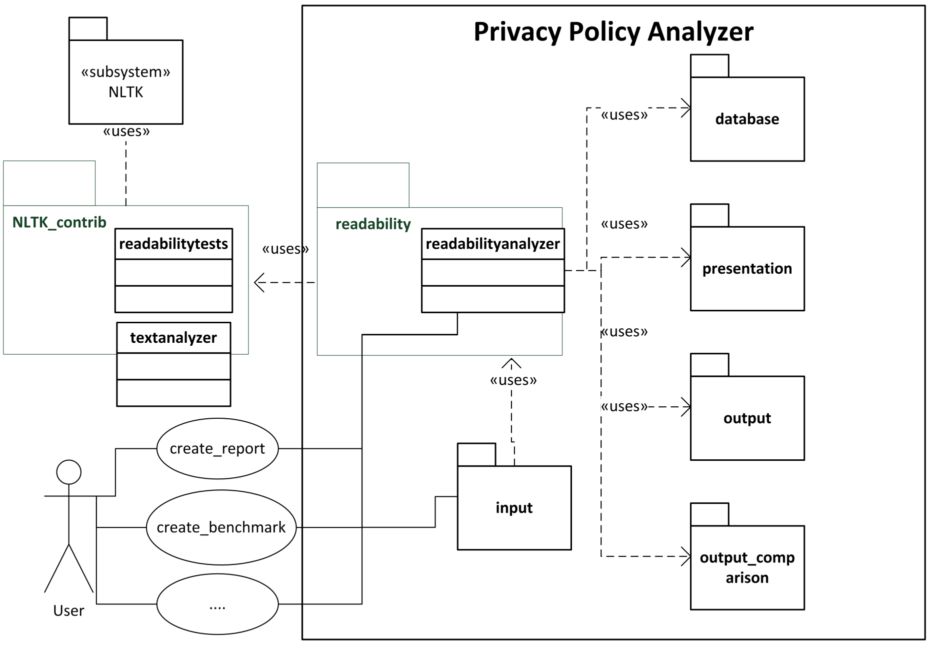


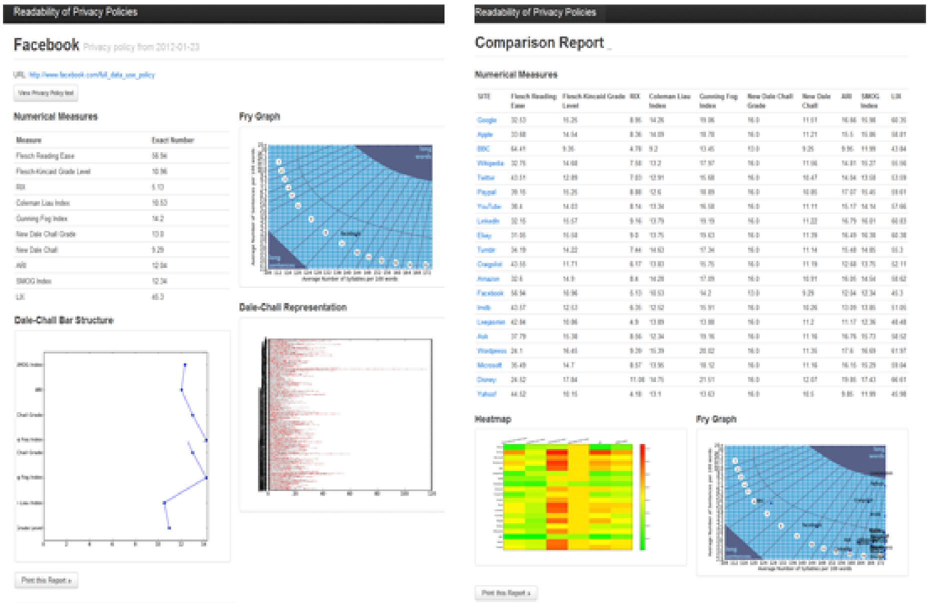
Fig. 4. Architecture of the Privacy Policy Analyzer.

shows the results of the analysis of an single privacy policy. It contains some metadata, like name and date of the polices and the original text as well. After the metadata a table with the results of the different measures and the generated visual results are following. The second report on the right hand in figure 5 is a comparison report, which compares multiple privacy policies against each other. The structure is similar to the single report, including metadata, an overview table of the numeric measure results and some visualizations of the comparison.

3.5 Adjusting New Dale-Chall Readability index for assessing readability of privacy policies

The Dale-Chall Readability index is based on the average length of sentences on the one hand and a word list on the other (see 2.2). The word list with 3000 words represents the most *elemental words* in the English language. The Dale-Chall is as mentioned viewed as one of the most accurate classical readability measures. However, when applied to the privacy policies the scores are all extremely high and show just few variability among the privacy policies.

