Sorry, I Don't Get It: An Analysis of Warning Message Texts

Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith

Distributed Computing and Security Group, Leibniz University Hannover
Schlosswender Str. 5, 30159 Hannover, Germany
{harbach,fahl,yakovleva,smith}@dcsec.uni-hannover.de

Abstract. Security systems frequently rely on warning messages to convey important information, especially when a machine is not able to assess a situation automatically. There is a significant body of work studying the effects of warning message design on users with numerous suggestions on how to optimise their effectiveness. Design guidelines and best practises help the developer to display urgent information. In this paper, we present the first empirical analysis on the extent of the influence of linguistic properties on the perceived difficulty of the descriptive text in warning messages. We evaluate warning messages extracted from current browsers and present linguistic properties that can improve a warning message text's perceived difficulty. Our results confirm that, while effects of attention, attitude and beliefs are at least as important as the linguistic complexity of the text, several steps can be taken to improve the text's difficulty perceived by the user.

Keywords: Usable Security, Comprehension, Warning Messages, Readability

1 Introduction

Designing and writing warning messages can be considered a form of art. In the past, users and IT professionals alike were confused by complicated warning and error messages that seemed to consist of only hex numbers and stack traces, such as the famous "blue screen of death". A considerable amount of work has continuously improved the quality of warning messages for many different applications and proposed guidelines on how to compose useful and understandable dialogues (e. g. [3, 6, 13]). However, users still seem to struggle with warnings on a regular basis, suggesting that there are still open problems in creating understandable and helpful warning messages.

The reception of warning messages by a user is often explained using Wogalter's Communication-Human Information Processing (C-HIP) model [15] or Cranor's extension of C-HIP: the human-in-the-loop (HITL) framework [4]. In both models, information is conveyed from a source through a channel to a human receiver. At the receiving end, the information first needs to gain sufficient attention before the information enters the comprehension stage. Afterwards, attitudes and beliefs as well as motivation further influence the information before the processing results in behaviour. A lot of work has been put into optimising colours, fonts, symbols and icons to attract attention and facilitate reception.

In this paper, we investigate the comprehension stage: does the structural composition (syntax and vocabulary) of a warning message's text influence the user's overall perception and support comprehension? Or, in other words: if a user chose to read a warning message, would he or she be able to extract the necessary information and find the text easy to parse and understand?

It has been recognised that the descriptive text provided in warning messages needs to convey important information about the problem and be understandable by most computer users at the same time. In 2011, Bravo-Lillo et al. [3] compiled a set of design guidelines and present rules for descriptive text, including:

- "describe the risk; describe consequences of not complying; provide instructions on how to avoid the risk;"
- "be brief; avoid technical jargon."

However, these guidelines are hard to quantify, especially since there is no example of a perfect warning message to date. Thus, judging whether or not the requirements and advice of the guidelines are sufficiently met usually needs an expert's opinion or dedicated testing through user studies. Consequently, there is considerable effort and knowledge involved in analysing and optimising warning messages. Small development efforts, such as start ups or app developers, often do not have the resources to thoroughly analyse the warning messages used in their products. They could benefit from more concrete and possibly objectively testable instructions on how to create useful warning messages in particular.

This paper investigates several methods to automatically assess warning message texts and analyses to which extent linguistic properties in general influence the user's perceptions of a warning message text. We will present an evaluation of existing readability measures on current browser security warnings as well as four empirical studies to assess the user's perceptions. Our results indicate that existing warning messages are too hard to read for the average user and that particular sentence structures as well as technical terms, which can be found in indexes of computer security textbooks, significantly correlate with the perceived difficulty of warning messages.

To the best of our knowledge, there has not been any work that empirically investigates the role of text comprehension and readability for computer warning messages to date. This work expands on preliminary results that have provided an overview of warning message readability using existing warning measures [8].

We offer three main contributions:

- We validate whether or not existing readability measures are suitable to judge warning message texts and determine the linguistic difficulty of existing warning messages.
- We investigate the effect of linguistic properties of warning message texts on the users' perceptions and provide empirical evidence for the influence of grammatical structures and vocabulary on warning message comprehension.

- We present quantifiable properties of text that influence warning message readability and comprehension.

The remainder of this paper is structured as follows: First, we introduce related work, before summarising readability of browser messages using a set of existing readability measures and an analysis thereof. Section 4 reports on the results of a user study that assess the applicability as well as the results of these readability measures. Sections 5 and 6 describe two online studies, collecting users' ratings of warning messages and comparing them to several linguistic properties. Additionally, effects of translation and a comparison between different software products are presented. Section 7 presents the results of interviews that discussed particular problems on a word and sentence level with users. Section 8 discusses limitations before Section 9 finally summarises the implications of our results and concludes this paper.

2 Related Work

A considerable amount of research has investigated warnings in the digital realm. Cranor's Human-In-The-Loop (HITL) framework [4] is a specialisation of Wogalter's C-HIP model [15] and describes how interactions between computers and humans can cause security problems.

