Development of a Scale to Assess the Linguistic and Phonological Difficulty of Passwords*

Jennifer Romano Bergstrom[1], Stefan A. Frisch[2], David Charles Hawkins[1], Joy Hackenbracht[1], Kristen K. Greene[3], Mary F. Theofanos[3], and Brian Griepentrog[1]

[1] Fors Marsh Group, Arlington, VA, USA
{jbergstrom,dhawkins,jhackenbracht,bg}@forsmarshgroup.com
[2] University of South Florida, Tampa, FL, USA
sfrisch@usf.edu
[3] National Institute of Standards and Technology, Gaithersburg, MD, USA
{kristen.greene,mary.theofanos}@nist.gov

Abstract. Institutions often require or recommend that their employees use secure, system-generated passwords. It is not clear how well linguistic and phonological language properties map onto complex, randomly-generated passwords. Passwords containing a mix of letters, numbers, and other symbol characters may or may not be similar to common patterns in spoken or written English. The Linguistic Phonological Difficulty (LPD) scoring rubric was created by considering the extent to which a string of characters in a password resembles ordinary spoken or written language patterns. LPD is a score calculated through a six-rule process that considers these spoken and written patterns of English as well as memory load. These rules can be applied to any password. Our research explores mapping linguistic and phonological language properties onto complex randomly generated passwords to assess behavioral performance.

Keywords: passwords, memorability, linguistics, phonology.

1 Introduction

The username-password authentication system is used to protect information from unwanted or illegal use and inspection [1]. Generally organizational password policies require criteria such as minimum number of characters, inclusion of symbols, capital letters and numbers to increase the complexity and security of passwords. However, there is typically an interaction between password complexity, memorability and security [2]. Specifically, people are more likely to remember passwords when they are required to generate items instead of just read them [3] and when deep information processing and elaboration are necessary [4]. Deeper processing increases recall because more connections are made between elements in a password, thus providing more neural pathways to aid in recollection. Additionally, users tend to generate

* The rights of this work are transferred to the extent transferable according to title 17 U.S.C. 105.

P.L.P. Rau (Ed.): CCD/HCII 2014, LNCS 8528, pp. 131–139, 2014.

passwords that are easy to remember, such as biographical data (e.g., birth date, names of significant others, favorite items), and these passwords can be guessed or broken by other individuals or computer programs. Increasingly, organizations are suggesting the use of system-generated passwords to avoid such issues with user generated passwords.

Many existing password-generation software programs have options available for constructing pronounceable random passwords (e.g., non-word consonant-vowel-consonant groups) [5], but it is unclear whether such programs use algorithms based on the linguistic and phonological properties of natural language (in our case, English). While it seems obvious that leveraging the natural properties of a given language should aid users in better remembering passphrases—indeed, there are numerous studies on passphrase memorability [6, 7]—it is less clear how linguistics and phonological language properties map onto complex, randomly-generated passwords.

Our research explores this potential mapping by examining randomly-generated passwords as linguistic units. In this paper, we attempt to quantify the linguistic and phonological complexity of 10 randomly generated passwords [8] to measure how password complexity impacts behavioral performance. We aimed to answer the following questions:

1. Is there a positive relationship between linguistic and phonological complexity and password entry errors, such that as linguistic and phonological complexity increases, errors increase?
2. Is there a positive relationship between linguistic and phonological complexity and the average time it takes to enter a password, such that as linguistic and phonological complexity increases, entry time increases?
3. Is there a positive relationship between linguistic and phonological complexity and self-reported password difficulty, such that as linguistic and phonological complexity increases, perceived difficulty increases?

2 Linguistic Phonological Difficulty

We quantified the complexity of 10 experimental passwords [8] by creating linguistic phonological difficulty (LPD) scores. The LPD scoring rubric was created by considering the extent to which a string of characters in a password resembles ordinary spoken or written language patterns. The over-arching rationale is that any password character string that has the structure of an English sentence or pronounceable unit would benefit from connections to familiar long-term memory chunks [9].