An analysis of the word frequency distribution shows that among the most frequent words (after elimination of all Dale-Chall words) in the privacy policies are on the one hand the names of the corresponding company, and also a set of vocabulary that is frequently used among the policies. Among some of the site's



Readability of Privacy Policies

Facebook Privacy policy from 2012-01-23

URL: http://www.facebook.com/full_data_use_policy

View Privacy Policy text

Numerical Measures

Measure	Exact Number
Flesch Reading Ease	56.54
Flesch-Kincaid Grade Level	10.96
RIX	5.13
Coleman Liau Index	10.53
Gunning Fog Index	14.2
New Dale Chall Grade	13.0
New Dale Chall	9.29
ARI	12.04
SMOG Index	12.34
LIX	45.3

Readability of Privacy Policies

Comparison Report

Numerical Measures

SITE	Flesch Reading Ease	Flesch Kincaid Grade Level	RIX	Coleman Liau Index	Gunning Fog Index	New Dale Chall Grade	New Dale Chall	ARI	SMOG Index	LIX
Google	32.53	15.25	8.96	14.26	19.06	16.0	11.51	16.66	15.98	60.35
Apple	33.68	14.54	8.36	14.09	18.78	16.0	11.21	15.5	15.86	58.81
BBC	64.41	9.35	4.78	9.2	13.45	13.0	9.25	9.95	11.99	43.84
Wikipedia	32.76	14.68	7.58	13.2	17.97	16.0	11.06	14.81	15.27	55.06
Twitter	43.51	12.89	7.03	12.91	15.68	16.0	10.47	14.34	13.58	53.09
Paypal	39.16	15.25	8.88	12.6	18.89	16.0	10.85	17.07	15.45	59.61
YouTube	38.4	14.03	8.14	13.34	16.58	16.0	11.11	15.17	14.14	57.66
LinkedIn	32.15	15.57	9.16	13.79	19.19	16.0	11.22	16.79	16.01	60.83
Ebay	31.85	15.58	9.0	13.75	19.63	16.0	11.39	16.49	16.38	60.38
Tumblr	34.19	14.22	7.44	14.63	17.34	16.0	11.14	15.48	14.85	55.3
Craigslist	43.55	11.71	6.17	13.83	15.75	16.0	11.19	12.68	13.75	52.11
Amazon	32.6	14.9	8.4	14.28	17.09	16.0	10.91	16.05	14.64	58.62
Facebook	56.54	10.96	5.13	10.53	14.2	13.0	9.29	12.04	12.34	45.3
Imdb	43.57	12.53	6.35	12.52	15.91	16.0	10.26	13.09	13.85	51.05
Livejournal	42.84	10.86	4.9	13.89	13.88	16.0	11.2	11.17	12.36	48.48
Ask	37.79	15.38	8.56	12.34	19.16	16.0	11.16	16.76	15.73	58.52
Wordpress	24.1	16.45	9.39	15.39	20.02	16.0	11.35	17.6	16.69	61.97
Microsoft	35.49	14.7	8.57	13.95	18.12	16.0	11.16	16.15	15.29	59.04
Disney	24.52	17.84	11.08	14.75	21.51	16.0	12.07	19.85	17.43	66.61
Yahoo!	44.52	10.15	4.18	13.1	13.63	16.0	10.5	9.85	11.99	45.98

Fig. 5. Two reports generated by the Privacy Policy Analyze. Left a comparison, right a single report.

names, some of the most frequent words are e.g. information, personal, services, data or privacy. These, and the other words, could be considered as essential in order to write a privacy policy. Assuming that you cannot write a privacy policy without a certain set of words it would not to be right to ”punish” a site with a bad readability score, when using it. This set of words should be viewed as *elemental words* in a privacy policy, and hence be required as known vocabulary to anybody attempting to read it.

Figure 6 shows the 80 most frequent words obtained from a sample of 20 of the most popular web sites (see 3.6), after deleting the Dale-Chall words. After the deleting procedure about 18,000 words were left. The plot shows the cumulative count of the 80 most frequent words. For this attempt to enhance the measurement we chose a cumulative percentage of 33% of the words left after deleting the Dale-Chall words. This left a list of 54 words that would for this purpose be added to the Dale-Chall list of words. Hence those 54 words make up about 6,000 of the 18,000 words left.

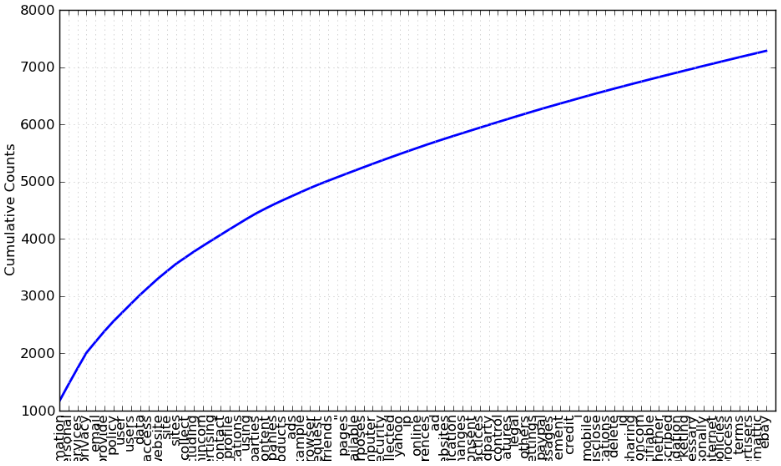


Fig. 6. Word frequency for Top 20 privacy policies without Dale-Chall words.