Egelman et al. [6] presented a first study on warning efficacy for phishing prevention in 2008. They found that a large part of their test subjects chose to heed warnings that required interaction from the user and offer guidelines to improve warnings. According to their results, effective warnings need to interrupt the primary task, provide clear choices, fail safely and prevent habituation.

In a similar fashion, Sunshine et al. [13] tested the efficacy of certificate warnings presented by browsers and tried to improve the state of the art by modifying colours based on context and providing more detailed and interactive information on risks. While their changes improved efficacy, they concluded that the warnings still leave users vulnerable to man-in-the-middle attacks. Maurer et al. [9] also showed that warnings based on user input data types can help to prevent phishing and decrease habituation by increasing the context of a warning.

Bravo-Lillo et al. [3] provided another perspective on improving warning messages. They found that design changes can improve understanding and motivation but also realised that warning messages were not able to help users to differentiate between low and high-risk situations. Understanding and motivation were also found to be strongly connected and important factors in safely responding to warnings. Additionally, Bravo-Lillo et al. [2] offer qualitative insight into warning assessment by users of different skill levels and conclude that all aspects of warning design need to be considered in order to improve warnings. They also explicitly mention that the process of reading a warning is a central concern for warning message reception.

In another line of work, previous research has empirically investigated readability issues of end-user license agreements [7] and found shortcomings in informing the user before demanding consent.

The related work has conferred many valuable insights into the effectiveness and design of warning messages as well as problems with readability. We hope that the analyses presented in this paper complement the existing results by investigating the role of linguistic properties for the comprehension of warning message texts.

3 Readability Measures

In a previous publication [8], we explored the application of readability measures from the domain of educational psychology for computer warning messages. These measures take a piece of text and predict a level of reading skill necessary for comprehending the contents. For example, obtaining a value of 11 from a readability measure, such as SMOG [10], for a piece of text implies that an average reader needs to have the reading level of a student in 11th grade to be able to process the linguistic structure of this text. It is important to note that readability measures do not address the semantic difficulty of a text, but focus on linguistic difficulty, which is related to complicated sentence construction, long or polysyllabic words and similar properties. However, a text can be deemed to be "readable" using a certain measure but still confuse a reader. Yet, the linguistic difficulty is an important precursor for the overall comprehension of a text and therefore a useful indicator. If readability, as obtained from a suitable readability measure, is bad, the semantic information is harder to extract. *In the remainder of this paper, we generally address linguistic difficulty as described above, as opposed to semantic difficulty or other aspects of text layout, such as typesetting.*

Previously, we presented an analysis of security warnings based on warning messages from the two most common open-source browsers, Google Chrome and Mozilla Firefox. We extracted 24 English warning texts (15 for Chrome, 9 for Firefox) and added another four certificate warnings (hostname verification or unknown root CA warnings) from Internet Explorer 8, Safari, Outlook and iTunes to our sample to offer a broader cross-product comparison for a particularly common warning message. Warnings include certificate and phishing warnings, as well as messages indicating connectivity problems or unreachable servers. We also collected the same warnings in German. The selected warnings have at least about 50 words, because the readability measures we used are not validated for shorter samples of text. An abbreviated list of the warnings can be found in Appendix A.

We found that the predicted reading skills for this set of warnings differ depending on which measure is applied. However, all measures suggested at least an average reading level of an eighth grade student, while the SMOG measure, which is most suitable for warning messages due to its construction, even predicted the reading level of a first year college student for the average warning message. Details can be found in [8]. The extent to which these values are appropriate and useful is discussed in the following section.

4 Exploratory Study

To validate the readability results described above, we conducted an exploratory study of readability and linguistic comprehension. In order to minimise the effects of differences in language skills, we decided to test only native speakers. Since the study was conducted in Germany, we used the German versions of the set of 28 warning messages introduced above.

4.1 Design

Participants took a standard reading ability test to judge their individual reading level (Metze's "Stolperwoerter" test [1]). Next, they were presented with a cloze test (a piece of text where every fifth word is removed and has to be filled in by the participant) on six selected warning messages and scored based on their success rate. Cloze tests are commonly used as comprehension tests for the construction of the existing readability measures [5]. We selected four German warnings from Chrome and two from Firefox, since their readability scores (Amstad's measure for German texts) were distributed across the range we found in the tests described above. We stripped the warnings of all identifying and distracting features, using the same font and background for all messages. We introduced a fictitious browser named *InterBrowse*, as well as a fictitious banking website *mybank.com*, and replaced all references to the original software and websites with these names. Participants were given a simple working scenario stating that they were trying to surf to www.mybank.com using InterBrowse and then encountered a warning. We also reminded them that we intended to test the messages and not the participants' performance. After completing the cloze tests, participants re-read the full messages and sorted the texts by their feeling of comprehension. We pre-tested our protocol in a laboratory setting, discussed in previous work [8].