Passwords containing a mix of letters, numbers, and other symbol characters may or may not be similar to common patterns in spoken or written English. For example, punctuation generally comes at the ends of written phrases or sentences and never at the beginning of a sentence. Even in spoken English, final boundaries are marked with lengthening and pitch change [10]. Thus, we might expect a password to be easier to process if it has symbols that break the password into smaller chunks in the same manner as punctuation breaks up a sentence. Similarly, in written English, capitalization is used at the beginning of sentences, the beginning of some words, or throughout an acronym or abbreviation. Thus, we might expect a password with chunks that have

initial capitalization or capital letters throughout (e.g., Pfd, SQEF) to be easier to process than a chunk with mixed capitalization (e.g., pFd, sqeF).

Less obvious patterns of language use can also be violated in randomly generated passwords. Letters and digits usually appear in separate chunks (e.g., an address like 4401 Crater Lake Road, Tampa, FL, 33624), and so we might expect passwords with a mixing of letters and numbers to be more difficult. Further, most random strings such as d3bn can only be remembered by remembering each unit; however some arrangements can create pronounceable chunks even if they are nonsensical (e.g., b8er as "baiter").

LPD is a score calculated through a six-rule process that considers these spoken and written patterns of English as well as memory load. These rules can be applied to any password. The six-rule coding process is as follows, and the LPD score is the sum of the point score of each of the rules.

> Rule 1: Parse by symbol. If the string begins with a symbol, then put phrase boundaries before each symbol, and score 1 point for un-sentence-like parsing difficulty. Otherwise, put phrase boundaries after each symbol and score 0 points for difficulty. (Symbol start)
>
> Rule 2: Score for memory load difficulty based on the number of phrases – 1-2 phrases = 0 points, 3 phrases = 1 points, 4 phrases = 3 points, 5 phrases = 5 points, 6 phrases = 7 points, such that for each additional phrase = 2 points. (Number of chunks)
>
> Rule 3: Score for memory load difficulty based on the size of each phrase – For each phrase of 1-3 characters = 0 points, 4-5 characters = 1 point, 6-7 characters = 2 points, 8-9 characters = 3 points, 10-11 characters = 4 points, 12-13 characters = 5 points. (Number of characters)
>
> Rule 4: Score 1 additional difficulty point for each letter string (i.e., two or more letters in a row) within a phrase with un-sentence-like capitalization (e.g., a capital letter anywhere in the string other than the initial position). NOTE: A string with all capital letters = 0. (possible points: 1 or 0 per string – there may be more points if there is more than one string within a password). (Unsentence like capitalization)
>
> Rule 5: Score 1 additional difficulty point for each phrase with mixed character strings (e.g., 8f1) as opposed to segregated letter strings (e.g., 81f or f81) for strings of 3 or more characters. (Mixed character string)
>
> Rule 6: Subtract 1 difficulty point for each phrase with a pronounceable character sequence (e.g., kef is a pronounceable syllable, kfe is not; this also can be applied to a small set of number and letter mixes, for example 8t1 is 'eighty one'). (Pronounceability chunking)

The 10 experimental passwords assessed in this study and their LPD scores are displayed in Table 1. The decriptive characteristics (i.e., how often the rules occurred across all 10 passwords) are displayed in Table 2. The passwords were those used in [8].

Table 1. The 10 experimental passwords and LPD scores

Password	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6	LPD Score
5c2'Qe	0	0	1	0	1	0	2
m#o)fp^2aRf207	0	3	2	1	0	-1	5
m3)61fHw	0	0	1	1	0	0	2
d51)u4;X3wrf	0	1	2	0	1	0	4
p4d46*3TxY	0	0	3	1	1	0	5
q80<U/C2mv	0	1	2	0	1	0	4
6n04%Ei'Hm3V	0	1	2	0	2	-1	4
4i_55fQ$2Mnh30	0	1	3	1	0	0	5
3.bH1o	0	0	1	1	1	0	3
a7t?C2	0	0	1	0	1	0	2