The limitation of this approach of adapting the readability lay in the missing validation through objective assessment of the readability. As discussed in section 2.2 this could be done through a study using a comprehension test. This however is out of scope for this paper. An attempt of concurrent validity is given by analyzing correlation with other readability scores in section 4.3.

3.6 Data

With the goal of having a data basis to analyze, we decided to collect the privacy policies from the most used English websites. As an indicator for most used respectively most visited webpages we relied on the Alexa Web Rank.[7] The Alexa Web Rank provides the top websites on a three-month schedule, which is sufficient for our purpose. We collected 20 of the top 50 policies in our database, as metadata we also saved the date, URL and name of the website. Many of the top 50 where left out, as there are many "dublicates" within the top 50 due to different country versions of site, or sites in languages other than English. We decided to use the top 20 web pages because we consider the privacy policy to be more important as the usage of the site increases.

For comparison reasons we decided to add four non privacy policy texts. We added two different classical children's stories. The first one is "A Christmas Tree" written by Charles Dickens [9] and the second one is "The Giving Tree" by Shel Silverstein [27]. Additionally we added also two old legislative texts, the "Declaration of Independence" from 1776 [3] and the "Bill of Rights" [31] written in 1789.

4 Results

4.1 Top 20 Websites

The table 3, the heat map in figure 7, the Fry Chart in figure 8 and the Dale-Chall chart in figure 9 show the results of an evaluation with the Privacy Policy Analyzer. The table displays the original numerical results that were obtained and the heat map is provided for a better comparison of the results by visualizing them. The Fry Chart and the Dale-Chall Chart are further visual aids to better compare the readability as mentioned in 3.3.

In the table 3 the bold marked numbers indicate the simplest text difficulty for each score. Italic marked numbers in contrast indicate the hardest text difficulty for each measure. The fact that the hardest 2 are generally distributed among 2 policies and the worst are distributed among 3 indicates that the scores show some correlation among the various measures. There are however some deviations existent. These are strongest for the Dale-Chall and the enhanced version of the Dale-Chall, but over all one can see that the very easiest, and respectively, the very hardest readable policies and texts stay consistent. For a thorough analysis of the correlation between the measures see section 4.3. The heat map in figure 7 also underpins this relationship described.

Both the table and the heat map show that Wordpress.com and Disney.com consistently are the hardest-to-read policies from our database. The best results for readability of their privacy policies have the websites Yahoo.com, Facebook.com and BBC.co.uk. At least for BBC.co.uk and Facebook.com the Fry