4.2 Participants

Based on this study protocol, we invited 1,486 students on a university-wide mailing list to participate in an online study. We advertised a study on browsing behaviour that would take 20 to 25 minutes and offered participants the chance to win a lottery of two 100 € Amazon vouchers as motivation. We received 311 complete responses, after removing non-native speakers and respondents with IT-related majors. The participants' average age was 22.8 and 130 came from the faculty of arts (cf. Table 2 in the Appendix). Technical experience among our participants was rather high, with an average of 2.29 on a scale from 1 (high expertise) to 5. Upon completion of the tasks, 216 participants (69.5 %) reported that they had seen one of the six warnings before and 49 (15.8 %) were unsure.

4.3 Results

For each participant, we collected the Stolper score, i. e. the individual's reading level, the cloze performance, i. e. how many of the gaps in the text were filled in correctly, the time taken per cloze text, and each participant's ranking in terms of subjective readability of the six presented warning messages, i. e. which messages did the participant find harder or easier to read and understand. Cloze performance was automatically assessed using a Levenshtein distance of 3 on the provided answers. Therefore, a word in a gap was counted as correct if the edit distance was equal or less than 3 compared to the original word, accounting for typos. This approach was chosen over an individual assessment of the semantics of the provided solution, since manually assessing each solution would have been too time consuming and could have biased results due to subjective scoring. To compensate for this strict assessment of performance, we chose a lower criterion score (see below).

We found significant differences in the cloze test performances between participants with high or low technical expertise. Since the cloze performances were found to be non-normally distributed (Kolmogorov-Smirnov Z between 1.579 and 2.862, $p < .031$ in all cases), we applied the Mann-Whitney U test and found significant differences in all messages (U between $5,762$ and $6,344$, Z between -2.301 and -3.144 with $p < 0.05$) except one (Message 6, $U = 7,595.5$, $Z = -.493$ and $p = .622$). While all other messages received higher scores from high-expertise participants, this particular message took the longest time to complete on average and received similar scores from both groups. The seldom seen message was about the use of a weak signature algorithm in a certificate and might therefore have been perceived as equally complicated by high- and low-expertise participants. Interestingly, this message also received the best average performance across all warnings, which suggests that complicated messages can be understood if enough time is spent.

In our reading ability test (Stolper-Test), the 311 respondents achieved an average score of 77.85 % (sd = 17.95), which is above average for their age group. The average score for participants between 21 and 25 years is 70.7 % and for people of 26 years and older is even lower (66 %), according to [1]. This effect can be explained by the above-average education of students.

Readability Results: Using the participants' reading abilities, we calculated readability scores for each of the six tested warnings to compare with existing measures. This procedure was adopted from the original construction of other readability measures which use cloze tests on passages of selected texts to derive the readability formula through regression [5]. The scores are based on a criterion score or threshold of correct answers on the corresponding cloze test. A criterion score of 90 % or higher is necessary for important information that needs to be well understood by readers [5, 8]. However, since cloze performance was automatically assessed, we chose a criterion score of 70 % to account for synonyms. Using this criterion, we calculated readability scores for the six warnings as the average reading level (Stolper score) of participants that performed better than the criterion on a particular warning message. Therefore, lower values for the readability score indicate higher readability.

According to the results (cf. Table 3 in the Appendix), our score correlates highly with the number of words in a message ($\rho = .943$, $p = .005$). While there are no other significant correlations due to the small sample size, we found indications of potential correlations with Amstad ($\rho = .714$, $p = .111$) and LIX ($\rho = -.600$, $p = .208$) scores. However, the implied direction of correlation is conflictive: These numbers suggest that better readability according to our Stolper-score-based measure is connected with worse readability according to Amstad and LIX. We could not find a significant correlation with the participants' rankings of messages either.

Because of the small number of warnings in this exploration, we cannot generally reject the applicability of readability measures for warning messages. However, the results suggest that the existing measures for German texts (i. e. the Amstad and LIX scores) do not fit the scores we collected directly from participants.

Another important trend is that for those students achieving 70 % or more correct answers in cloze testing, the mean reading ability is considerably higher ($> 79\,\%$) than the average score in their age group and older age groups ($66 - 70\,\%$). This implies that the average person would find these warnings hard to read.

The results also suggest that the readability scores we derived from Stolper scores somewhat mirror the participants' perceptions: scores are higher for messages rated as having the best subjective readability and lower scores for those perceived as worst. Another interesting implication of our results is that we did not find any correlation at all between the existing readability measures for German texts and the participants' subjective ratings of warning comprehension. The next section investigates this further.

5 Rating Study

The study described in the previous section focused on gaining direct measurements of text readability to evaluate the applicability of readability measures. The results suggest that the readability scores obtained from existing measures may not mirror the participants' perceptions of warning messages.

With the study presented in this section, we aimed to gather how easy people perceive understanding a warning message text to be. If a text is easier to read, the problem of users not reading or skimming warning messages might be alleviated. Therefore, we collected user ratings for the 28 warning messages introduced in Section 3. Again, we used the German versions of the texts and tested native speakers, to minimise effects of language skill levels.