Table 2. Descriptive characteristics of LPD of the 10 experimental passwords

Variable	***Mean***	***SD***	**Min**	**Max**
Rule 1-Symbol Start	0	0	0	0
Rule 2-No. of chunks	0.7	0.90	0	3
Rule 3-No. of characters	1.8	0.75	1	3
Rule 4-Unsentence like capital	0.5	0.50	0	1
Rule 5-Mixed char. string	0.8	0.60	0	2
Rule 6-Sounds like	-0.2	0.40	-1	0
LPD Score	**3.6**	**1.20**	**2**	**5**

2.1 LPD Coding: Inter-rater Agreement

Before LPD scores were calculated for the 10 experimental passwords (shown in Table 1), we conducted training in which two coders (with Human Factors and Industrial/Organizational Psychology backgrounds), blind to the purpose of the study, were taught how to rate LPD. One of the main researchers trained the two coders by reviewing the rules and using examples that were similar in nature to the experimental passwords to demonstrate the coding. The two coders then worked on passwords independently and then the three (researcher and coders) reviewed the independent coding and discussed discrepancies. The coders engaged in this frame of reference training on three practice sets (8, 11, and 11 passwords, respectively) until consensus was reached on the scoring rules. After the frame of reference training, the coders received a final set of 78 passwords and coded them independently. They coded the

individual rules and summed them for the total LPD score. An agreement for each rule was calculated by dividing the number of agreed upon scores by the total number of passwords. Inter-rater agreement for the each of the six LPD scoring criteria was at or above 95%. Seven passwords were scored differently on no more than two rules – the coders and the researcher discussed the discrepancies, and the coders came to a consensus. There was one disagreement on number of characters, four on unsentence-like capitalization, and three for pronounceability chunking. The pronounceability chunk is more subjective than the other rules; researchers expected for some scoring variability to exist in this rule. The discrepancy for number of characters can be attributed to a simple error, but the number of errors in unsentence-like capitalization could suggest that some ambiguity remained on what qualified as a letter string. Despite these discrepancies, the high degree of inter-rater agreement suggests that it is possible for two independent coders to be trained in the LPD scoring criteria and reach consensus for LPD on a large set of randomly generated passwords. Finally, the two coders independently rated the 10 passwords used in this study. They agreed on all but one rule for one password (6n04%Ei'Hm3V); they discussed the discrepancy ("Hm" can be considered a pronounceable syllable; as if one was pondering) and came to a consensus on the score.

3 Applying the LPD Scale

In order to test the efficacy of the LPD scale we used the performance data (times, errors, and self-reported difficulty ratings) reported in [8], a password typing entry and memorization study performed on smartphones and tablets. In [8], participants were asked to enter a randomly generated password 10 times, on the entry screen. The passwords are shown in Table 1.

The errors were operationalized as the total number of times the password was entered incorrectly during this entry phase. For entry time, we computed an index of the average time (in seconds) participants spent entering each password during this particular phase. Participants in [8] completed a difficulty-rating questionnaire in which they rated the perceived difficulty of each password on a 5-point Likert scale (1=Very Difficult; 5=Very Easy). To ease in interpreting effects consistent with the error and entry time data, the difficulty ratings were reverse scored, such that 5=*Very Difficult* and 1=*Very Easy*.

3.1 Smartphone

Table 3 displays the descriptive statistics for the smartphone performance data used to test the LPD scale. To examine the relationship between LPD and each of the performance outcomes, we first estimated the zero-order correlations between LPD and performance, in terms of the number of errors, average entry time, and self-reported difficulty. As shown in Table 4, as LPD increased so did the number of entry errors (r = .28), the longer the average entry time (r = .60), and the greater the perceived difficulty (r = .55).

