POSTER: Towards Measuring Warning Readability

Marian Harbach, Sascha Fahl, Thomas Muders, Matthew Smith
Distributed Systems and Security Group, Dept. of Computer Science
Leibniz Universität Hannover, Germany
{harbach,fahl,muders,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. For a long time, researchers have investigated the effects of warning messages to optimise their reception by a user. Design guidelines and best practises help the developer or interaction designer to adequately channel urgent information. In this poster, we investigate the application of readability measures to assess the difficulty of the descriptive text in warning messages. Adapting such a measure to fit the needs of warning message design allows objective feedback on the quality of a warning's descriptive text. An automated process will be able to assist software developers and designers in creating more readable and hence more understandable security warning messages. We present an initial exploration of the use of readability measures on the descriptive text of warning messages. Existing measures were evaluated on warning messages extracted from current browsers using an experimental study with 15 undergrad students. While our data did not yield conclusive results yet, we argue that readability measures can provide valuable assistance when implementing security systems.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation (e.g. HCI)**]: User Interfaces - Evaluation/Methods.

Keywords

Usable Security, Warning Messages, Readability, User Interface Design

1. INTRODUCTION

Designing and writing warning messages can be considered a form of art that is often supported by engineering guidelines. A sizeable amount of research has evaluated different strategies to create effective warnings in the physical as well as the digital world [9, 5, 7]. These evaluations often address the entire warning message and focus on overall construction to gain the user's attention. The reception of warning messages by a user is often explained using Wogalter's Communication-Human Information Processing (C-HIP) model [10].

CCS'12, October 16–18, 2012, Raleigh, North Carolina, USA.
ACM 978-1-4503-1651-4/12/10.

It has been recognised that the descriptive text provided in warning messages needs to be comprehensive and understandable by most computer users. 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."

Judging whether or not these goals are sufficiently met is however usually left to an expert's opinion or to testing through user studies. Consequently, there is considerable effort and knowledge involved in analysing and optimising warning messages.

For over 60 years, educational research has developed and studied automatic measures to analyse text readability and suitability. Formulas, such as the Flesch Reading Ease, the Gunning Fog Index or the New Dale-Chall Formula, are compiled from empirical analyses and allow a rough estimation of the number of years of education a reader has to have had in order to be able to comprehend a given text to a certain degree.

The ongoing work presented in this poster examines the possibility of using automatic readability measures to support the analysis and creation of end-user warning messages in computer software. We will present an initial analysis of browser security warnings using existing measures as well as a first explorative study with 15 participants to analyse the applicability of these measures. To the best of our knowledge, there has not been any work investigating the application of readability measures for computer warning messages to date.

2. READABILITY MEASURES

Readability measures have been investigated for decades. Examples include the Flesch-Kincaid Grade Scale, the Dale-Chall Formula, FORCAST, the Gunning-Fog Index and SMOG. Traditional readability measures are also called *surface* or *shallow* measures, because in contrast to *deep* measures, they only use properties such as average number of words per sentence, syllables per word or average word length to judge readability. They are generally based on statistical regression against a certain population and therefore have the advantage of being easily computable.

Deep measures, such as the CohMetrix [6] or the DeLite Readability Checker [8], use more elaborate analyses to judge the readability of a text. However, while the traditional surface measures were made to be computed manually by ed-

ucators in an age where few computers were available, the deep measures are computationally expensive and often need machine-learning-based training. It has also been shown repeatedly that the shallow measures have strong correlations to deep measures [2]. We hence focus our preliminary analysis on the traditional measures. A recent overview of work in the area of text readability can be found in [2].

For this work, we computed seven different readability measures for the warnings we analysed. While these measures use different text properties and training populations, all take a piece of text and compute a score that usually represents the number of years of education a reader has to have had in order to read and understand that piece of text. We applied the Flesch-Kincaid readability test (Flesch-Reading-Ease converted to grade scale), the Gunning-Fog Index, the New Dale-Chall Formula, FORCAST and SMOG as well as the Amstad Formula (an adaption of Flesch-Reading-Ease) and the DeLite Readability Checker for German texts.