5.1 Design

We prepared an online survey that presented each participant with six out of our set of 28 warning messages. Participants were primed with the same scenario as in the previous study. The order and selection of the messages was randomised for each participant. For each warning message, participants were asked to read the message, to summarise the contents roughly in one sentence and then rate their perception of the warning message with four items on a 7-point scale from "I completely agree" to "I completely disagree". The items addressed comprehension of the entire message, the words used in the message, previous exposure and understanding of why the message appears. We also added two additional items, which were semantically inverse to two items in the original set. Before starting the rating exercise, we asked participants an attention question, that required participants to answer "No" even though the correct answer was obviously "Yes". At the end of the survey, we collected demographics.

5.2 Participants

We invited 1,522 students of the same mailing list[1] to participate in the survey. The study was advertised as a follow-up of the previous study that would take 8 to 12 minutes to complete, welcoming new and returning participants. Once again, we offered participation in a lottery for two 50 € Amazon vouchers as compensation. 250 participants successfully completed the survey. First, we removed participants that wrongly answered the attention question with "Yes" instead of the required "No". We also removed records of participants that study IT or a related subject, whose native language was not German and whose browser language was not German, to remove effects stemming from the level of language skill as well as daily exposure to warnings in different languages. Furthermore, responses that had a mean difference of three or more between the two inverse items and the corresponding original items were removed. Lastly, we filtered respondents that always chose the same answers on the rating items and those who either entered nonsensical summaries or copy-and-pasted parts of the warning message.

After filtering, 119 complete and validated responses remained. 40.3 % of our participants were female, 51.3 % had participated in the study described above and 60.5 % reported to have seen one of the warnings they were shown before (cf. Table 4 in the appendix). On average, it took the participants about 16 minutes to complete the survey, which is considerably longer than anticipated by pretesting in a laboratory setting.

5.3 Results

Initially, we checked for demographical imbalances in our rating results, using the non-parametric Mann-Whitney U test, since normality testing indicated significant deviations from the normal distribution in many of the rating variables. We found a few imbalances on the item for message comprehension: Messages 5, 21 and 27 were rated significantly better by participants that had previously participated in the first study. Message 12 received better ratings from men and

[1] The number of subscribers increased between studies.

messages 18 and 22 received significantly different ratings by participants that stated they had seen some of the warnings before. Since there was no obvious pattern in these differences, we accept them for further analysis.

We used Spearman's rho as a robust measure to test the monotonic relationship between rating ranks. The average ratings for comprehension and understanding the cause are strongly correlated ($\rho = .937$, $p < .001$) as is comprehension and difficulty of vocabulary ($\rho = -.797$, $p < .001$). Additionally, there is a relationship between previous exposure and the three other items ($\rho = -.65$, $\rho = .76$ and $.80$, $p < .001$): having more experience with a warning may support comprehension and understanding the cause.

Linguistic Properties: To see if particular linguistic properties of a warning message influence the users' perceptions, we used the Stanford Parser [11] and Part-of-Speech (POS) tagger [14] for German texts to analyse the structure of the warning texts. We gathered frequencies for 54 types of tags from the "Stuttgart-Tübingen-Tagset", as well as parse-tree parameters, including average number of nominal and verb phrases per sentence, as well as maximum and average parse-tree depth.

Several POS tag types showed medium to strong correlations with the ratings: Articles (ART, $\rho = .593$, $p = .001$) and the participle perfect (VVPP, prefix or infix "ge", $\rho = .564$, $p = .002$) appear to positively correlate with ratings, while the occurrence of the particle "zu" (english: "to") in front of an infinitive (PTKZU, $\rho = -.63$, $p < .001$) showed a negative correlation. Linear regression showed that VVPP and PTKZU can explain 54.7 % of the total variance in the participants' comprehension rating. Additionally, we did not find any meaningful correlation with the existing readability measures Amstad and LIX.

We also found correlations between the readability score we calculated based on cloze testing in the previous study with the maximum parse-tree depth ($\rho = -.872$, $p = .054$) and the number of attributive adjectives per sentence (ADJA, $\rho = -.90$, $p = .037$), but not with the ratings collected in this study. However, these correlations lack power, since the previous study only investigated six warning messages.

6 English Rating Study

In order to explore if similar effects exist for English warnings, we ran an additional rating study with the same setup on Amazon's Mechanical Turk (MTurk). Furthermore, warning messages for international software projects, such as Firefox and Google Chrome, are usually written in English and then translated into the different languages for localisation. It is possible that translation may cause the resulting warning messages to have a different linguistic structure compared to one written directly in the target language. Thus, we also used this study to compare the results of the translated warning texts with their original counterparts to see if translation has any effects on the ratings.

6.1 Design

We used the English versions of the set of 28 warning messages and created a HIT that advertised a task to rate ten browser warning messages on MTurk. We offered 1.50 $ as compensation for each successful completion and stated that only non-random and honest answers would receive the compensation. The study included the same validation questions as before and presented ten randomly selected warnings to each participant after introducing the InterBrowse and mybank.com scenario.