Table 3. Smartphone Performance data

Variable	*Mean*	*SD*	Min	Max
Entry Errors	2.2	3.03	0	10
Mean Entry Time	15.26	8.87	3.38	80.73
Difficulty Rating	3.09	1.31	1	5

Because number of errors and mean entry time positively correlated with participants' age and gender (rs > .23 and > .08, respectively) and negatively correlated with password order (rs > -.07), we ran a series of linear regressions to test the relationship between LPD and each of the performance outcomes controlling for age, gender, and order. To account for the repeated measures aspect of the design, we adjusted for dependence when estimating standard errors. The results are the same as reported in Table 4.

Table 4. The correlation of LPD with performance outcomes for Study 1. $^{***}p < .001$

	LPD Score	Entry Errors	Mean Entry Time	Difficulty Rating
LPD Score	1			
Entry Errors	$.28^{***}$	1		
Mean Entry Time	$.60^{***}$	$.39^{***}$	1	
Difficulty Rating	$.55^{***}$	$.35^{***}$	$.46^{***}$	1

Next, instead of focusing on the overall LPD score, we examined the relationship between the individual LPD rules and each of the performance outcomes. All rules were entered as simultaneous predictors in a linear regression, with the exception of Rule 1 (Symbol Start) which was excluded because it had zero variance among the ten passwords. The standardized β coefficients for the individual LPD rules are shown in Table 5. These results suggest that the number of characters per chunk (Rule 3), relative to the other LPD rules, appears to be most related to the number of entry errors, mean entry time, and perceived difficulty (βs = .24, .42, .50, respectively).

Table 5. Standardized β coefficients regressing LPD onto performance outcomes in Study 1. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

	Entry Errors	Mean Entry Time	Difficulty Rating
Rule 2-No. of chunks	.01	$.24^{***}$	-.00
Rule 3-No. of characters	$.24^{***}$	$.42^{***}$	$.50^{***}$
Rule 4-Unsentence like capital	$-.12^{***}$	$-.21^{***}$	$-.26^{***}$
Rule 5-Mixed char. string	$-.16^{*}$	$-.20^{***}$	$-.33^{***}$
Rule 6-Sounds like	$-.19^{**}$	$-.31^{***}$	$-.39^{***}$

3.2 Tablet

Table 6 displays the descriptive statistics for the performance data for tablet.

Table 6. Descriptive statistics for performance data for Tablet

Variable	*Mean*	*SD*	Min	Max
Entry Errors	1.21	2.46	0	10
Mean Entry Time	14.81	12.62	3.33	154.01
Difficulty Rating	3.06	1.28	1	5

Our analyses were the same as in smartphone. Again, we estimated zero-order correlations between LPD and each of the performance outcomes. As shown in Table 7, the results were the same as in smartphone. Specifically, as LPD increased, so did the number of entry errors (r = .14), the longer the average entry time (r = .42), and the greater the perceived difficulty (r = .52).

Table 7. The correlation of LPD with performance outcomes for Study 2. $^{***}p < .001$

	LPD Score	Entry Errors	Mean Entry Time	Difficulty Rating
LPD Score	1			
Entry Errors	.14***	1		
Mean Entry Time	.42***	.32***	1	
Difficulty Rating	.52***	.28***	.40***	1

Second, we again estimated a series of linear regressions to test the relationship between LPD and each of the performance outcomes controlling for age, gender, and order. To account for the repeated measures aspect of the design, we adjusted for dependence when estimating standard errors. The results are the same as reported in Table 7.

Table 8. Standardized β coefficients from a regression of the LPD onto performance outcomes in Study 2. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

	Entry Errors	Mean Entry Time	Difficulty Rating
Rule 2-No. of chunks	-.03	.07	-.03
Rule 3-No. of characters	.11**	.33***	.48***
Rule 4-Unsentence like capital	-.09	-.18***	-.27***
Rule 5-Mixed char. string	-.13	-.16***	-.33***
Rule 6-Sounds like	-.20**	-.29***	-.41***

Last, we examined the relationship between the individual LPD rules and each of the performance outcomes. The standardized β coefficients for the individual LPD rules are shown in Table 8. These results suggest that the number of characters per chunk (Rule 3), relative to the other LPD rules, appears to be most related to mean entry time and perceived difficulty ($\beta s =$ .33, .48, respectively).