[7] `http://www.alexa.com/topsites`

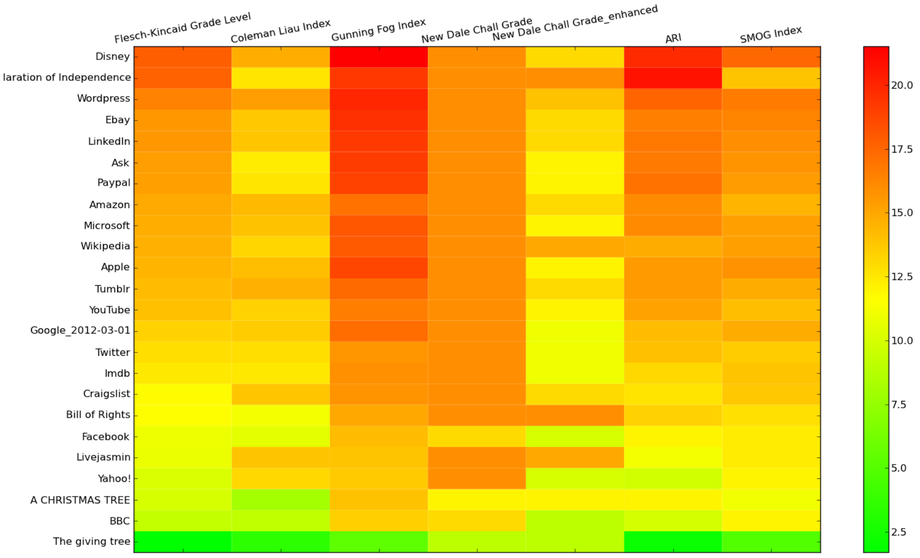


Fig. 7. Heatmap for Top 20 websites

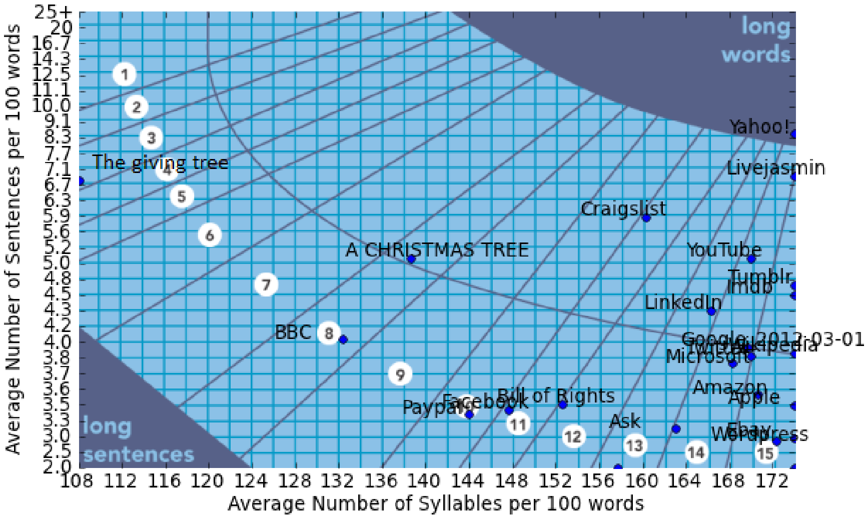


Fig. 8. Fry Chart with the Top 20 websites and four comparison texts.

Table 3. Measure Results for the Top 20 websites

	Flesch Reading Ease	**Flesch-Kincaid**	**RIX**	**Coleman-Liau**	**Gunning Fog**	**New Dale Chall**	**New Dale Chall Grade**	**New Dale Chall Score en-hanced**	**New Dale Chall grade en-hanced**	**ARI**	**SMOG**	**LIX**
Wordpress	*24.03*	*16.46*	*09.39*	*15.39*	*20.05*	11.35	*16*	09.50	14	*17.60*	*16.71*	*61.97*
Disney	*24.52*	*17.84*	*11.08*	*14.75*	*21.51*	*12.07*	*16*	09.27	13	*19.85*	*17.43*	*66.61*
Ebay	31.05	15.58	09.00	13.75	19.63	11.39	*16*	09.28	13	16.49	16.38	60.38
LinkedIn	32.15	15.57	09.16	13.79	19.19	11.22	*16*	09.30	13	16.79	16.01	60.83
Amazon	32.60	14.90	08.40	14.28	17.09	10.91	*16*	09.05	13	16.05	14.54	58.62
Wikipedia	32.75	14.68	07.58	13.20	17.97	*11.56*	*16*	*09.89*	*15*	14.81	15.27	55.56
Apple	33.68	14.54	08.36	14.09	18.78	11.21	*16*	08.60	12	15.50	15.86	58.81
Tumblr	34.19	14.22	07.44	14.63	17.34	11.14	*16*	09.11	13	15.48	14.85	55.30
Microsoft	35.49	14.70	08.57	13.95	18.12	11.16	*16*	08.94	12	16.15	15.29	59.04
Ask	37.79	15.38	08.56	12.34	19.16	11.16	*16*	08.92	12	16.76	15.73	58.52
YouTube	38.40	14.03	08.14	13.34	16.58	11.11	*16*	08.97	12	15.17	14.14	57.66
Google_2012-03-01	39.05	13.32	07.13	13.57	17.28	11.00	*16*	08.47	11	14.23	14.82	54.32
Paypal	39.15	15.25	08.88	12.60	18.89	10.85	*16*	08.90	12	17.07	15.45	59.61
Livejasmin	42.84	10.86	04.90	13.89	13.88	11.20	*16*	*09.96*	*15*	11.17	12.36	48.48
Twitter	43.51	12.89	07.03	12.91	15.68	10.47	*16*	08.37	11	14.04	13.58	53.59
ref Craigslist	43.55	11.71	06.17	13.83	15.75	11.19	*16*	09.09	13	12.68	13.75	52.11
Imdb	43.57	12.53	06.35	12.52	15.91	10.26	*16*	08.27	11	13.09	13.85	51.05
Yahoo!	44.52	**10.15**	**04.18**	13.10	**13.63**	10.50	*16*	07.92	**10**	**09.85**	**11.99**	45.98
Facebook	**56.94**	10.96	05.13	**10.53**	14.20	**09.29**	**13**	**07.70**	**10**	12.04	12.34	**45.30**
BBC	**64.41**	**09.35**	**04.78**	**09.20**	**13.45**	**09.25**	**13**	**07.14**	**09**	**09.95**	**11.99**	**43.84**
A christmas tree	71.15	10.03	04.30	08.10	13.99	08.62	12	08.65	12	12.04	11.07	42.89
The giving tree	104.96	01.70	01.39	03.46	05.34	07.27	09	07.27	09	02.11	04.88	23.63
Decl. of Independence	35.31	17.70	09.97	12.56	19.30	10.71	16	10.84	16	20.75	13.92	64.15
Bill of Rights	55.25	11.62	06.37	11.18	14.97	10.25	16	10.31	16	13.33	12.77	50.51

Charts support their good readability score (see figure 8). For Yahoo.com the Fry Chart shows that the policy contains short sentences but long words. This fact is also displayed in the Dale-Chall representation of Yahoo.com in figure 9, where it is compared to Disney.com, as one of the hardest sites, and two of the comparison texts. Here one can see that it has considerably shorter sentences as Disney.com. It shows however, that it is still a lot more difficult than the children's story "The giving tree".

The heat map and the Fry Chart also very clearly show that the children's stories stand out in comparison even with the easiest policies. The "Declaration of Independence" in contrast scores among the most difficult texts analyzed.