6.2 Participants

Our HIT was completed by 120 workers and took an average time of 20 minutes and 13 seconds (sd = 12 minutes and 29 seconds). We applied the same filtering methods as described in the previous study and hence retained 68 valid responses. Each message received an average of 24.3 ratings, ranging from 15 to 32. The average age of participants was 37 years ($sd = 12.7$), exactly half were female, and the overall self-reported technical experience was 2.44 ($sd = 1.01$). Respondents stated their occupation as student (8.8 %), full-time employee (14.7 %), part-time employee (47.1 %), self-employed (20.6 %) and other (8.8 %), including unemployed and homemakers.

6.3 Results

Similar to the results above, many of the rating variables showed significant deviations from a normal distribution (Kolmogorov-Smirnoff Test). We therefore ran the remaining analysis using non-parametric tests. First of all, the data was checked for demographical imbalances. For the comprehension rating, we found that messages 25 and 26 were perceived to be more difficult by younger participants. Interestingly, as in the results for the German versions of the messages, message 12 was perceived as being significantly easier to comprehend by men (Mann-Whitney $U = 48$, $Z = -2.297$, $p = .026$). Similar to above, the different ratings show significant correlations, although the strength is slightly weaker.

To identify structural features that influence ratings in English messages, we again applied the Stanford Parser and POS tagger for English texts to the English warnings. We used the 36 POS tags of the Penn Treebank Tagset[2], as well as the number of nominal and verb phrases, number of words per sentence, maximum number of words in a sentence, and (maximum) parse-tree depth. In contrast to before, we found only two correlations: the number of determiners (DT, similar to articles, $\rho = -.60$, $p < .001$) negatively influenced the ratings on difficulties with the vocabulary and the comprehension rating ($\rho = .491$, $p = .008$). In this case, linear regression was able to explain 46.2 % of the variance in the comprehension rating, using the number of words in the longest sentence as well as the number of wh-determiners (WDT, e. g. "which") and co-ordinating conjunctions (CC, e. g. "and"). There also was no meaningful correlation with the existing readability measures for English texts.

[2] `http://www.cis.upenn.edu/~treebank/home.html`

Comparison With German Results: We found a medium to strong correlation between the ranks for the German messages from the previous study to the English pendants (ρ between .68 and .78 for the four rating items, $p < .001$), indicating that messages perceived as complicated in German were also perceived as such in English and vice versa. Therefore, we conclude that the effects observed in the German messages do not purely stem from translation.

Next, we ranked all messages according to the three rating categories comprehension, understanding the cause and difficulty of vocabulary in the respective language. Based on the top and bottom five messages in each category, we found that three messages performed very well and four messages performed very poorly in both languages. Messages 18 and 19, (Firefox: "Reported Attack Page" and "Suspected Web Forgery"), and 28 (Safari, "Invalid Certificate") were consistently among the highest ratings. These warnings use easy, non-technical vocabulary and give direct recommendations on possible actions for the user.

The four messages receiving consistently bad ratings comprise three messages from Chrome ("Weak Signature Algorithm", "Unlisted Server Certificate", and "No Revocation Mechanism"), as well as one from Firefox ("SSL Disabled"). These messages address very technical issues and have probably never been seen by any of our participants: they also received very low previous exposure ratings.

Comparison Between Products: Between the six certificate warning messages of different products that we included in the set of warnings, results showed that the Safari message was consistently found to be the easiest to comprehend and to use the easiest words. Likewise, we found that the message from Internet Explorer 8 was consistently rated worst. While the messages have comparable length (42 and 59 words respectively), the Internet Explorer message repeatedly uses the word "certificate" and other technical terms. The Safari message, in contrast, uses simple language, states a cause, the involved risk and asks the user to decide on a course of action.

Two Chrome warnings in our set differed only by their headline. One read: "This is probably not the site you are looking for!" and the other said "The site's security certificate is not trusted!". The message that did not mention certificate in the headline received consistently better ratings in both languages. Even though the difference is not statistically significant, this trend may imply that technical terms at the very beginning of a warning message can negatively influence the users' perceptions. To further investigate which factors influence users' perceptions of a warning message text in particular, we conducted interviews.

7 Interview

The previous studies have shown that there can be particular linguistic properties that may influence a user's perception of a warning message. The use and placement of technical terms as well as specific grammatical constructs showed correlations with the user ratings. We conducted interviews to directly analyse the participants' perceptions of technical terms and linguistic features, such as sentence composition.