3. COMPUTER SECURITY WARNINGS

We analysed security warnings of the two most common open-source browsers, Google Chrome and Mozilla Firefox. From the source code repositories, we were able to extract 26 English warning texts (16 for Chrome, 10 for Firefox) with more than 50 words, having an average length of 159.65 words ($sd = 19.2$, ranging from 51 to 360). These warnings include certificate and phishing warnings as well as messages indicating connectivity problems or unreachable servers. We only selected warnings with 50 or more words, because the measures do not perform reliably for short samples of text. Figure 1 provides a graphical overview and Table 1 gives details of the obtained readability scores for all tested measures. We also tested 14 German warnings, using the Amstad measure for German texts, which yielded similar results.

Table 1: Mean readability scores and statistics using different measures.

Measure	*mean*	*sd*
Flesch Reading Ease (FRE)[1]	60.37	11.95
Amstad[2]	54.33	8.79
Flesch-Kincaid	9.61	.55
FORCAST	17.40	.09
Gunning Fog	14.63	.53
New Dale-Chall	11.20	.12
SMOG	13.49	.36

Flesch-Kincaid, Fog and SMOG have significant and strong correlations ($r > .9$, $p < .001$). FORCAST has medium to strong negative correlations with those three ($r = -.508$ to $-.76$, $p < .01$) and New Dale-Chall has no correlation at all. These two measures probably behave differently due to their construction: FORCAST was developed for the U.S. army and is based only on the number of single-syllable words in a 150-word sample; The New Dale-Chall formula uses a set of 3,000 easy words and penalises the use of words not in that list.

[1]FRE and Amstad scores closer to 100 indicate better readability. The other measures score the number of years of education to be had by a potential reader. Flesch-Kincaid transforms the FRE to grade scale.

[2]Amstad is the adaption of FRE to German. This measure was applied to the 14 German warnings.

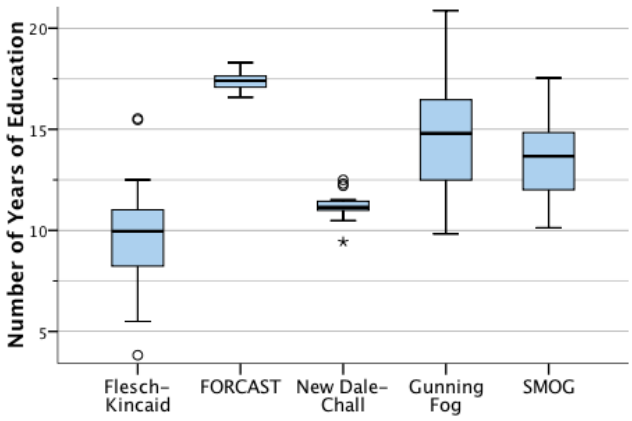

Figure 1: Boxplots for readability scores.

From the measures' construction, the SMOG measure is best suited to be applied to security warnings. It uses the average grade of readers that scored 100% of correct answers in a comprehension test, whereas Dale-Chall uses value of 50% as criterion score, Flesch-Kincaid uses 75% and Gunning Fog uses a 90% score. Readability literature suggests that "for unassisted reading, especially where [...] safety issues are involved", measures with high criterion scores may be more appropriate [4].

Overall, the data suggests that the reader of an average warning message needs to have at least 10 years of education to understand the messages, even 13 or 17 when applying SMOG or FORCAST. For the average warning message, the SMOG measure suggests an average grade level of 13.49, which is equal to college education. Whether or not these values are appropriate and useful is subject to ongoing work.

4. EXPLORATORY STUDY

To begin to evaluate the obtained results, we conducted an exploratory study. 15 undergrad students (average age 22.3, $sd = 2.19$, 5 female, 10 male, from different disciplines except languages and IT) took a standard reading ability test to judge their individual reading ability (Metze's "Stolperwoerter" test [1]). Next, they were presented with a cloze test on six selected warning messages and scored based on their error rate. Synonyms and words that did not alter the meaning of a sentence were counted as correct. We selected four German warnings from Chrome and two from Firefox. Their readability scores (Amstad's measure for German texts) were distributed across the range we found in the tests described above. After completing the cloze tests, participants were given the full messages and asked to rate their comprehension subjectively as well as to answer multiple choice questions concerning the warnings' contents. Finally, they chose which message they found to be the most and least readable.

