Poster: Towards Measuring Warning Readability

SOUPS '12 Poster Abstract

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

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 [7, 5, 6]. 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 the following rules for descriptive text: "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 compiled from empirical analyses, 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 of 15 students 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

The 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. While these properties can hardly capture all facets of a piece of text, it has been shown repeatedly that the shallow measures have strong correlations to deep measures [2]. Shallow measures also have the advantage of being easily computable. A recent overview of work in the general 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 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 of the obtained readability scores for all tested measures. We also tested German warnings, using the Amstad measure for German texts, which yielded similar results.

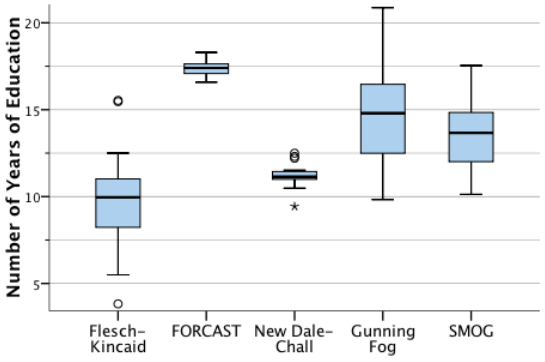

Figure 1: Boxplots for readability scores.

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.

From the measures' construction, the SMOG measure is best suited to be applied to security warnings. It is constructed using the average grade of readers that scored 100% of correct answers in a comprehension test, whereas Dale-Chall uses a 50% 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.

4. EXPLORATORY STUDY

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 skills, using Metze's "Stolperwoerter" test [1]. Next, they were presented with a cloze test on six selected warning messages. We selected 4 German warnings from Chrome and 2 from Firefox with readability scores distributed across the entire spectrum indicated by the above tests. Afterwards, they were given the full messages and asked to rate their comprehension as well as answer multiple choice questions concerning the warnings' contents. Finally, they chose which message they found to be the most and least readable.

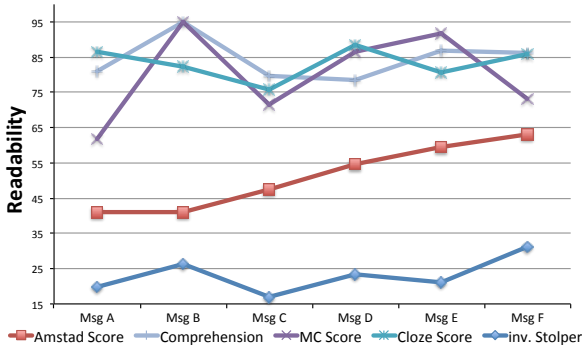

Figure 2: Results of the experimental study, ordered by Amstad readability score.

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 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.

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. This can help to improve the warning message design process. However, further analysis is necessary to make useful predictions.

For instance, readability analysis has limitations that require further research: First, the traditional 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. Second and most importantly, the readability measures do not analyse whether or not a sentence is grammatically correct or makes sense. Therefore, readability measures should only be used as supportive tools during the design process.

In our next steps, we are going to build on the preliminary results using a more comprehensive study with a larger sample size. Most importantly, we would 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. Lastly, 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.

6. REFERENCES

[1] A. Backhaus, H. Brügelmann, S. Knorre, and M. W. 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] 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.

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