7.1 Design

The interview was introduced to the participants as an investigation of readability in Internet browser warning message texts. We reminded them that this test was not about their abilities to comprehend the warnings but that their insights as to why a certain message might be hard to understand was of interest. Participants were presented with six warning messages as well as our InterBrowse scenario and would then be asked to carefully read the message. Next, we queried which sentences or parts of sentences were hard to read and their explanation. Afterwards, participants ranked all 6 warnings according to the perceived level of complexity. In a last task, they were provided with three highlighters and the same set of warning messages once more: we asked them to use a green highlighter to mark easy and clear words, a yellow one for words of medium difficulty that they still knew the meaning of and a red one for unclear and hard words. While they were working, we asked participants to offer their reasoning and collected their comments.

7.2 Participants

The participants were randomly recruited by phone from the database of more than 1,500 students also used above. Non-native speakers, students of German and Literature or Computer Science were excluded. We offered a compensation of 10 € and interviewed eight students (three female, 19 to 24 years old, four from the faculty of arts and four from the sciences) before our results reached saturation. Two participants had taken part in one of our previous studies, seven stated that they had seen one of the warning messages before or were unsure, four mainly use Firefox while two use Safari, one Chrome and one IE. The mean self-reported technical experience was 2.87 ($sd = .64$).

7.3 Results

Participants' comments can be divided into three main categories, detailed below. Participants are referred to as $P1, \ldots, P8$.

Headlines: Seven respondents stated that a warning's title should be short and precise. Additionally, five claimed that technical terms should not be in a headline. Four participants offered that "if I only looked at the heading, I wouldn't have had any clue what the error message is about" (P7). Participants agreed that an ill-conceived headline would deter them from continuing to read.

Positive properties of sentences: Short, precise sentences with an easy structure were appreciated by all respondents. Four of them explicitly requested that a simple sentence structure should be used: "[This] makes the message more colloquial and perfect for people who aren't experts" (P8). All participants offered that technical terms used in error messages hamper the understanding and awareness of the potential problem. The text marking tasks also showed that short sentences are preferred, yet, according to the comments, longer and more complex structures do not necessarily lead to readability problems.

Negative properties: All participants agreed that the use of technical terms (see below) discourages them from reading (on) and trying to understand the scenario. P2 added: "One has to be really desperate to read this passage thoroughly". In a similar fashion, half of the participants stated that in daily life, they would simply ignore paragraphs with many technical details. Six participants attempted to decode the meaning and the possible impact of the information in some of the warning messages, but failed. They felt "insufficiently informed" (P6) by the messages. P1 stated: "You simply want to get to the desired website and I don't understand the problem itself nor when or how it will get solved". These findings generally confirm the general preconceptions and the results of previous work.

Word-level Observations: During the word marking exercise, participants often indicated words as hard that had a technical background or referred to unclear concepts. The list included words such as "certificate" or "entity", but also simple adjectives, including "attacking" and "weak". Table 5 in the Appendix provides an overview of all words mentioned by participants.

Using this list of words, we counted occurrences in our set of 28 German warning messages. Again using Spearman's Rho, the counts of hard words showed a correlation of .559 ($p = .002$) with the ratings of comprehension obtained in the studies described above, even though the list of words was only obtained on 6 of the 28 messages. Expanding on the implications of these results, we used the index terms of a computer security textbook [12][3] as an extended word list. The count of words from this list found in the 28 warning messages provided a slightly stronger correlation with ratings ($\rho = .646$, $p < .001$). The three best-rated and four worst-rated messages identified in section 6.3 also consistently received corresponding index-word counts of one match or less and three matches or more respectively. The same holds for headlines: the best-rated messages only used "website" in their headings while the worst-rated messages used technical terms (e. g. "certificate" or "revocation").

8 Limitations

There are several limitations which need to be taken into account: First, our participants were either students or Mechanical Turk workers, which both represent a special group of people. Especially the students may present a best-case scenario for text comprehension, due to the exposure to difficult reading assignments in many subjects. However, the groups are quite different in terms of age and education, as well as professional background. Yet, we still found similar results in both studies.

Second, collecting self-reported measures likely causes a certain amount of bias. However, we implemented measures to try and mitigate these effects, by randomising messages and their order, as well as using only relative comparisons.

[3] We chose this textbook because it was the most recent security textbook digitally available at our library with an index.

Finally, we did not address the efficacy of warning messages explicitly, but used user ratings. While "pleasant" readability is a goal within itself, the correlation between readability and efficacy needs to be explicitly studied in future work. As noted above, related work suggests that facilitating the understanding of warning messages can predict user behaviour [13].

9 Discussion

During the course of our investigations, we found several aspects of warning message texts that influence their reception by users. First, cloze testing indicated that the required average reading level for warning messages is higher than the average reading level of most adults, mirroring the common image of warning messages often being too complicated. Results also hinted at the possibility that complicated messages can be understood by many readers if they spend enough time. However, these tests also indicated that the set of existing readability measures does not predict warning message difficulty accurately.