Generally the analysis shows that for almost every policy the scores result in grade levels (Flesch-Kincaid, Coleman-Liau, Gunning Fog, New Dale Chall, ARI, SMOG) of 13 and higher, which means that a reading capability at college level is required to comfortably read these privacy policies.

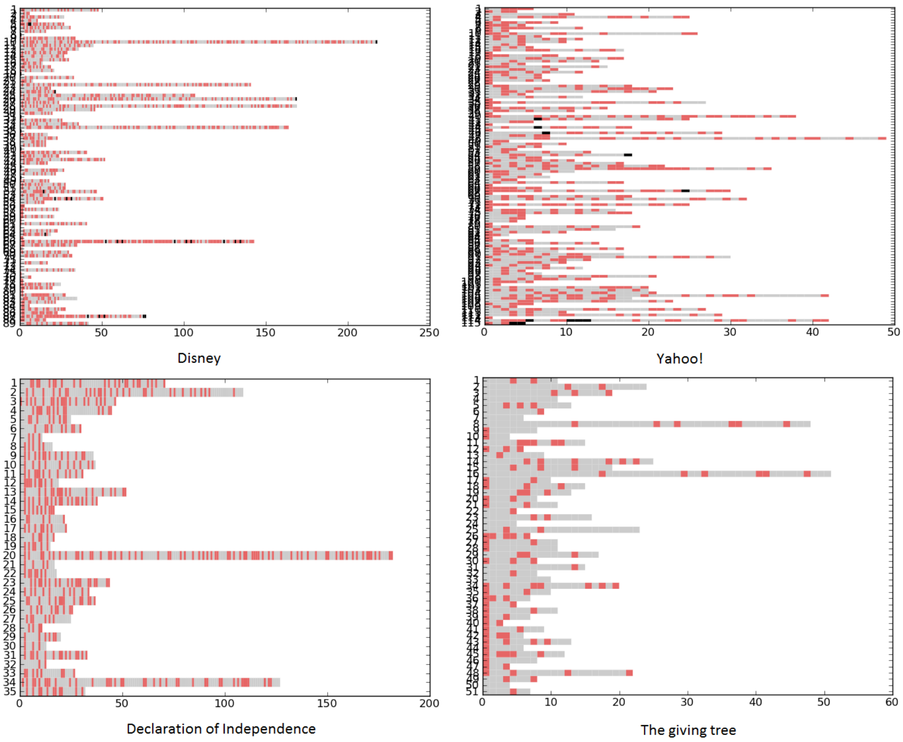


Fig. 9. Dale Chall Chart for four texts.

4.2 Google privacy policy over time

Table 4. Google privacy policies over time scores.

Date of Policy	Flesch Kincaid	Coleman Liau	Gunning Fog	NDC Score enh.	NDC	NDC Grade enh.	ARI	SMOG
2000-08-14	11,15	11,44	13,69	8,48	10,42	11	12,01	12,12
2004-07-01	12,23	12,74	16,51	8,1	10,29	11	12,93	14,3
2005-10-14	15,79	13,97	19,99	8,96	11,49	12	17,13	16,6
2008-08-07	15,81	14,08	19,97	8,92	11,59	12	17,05	16,6
2009-01-27	15,71	14,06	19,82	8,89	11,56	12	16,92	16,51
2009-03-11	16,02	14,21	20,18	8,96	11,67	12	17,32	16,75
2010-10-03	15,31	14,21	19,11	9,01	11,53	13	16,63	16,02
2011-10-20	15,25	14,26	19,06	9,04	11,51	13	16,66	15,98
2012-03-01	13,32	13,57	17,28	8,47	11	11	14,23	14,82

In the previous section we could only analyze a snapshot of the current privacy policies of the websites. As Sumeeth et al. [29] we were also interested in how privacy policies developed over time. Google provides an archive of their privacy policy since August 2000.[8] Since Google is currently the most popular web site we want to provide a thorough analysis of the readability of their privacy policy. They state that they want to be transparent as possible regarding changes. We used this possibility to compile a comparison report on the English version of the Google privacy policy to analyze changes over time with the Privacy Policy Analyzer. The input data for the analysis are nine different versions starting in 2000 until 2012 (as seen in table 4). Results are the various measures that our tool provides and visual representation through the Fry Graph in figure 11.