In the analysis, we found no significant correlations between the existing measures, the multiple choice or cloze scores and self-reported comprehension. Messages B and D were selected as most readable while messages A and C were deemed least readable. Figure 2 summarises the results. All

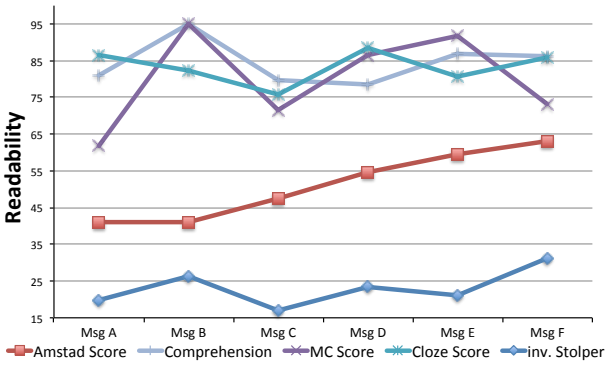

Figure 2: Results of the experimental study, ordered by Amstad readability score.

scores are normalised to the 0-100 interval with 100 indicating best readability according to the corresponding measure.

Due to the small sample size in this exploration, we cannot draw general results from the data. However, the preliminary results suggest that the existing measures for German text (i. e. the red Amstad scores in Fig. 2) do not fit the patterns we observe in the measures collected directly from participants. Another important trend is that for those students achieving 90% or more correct answers in cloze testing, the mean reading ability (Stolper score) is considerably higher than the average score in their age group. This indicates, that the average person might find these warnings hard to read. The results also suggest that the Stolper score mirrors the participants' perceptions: scores are higher for messages perceived as having the best subjective readability and lower scores for those perceived as worst.

5. CONCLUSION AND FUTURE WORK

Applying readability measures to warning messages has the potential to provide developers and designers with an automatic tool that can estimate how readable and understandable a warning will be for their target audience. This can help to improve the warning message design process, especially for those developers that cannot afford specialist help. However, further analysis is necessary to give useful and reliable predictions. It is important to note that a tool for warning message text analysis cannot relieve the creator of warnings of his responsibilities. Readability measures do not analyse whether or not a sentence is grammatically correct or makes sense. They cannot take context and other important aspects of the warning message design process into account. Therefore, readability measures should only be used as supportive tools during the design process.

Readability analysis has limitations that require further research: Traditional readability measures are usually defined through regression of reading comprehension scores of readers of a particular grade, using a small number of text properties. The measures therefore depend on their training population and the chosen properties. Populations could be varied to suit possible application audiences (e. g. browsers vs. scientific tools). We will also investigate the specific properties of security warning texts, their role in readability for users and their suitability as a basis for warning readability measures.

In our next steps, we are going to build on the preliminary results using a more comprehensive study with a larger sample size. Additionally, we would also like to conduct the study with English native-speakers to test the applicability of measures for English text. In a further step, we plan to extend the population to investigate warning readability for a more average computer user. We will conduct interviews to assess user perceptions when reading security warning messages. Additionally, traditional readability measures have problems to analyse short pieces of texts. During our exploration, we came across a large number of security warnings that consisted of less than 50 words. We would like to explore whether or not a useful measure can be found predicting readability of short warnings as well and whether or not short security warning can be useful at all.

6. 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] R. G. Benjamin. Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review*, 24:63–88, 2012.

[3] C. Bravo-Lillo, L. F. Cranor, J. Downs, S. Komanduri, and M. Sleeper. Improving Computer Security Dialogs. In *INTERACT 2011*, pages 18–35, 2011.

[4] W. H. DuBay. The Principles of Readability. `http://www.impact-information.com/impactinfo/readability02.pdf`.

[5] 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 CHI 2008*, pages 1065–1074, 2008.

[6] A. Graesser, D. S. McNamara, M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavioral Research Methods, Instruments, and Computers*, 36:193–202, 2004.

[7] 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.

[8] T. vor der Brück and S. Hartrumpf. A Readability Checker Based on Deep Semantic Indicators. In *Human Language Technology. Challenges of the Information Society*, Lecture Notes in Computer Science, pages 232–244. Springer, 2009.

[9] M. S. Wogalter, editor. *Handbook of Warnings*. Lawrence Erlbaum Associates, London, 2006.

[10] 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.