We then conducted the rating studies to collect users' ratings of warning messages and analyse if there are linguistic properties that can explain the rating differences. In both English and German warning texts, linguistic properties were able to explain about half of the variance in the ratings. Grammatical constructs that increase the information content of a sentence, for example coordinating conjunctions in English texts and German infinitive constructions, as well as grammatical tenses, such as the participle perfect in German texts, cause texts to be perceived as harder to understand. Additionally, we found that in both German and English versions of the warnings, messages with easy and non-technical vocabulary consistently received positive ratings while those that addressed specific technical problems consistently received negative ratings. A comparison between warnings from different products showed similar results.

Finally, we interviewed users and gathered aspects of warning message texts that may influence comprehension: headlines, non-technical vocabulary and short sentences were among the most frequently stated issues influencing the users' perceptions of warning message texts. Interestingly, the stated need for precise statements can cause conflicts: technical vocabulary is commonly used to make statements precise and short.

We were able to show that the linguistic properties identified in our studies can also be found in the best and worst message texts, according to the collected ratings. The set of words extracted from our interviews as well as a computer security textbook's index showed significant correlations with the ratings.

As stated above, our findings were able to explain about half of the variance in ratings using linguistic properties. We thus conclude that the linguistic properties of warning message texts and consequently issues that users might have with complicated sentence structures or difficult compounded words are one part of the larger puzzle, which entirely needs to be taken into account when designing new warning messages. Additional factors, such as missing context, previous exposure, unclear semantics, and effects of attitudes and beliefs can also strongly influence the users' perceptions of warning messages and their text.

Altogether, we found quantitative empirical evidence that linguistic properties can help to improve warnings: keeping headlines simple, using as few technical words as possible and creating short sentences without complicated grammatical constructions makes warning messages more pleasant for the user. A final take-away is that warning messages should not contain words that can be found in IT security textbook's indexes. It is of course a challenge to describe the warning without such terms, however our results suggest it is a challenge worth working on.

References

1. A. Backhaus, H. Brügelmann, S. Knorre, and W. Metze. Forschungsmanual zum Stolperwörter-Lesetest. `http://www.agprim.uni-siegen.de/lust/stolpermanual.pdf`, 2004.
2. C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security & Privacy Magazine*, 9(2):18–26, 2011.
3. C. Bravo-Lillo, L. F. Cranor, J. Downs, S. Komanduri, and M. Sleeper. Improving Computer Security Dialogs. In *Proc. INTERACT*, pages 18–35, 2011.
4. L. F. Cranor. A Framework for Reasoning About the Human in the Loop. In *Proc. 1st Conf. Usability, Psychology, and Security (UPSEC 08)*, 2008.
5. W. H. DuBay. The Principles of Readability. `http://www.impact-information.com/impactinfo/readability02.pdf`.
6. S. Egelman, L. F. Cranor, and J. Hong. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of Human Factors in Computing Systems*, pages 1065–1074, 2008.
7. J. Grossklags and N. Good. Empirical Studies on Software Notices to Inform Policy Makers and Usability Designers. In *Proc. USEC*. Springer, 2007.
8. M. Harbach, S. Fahl, T. Muders, and M. Smith. Poster: Towards Measuring Warning Readability. In *Proc. ACM CCS*, 2012.
9. M.-E. Maurer, A. De Luca, and H. Hussmann. Data Type Based Security Alert Dialogs. In *Extended Abstracts on Human Factors in Computing Systems*, 2011.
10. G. H. McLaughlin. SMOG Grading – a New Readability Formula. *Journal of Reading*, 12(8):639–646, 1969.
11. A. Rafferty and C. D. Manning. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *ACL Workshop on Parsing German*, 2008.
12. S. Spitz, M. Pramateftakis, and J. Swoboda. *Kryptographie und IT-Sicherheit.* Springer, 2011.
13. J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *USENIX 2009*, pages 399–416, Aug. 2009.
14. K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL*, pages 252–259, 2003.
15. M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson. Research-based Guidelines for Warning Design and Evaluation. *Applied Ergonomics*, 33(3):219–230, 2002.

A Warning Messages

Due to space constraints, the full set of warnings used in the study cannot be shown in the paper. In the following, we present an overview of messages while the full set can be downloaded from `http://benutzerstudie.dcsec.uni-hannover.de/warnings/`.

Table 1: Overview of warning messages used in the study.