As one can see in the plot in figure 10 and the Fry Graph in figure 11 the privacy policy was relatively readable when starting in 2000. Table 4 provides the numbers used in figure 10. In the time frame from 2000 to 2004 no changes were applied as opposed to the time after 2008 where the policy changed at least once a year. With the changes in 2004 and 2005 one can see a clear increase in the difficulty of readability. This level stays with small changes until the second change of the policy in 2009 (2009-03-11), where the peak of all measures except New Dale Chall is reached. After this peak level the readability gets easier with almost every change to the policy. The last update from 2012-03-01 had among the consolidation of many policies also the scope to improve the readability, as Google stated [33]. Expressed in readability measures a visible improvement in comparison to the previous versions exists, but it is still slightly more difficult than the 2004 version and the 2000 version anyway. One possible

[8] http://www.google.com/intl/en/policies/privacy/archive/

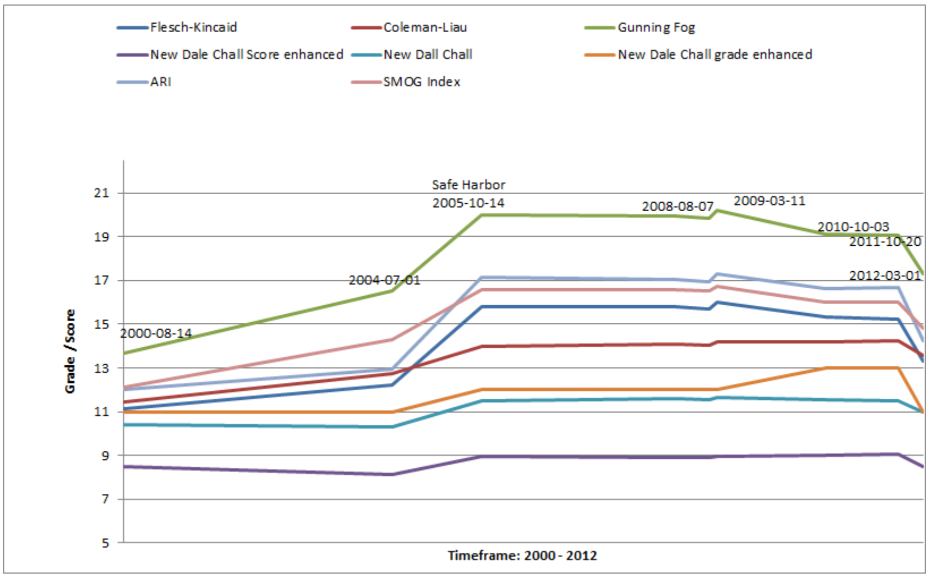


Fig. 10. Readability of Google Privacy Policies over Time. Based on data of table 4.

reason for the significant jump between 2004 and 2005 could be the fact that Google joined the U.S.-EU Safe Harbor Act in 2005 [1]. For the Safe Harbor certification a comprehensive and detailed privacy policy is mandatory [2]. Based on the limitations of this paper, the dependences between changes in the Google policy and the Safe Harbor certification are not further investigated.

The Google over time analysis also indicated the correlation between the different measures. Though they do not result in equal numeric results, they generally are all indicating the same leaps triggered by changes in the privacy policy text. The line chart gives a good visualization of this. For a deeper analysis regarding statistical correlation in general between the readability measures see section 4.3.

4.3 Validation of method

In this section we want to reflect on the method used. We want to verify the implementation of the readability measures on the one side and attempt to verify the adjustment of the Dale-Chall measure as discussed in section 3.5. As discussed in section 2.2 there are generally three widely used approaches to verify readability measures: correlation with the score of reading comprehension tests, the score of expert judgements or other readability scores. As the first two are out of the scope of this paper we will use the latter by analyzing the correlation between the implemented measures.

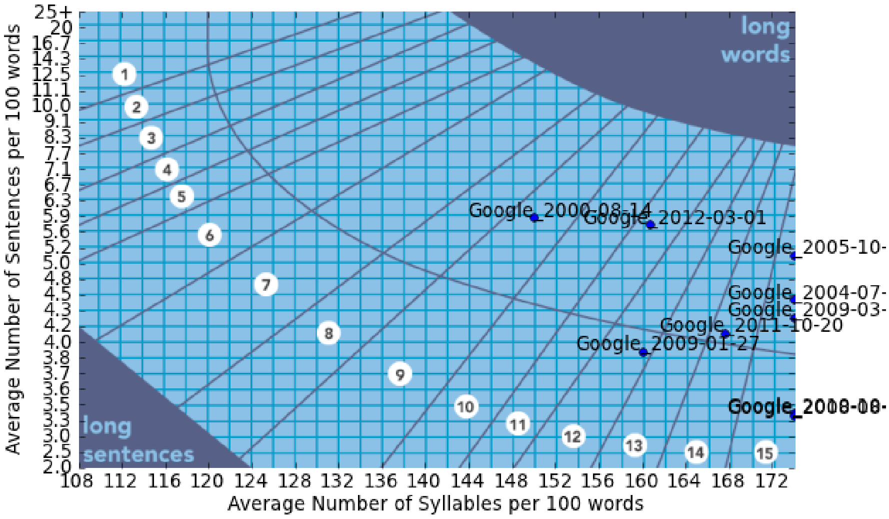


Fig. 11. Readability of Google Privacy Policies over Time.

Table 5 shows the Pearson correlation coefficients between the measures discussed in this paper and implemented the Readability Analyzer (exceptions are the visual measures such as the Fry Chart). The scores are based on the application to the 20 privacy policies and four "benchmark" texts as discussed in section 3.6. It shows that all coefficients are statistically significant at the 5% significance level (most even at the 1% level).

It shows that all the scores that are based on solely easier syntactic features, and word length as a semantic feature, correlate highly with coefficients of more than 0,8 and many even higher than 0,9. All correlate positively except with Flesch Reading Ease, as this score decreases as a text becomes more difficult as opposed to all the other scores. From those measures Coleman-Liau correlates the weakest with the other scores. Also the New-Dale Chall score with the word list as a semantic part (and its obtained grade) does correlate pretty highly with the other scores. The "enhanced" version however does not correlate much with the scores, with only low coefficients around 0,5 - 0,6. This much limits it's *objective* validity to measure the readability of Privacy Policies. To validate it further, therefore other objective approaches are needed to validate its applicability objectively.

Table 5. Pearson correlation coefficients for readability measures, where * indicates p-value of 5%, and ** indicates a p-value of 1%

	Flesch Reading Ease	**Flesch-Kincaid**	**Coleman-Liau**	**Gunning Fog**	**ARI**	**SMOG**	**LIX**	**RIX**	**NDC**	**NDC grade**	**NDC enhanced**	**NDC grade enhanced**
Flesch Reading Ease	1	-,930**	-,969**	-,922**	-,868**	-,949**	-,942**	-,865**	-,958**	-,919**	-,563**	-,508*
Flesch-Kincaid	-,930**	1	,830**	,983**	,987**	,944**	,990**	,965**	,854**	,807**	,630**	,557**
Coleman-Liau	-,969**	,830**	1	,823**	,757**	,882**	,859**	,752**	,950**	,934**	,521**	,480*
Gunning Fog	-,922**	,983**	,823**	1	,963**	,976**	,977**	,953**	,860**	,789**	,566**	,499*
ARI	-,868**	,987**	,757**	,963**	1	,897**	,973**	,965**	,785**	,744**	,653**	,576**
SMOG	-,949**	,944**	,882**	,976**	,897**	1	,950**	,907**	,903**	,834**	,484*	,432*
LIX	-,942**	,990**	,859**	,977**	,973**	,950**	1	,973**	,889**	,841**	,625**	,550**
RIX	-,865**	,965**	,752**	,953**	,965**	,907**	,973**	1	,812**	,726**	,591**	,512*
NDC	-,958**	,854**	,950**	,860**	,785**	,903**	,889**	,812**	1	,924**	,605**	,555**
NDC grade	-,919**	,807**	,934**	,789**	,744**	,834**	,841**	,726**	,924**	1	,599**	,544**
NDC enhanced	-,563**	,630**	,521**	,566**	,653**	,484*	,625**	,591**	,605**	,599**	1	,983**
NDC grade enhanced	-,508*	,557**	,480*	,499*	,576**	,432*	,550**	,512*	,555**	,544**	,983**	1

5 Conclusion

5.1 Research limitations

Due to the comprehensive aspects of research regarding readability of privacy policies some research limitations are existent. One very important component in any context with privacy policies are the legal aspects. We do not include any part of regulation and legislation into this paper because this is not in our scope. Another limitation are the state-of-the-art measures, which are not part of the Privacy Policy Analyzer, due to the fact that the training of the measures needs a big amount of pre-labeled data. A further limitation in this paper is the quality and quantity of privacy policies in our database. This state of the database is sufficient for the scope of this paper, but could be improved for further research, by collecting more policies.

5.2 Future Work

Most of the limitations provide interesting opportunities for future research efforts. Therefore we would suggest the following aspects for further investigation. There a some possibilities to improve on the topic readability of privacy policies by using our already developed tool Privacy Policy Analyzer. One could do a more sophisticated analysis with more statistical valid quantitative comparisons using a larger database of policies (e.g. number of policies > 100). The collection of a high number of privacy policies was out of scope of this paper, as insuring sufficient data quality is quite time consuming. One aspect which is not noted in this paper is the possibility to expand readability analytics with the way how a policy is presented, like how easy they are to find, which font and font-size on which background color is used etc. The legal aspects discussed above can be assumed to influence the readability of the privacy policies. The readability measures are usually aimed at grading texts at school level. Hence future work should focus on adjusting them to suit the task of rating privacy policies better. We provided one approach in adjusting the Dale-Chall word list in 3.5. Future research could focus on using approaches similar to those discussed in the section 2.3. Those require an empirical classification of the readability of the texts e.g. with surveys where privacy policies are manually rated. These results could then be used as training data for the state-of-the-art measures which can then produce results better suited for this specific task.

5.3 Findings

In this paper we developed the tool Privacy Policy Analyzer, which can analyze privacy policies based on different measures for text difficulty. We showed that the readability level of the most privacy policies from popular websites is at college level or above. In coherence with the average age of the internet user and their education one essential statement is, that in general privacy policies compared to the reading comprehension level of the internet user are to difficult. An improvement of readability is necessary to cover most internet users and give them a chance of understanding privacy policies easily. Using Google as an example we show that an improvement of readability can be achieved.