#	Browser	Beginning of Message
1	Chrome	The site's security certificate has expired! For a certificate which has not expired, the issuer of that certificate is ...
2	Chrome	The site's security certificate is not trusted! You attempted to reach mybank.com, but the server presented ...
3	Firefox	This Connection is Untrusted. You have asked InterBrowse to connect securely to mybank.com, but ...
4	Chrome	The site's security certificate is not trusted! You attempted to reach mybank.com but instead you actually reached ...
5	Chrome	Invalid Server Certificate. You attempted to reach mybank.com, but the server presented an invalid certificate. ...
6	Chrome	The server's security certificate is not yet valid! You attempted to reach mybank.com, but the server presented ...
7	Chrome	This is probably not the site you are looking for! You attempted to reach mybank.com but instead you actually ...
8	Chrome	The site's security certificate is signed using a weak signature algorithm! You attempted to reach mybank.com, but ...
9	Chrome	The server certificate contains a weak cryptographic key! You attempted to reach mybank.com, but the server presented ...
10	Chrome	The server's security certificate is revoked! You attempted to reach mybank.com, but the certificate that the server ...
11	Chrome	Unlisted Server Certificate. This site lists all its valid certificates in DNS. However the server used one which isn't listed. ...
12	Chrome	The server's security certificate has errors! When you connect to a secure website, the server hosting that site presents ...
13	Firefox	This Connection is Untrusted. You have asked InterBrowse to connect securely to mybank.com, but we can't confirm ...
14	Chrome	This webpage is not available. InterBrowse's connection attempt to mybank.com was rejected. The website may be down, or ...
15	Chrome	No revocation mechanism found. No revocation mechanism found in the server's certificate. When you connect to ...
16	Chrome	Unable to check whether the server's certificate was revoked. When you connect to a secure website, the server hosting ...
17	Chrome	Unknown server certificate error. An unknown error has occurred. When you connect to a secure website, the server ...
18	Firefox	Suspected Web Forgery. This page has been reported as a web forgery designed to trick users into sharing personal or ...

19 Firefox Reported Attack Page! This web page at mybank.com has been reported as an attack page and has been blocked based on ...

20 Firefox The certificate is not trusted because it is self signed. mybank.com uses an invalid security certificate. ...

21 Firefox Certificate will not be valid until date. mybank.com uses an invalid security certificate. The certificate will not be valid ...

22 Firefox The certificate expired on date. mybank.com uses an invalid security certificate. The certificate expired on ...

23 Firefox SSL protocol has been disabled. An error occurred during a connection to mybank.com. Can't connect securely because ...

24 Firefox Untrusted Connection Error. You have asked InterBrowse to connect securely to mybank.com, but we can't confirm that ...

25 MS IE 8 Security Certificate Problem. There is a problem with this website's security certificate. The security certificate ...

26 iTunes InterBrowse cant verify the identity of the server mybank.com. The certificate for this server was signed by ...

27 MS Outlook Problem with the site's security certificate. The information you exchange with this site cannot be viewed or changed ...

28 Safari InterBrowse can't verify the identity of the website mybank.com. The certificate for this website is invalid. You might ...

B Tables

Table 2. Demographics for the exploratory online study. Self-reported technical expertise was measured on a scale of agreement to the statement "I have a very detailed understanding of computer technology and the Internet" with 1 being complete agreement and 5 complete disagreement. The Stolper score indicates reading ability on a scale from 0-100% of successful completion of 35 reading tasks in five minutes.

N:	311
Age:	$22.8, sd = 4.1$
Tech. Expertise:	$2.29, sd = .92$
Area of Studies:	130 Arts (41.8%)
	181 Sciences and Other (58.2%)
Browser:	195 Firefox (62.7 %)
	56 Chrome (18.0 %)
	14 Internet Explorer (4.5 %)
	17 Opera (5.5 %)
	28 Safari (9.0 %)
	1 Other (.3 %)

Table 3. Results of cloze testing. Higher Amstad and lower LIX scores suggest better readability. The average rank indicates the position within the participants' subjective ordering of warnings (ranks closer to 1 indicate better subjective readability). Lower values of our readability score (70 % criterion score) indicate better readability. The last column shows the number of participants that were above the 70 % criterion score.

Message	Words	Amstad	LIX	Avg. Rank	Score 70	# Respondents
1	61	62.84	39.67	3.05	80.16	110
2	45	43.48	59.44	2.78	79.17	69
3	85	54.99	54.19	3.87	80.73	43
4	114	68.02	38.59	3.25	81.14	70
5	99	71.44	45.79	3.49	80.25	81
6	59	49.64	48.65	4.55	79.54	112

Table 4. Demographics for the Rating Study. Self-reported technical expertise was measured on a scale of agreement to the statement "I have a very detailed understanding of computer technology and the Internet" with 1 being complete agreement and 5 being complete disagreement.

N:	119
Age:	$22.7, sd = 4.02$
Tech. Expertise:	$2.34, sd = .98$
Area of Studies:	51 Arts (42.9 %)
	68 Sciences (57.0 %)
Browser:	82 Firefox (70.1 %)
	16 Chrome (13.7 %)
	8 Internet Explorer (6.7 %)
	3 Opera (2.5 %)
	8 Safari (6.7 %)
	2 N/A (1.7 %)

Table 5. Words mentioned by interview participants, arranged by difficulty and number of participants they were mentioned by. The category "high-one" was omitted because it was empty.

Medium Difficulty		**High Difficulty**
one	more than one	more than one
to confirm	weak	signature algorithm
to issue	attacking	security certificate
to forge	security settings	certificate
expiry	to expire	entity
to adapt	server	network administrator
to check	to present (a certificate)	proxy server
to contact	manipulation	proxy settings
operating system	security credentials	
to block	identity information	
private information	identification	
communicate	secure connection	