References

1. Safe harbor information: Google inc., http://safeharbor.export.gov/companyinfo.aspx?id=13346
2. U.s. eu safe harbor overview, http://export.gov/safeharbor/eu/eg_main_018476.asp
3. The declaration of independence (1776), http://www.ushistory.org/declaration/document/
4. Agrawal, R., Grosky, W.I., Fotouhi, F.: Ranking privacy policy. In: Proc. IEEE 23rd Int Data Engineering Workshop Conf. pp. 192–197 (2007)
5. Anton, A.I., Earp, J.B., He, Q., Stufflebeam, W., Bolchini, D., Jensen, C.: Financial privacy policies and the need for standardization. IEEE Security & Privacy 2(2), 36–45
6. Chall, J., Dale, E.: Readability Revisited: The New Dale-Chall Readability Formula. Brookline Books, Brookline, Massachusetts (1995)
7. Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. Journal of the American Statistical Association 68(342), 361–368 (June 1973)
8. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: HLT-NAACL. pp. 193–200 (2004)
9. Dickens, C.: A Christmas Tree. Slawson Communications, Inc. (1988), http://www.classicreader.com/book/876/1/
10. Dubay, W.H.: The principles of readability. CA Impact Information (949), 1–72 (2004), http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Principles+of+Readability#2
11. Dubay, W.H.: Smart Language: Readers, Readability, and the Grading of Text. Booksurge Llc (2007), http://books.google.de/books?id=3in5PAAACAAJ
12. Feng, L., Jansche, M., Huenerfauth, M., Elhadad, N.: A comparison of features for automatic readability assessment. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 276–284. COLING '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1944566.1944598
13. Fry, E.: Readability (February 2006), http://www.impact-information.com/impactinfo/fryreadability.pdf
14. Heilman, M.J., Collins, K., Callan, J.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Proceedings of the Human Language Technology Conference (2007)
15. Hochhauser, M.: The effects of hipaa on reserch consent forms. Patient Care Management (2002)
16. Huning, T.R.: Readability of privacy policies of german websites (2010), unpublished.
17. Jakobsen, T., Skardal, T.: Readability index (2007), http://www.mortengoodwin.net/publicationfiles/webmining_2007_reportgroup6.pdf

18. Jensen, C., Potts, C.: Privacy policies as decision-making tools: an evaluation of online privacy notices. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 471–478. CHI '04, ACM, New York, NY, USA (2004), http://doi.acm.org/10.1145/985692.985752
19. Kanungo, T., Orr, D.: Predicting the readability of short web summaries. In: WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining. pp. 202–211. ACM, New York, NY, USA (2009), http://dx.doi.org/10.1145/1498759.1498827
20. Karmakar, S., Zhu, Y.: Visualizing multiple text readability indexes. In: Proc. Int Education and Management Technology (ICEMT) Conf. pp. 133–137 (2010)
21. Karmakar, S., Zhu, Y.: Visualizing text readability. In: Proc. 6th Int Advanced Information Management and Service (IMS) Conf. pp. 291–296 (2010)
22. Kate, R., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R., Roukos, S., Welty, C.: Learning to predict readability using diverse linguistic features. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 546–554. Beijing, China (August 2010), http://www.patwardhans.net/papers/KateEtAl10.pdf
23. Mcdonald, A.M., Reeder, R.W., Kelley, P.G., Cranor, L.F.: A comparative study of online privacy policies and formats. In: Proceedings of the 9th International Symposium on Privacy Enhancing Technologies. pp. 37–55. PETS '09, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-03168-7_3
24. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Tech. rep. (2006)
25. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 523–530 (2005)
26. Si, L., Callan, J.: A statistical model for scientific readability. In: Proceedings of the tenth international conference on Information and knowledge management. pp. 574–576. CIKM '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/502585.502695
27. Silverstein, S.: The Giving Tree. Harper & Row (1964)
28. Singleton, S.M., Harper, J.: With a grain of salt: What consumer privacy surveys don't tell us. (2001), http://ssrn.com/abstract=299930
29. Sumeeth, M., Singh, R.I., Miller, J.: Are online privacy policies readable? IJISP pp. 93–116 (2010)
30. Taylor, W.L.: Cloze procedure: A new tool for measuring readability. Journalism Quarterly 30(4), 415–433 (1953), http://doi.apa.org/?uid=1955-00850-001
31. of the United States, C.: Bill of rights (1789), http://www.archives.gov/exhibits/charters/bill_of_rights_transcript.html
32. Vu, K.P., Chambers, V., Garcia, F., Creekmur, B., Sulaitis, J., Nelson, D., Pierce, R., Proctor, R.: How users read and comprehend privacy policies. In: Smith, M., Salvendy, G. (eds.) Human Interface and the Management of Information. Interacting in Information Environments, Lecture Notes in Computer Science, vol. 4558, pp. 802–811. Springer Berlin / Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-73354-6_88
33. Whitten, A.: Updating our privacy policies and terms of service (2012), http://googleblog.blogspot.com/2012/01/updating-our-privacy-policies-and-terms.html
34. Xing, F., Cheng, D., Pu, J.: A new approach to readability study based on information computing. In: Proceedings of the 2008 International Conference on Advanced

Language Processing and Web Information Technology. pp. 156–161. IEEE Computer Society, Washington, DC, USA (2008), http://dl.acm.org/citation.cfm?id=1441420.1441517