4 Conclusion

Organizational password policies often require criteria such as minimum number of characters, inclusion of symbols, capital letters and numbers to increase the complexity and security of passwords. The effectiveness of passwords is diminished by the fact that as passwords become more complex, they are harder to remember [2]. Complex passwords can be given to users; however it is more difficult for people to remember passwords that they do not create themselves [3]. We developed a scale for quantifying the linguistic and phonological complexity of randomly generated passwords. We tested the scale to examine the extent to which it predicted task completion time, error rates, and perceived difficulty from previously collected data [8]. Given the parsing of passwords into chunks by special characters (non-letter and non-digit), Rule 3 (number of characters per chunk) appeared to contribute the most to perceived difficulty, number of errors, and average entry time. While this rule was the most predictive, it may be possible to simplify the LPD scale by reducing the set of remaining rules. Nonetheless applying measures of linguistic and phonological difficulty to complex system generated passwords seems a promising avenue of future research.

One caveat to this work is that exploiting spoken or written language patterns is necessarily language specific. For example, Spanish questions and exclamations use punctuation before and after the sentence, unlike English. Similarly, the set of characters that create a possible word string in English (e.g., strab) may not be possible in another language (e.g., strab is not a possible word in Spanish).

Future research is needed with a larger number and variety of passwords. As shown in Table 1, there was minimal variance in the current 10 passwords on some measures, which makes it difficult to confidently assess the effects of the individual rules. The most predictive rule, Rule 3, was also one of the rules with the greatest range. The list of passwords in the present study did not explore extreme violations of the other rules in use of symbols, number of chunks, use of capitalization, or mixing of character strings. Future work should examine these elements and the impact on performance and the importance of LPD above and beyond traditional measures of complexity, such as number of characters/screens [8]. In addition, the point scores for the LPD rules implicitly weight the rules (e.g. Rule 2 adds two points per phrase while Rule 3 adds one points per character within a phrase). The different linguistic and phonological aspects of the rules may indeed have different weights that must be empirically determined.

Acknowledgements. This work was funded by the Comprehensive National Cybersecurity Initiative (CNCI). The authors gratefully acknowledge Dr. Yee-Yin Choong for her thoughtful review and feedback and Ashley Barbee for assistance in coding.

References

1. Zviran, M., Haga, W.J.: Password Security: An Empirical Study. Journal of Management Information Systems 15(4), 161–184 (1999)
2. Keith, M., Shao, B., Steinbart, P.J.: The Usability of Passphrases for Authentication: An Empirical Field Study. International Journal of Human-Computer Studies 65, 17–28 (2007)
3. Vu, K.L., Proctor, R.W., Bhargav-Spantzel, A., Tai, B., Cook, J., Schultz, E.E.: Improving Password Security and Memorability to Protect Personal and Organizational Information. International Journal of Human-Computer Studies 65, 744–757 (2007)
4. Craik, F.I.M., Lockhart, R.S.: Levels of Processing: A Framework for Memory Re-search. Journal of Verbal Learning and Verbal Behavior 11, 671–684 (1971)
5. Gasser, M.: A Random Word Generator for Pronounceable Passwords. Mitre Corporation Report MTR-3006 (1975)
6. Bonneau, J.: Linguistic Properties of Multi-Word Passphrases. In: USEC Workshop on Useable Security, Kralendijk, Bonaire, Netherlands (2012)
7. Bonneau, J.: The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In: IEEE Symposium on Security and Privacy (2012)
8. Greene, K.K., Gallagher, M.A., Stanton, B.C., Lee, P.Y.I.: Can't Type That! P@$$w0rd Entry on Mobile Devices. In: Proceedings of the Human Computer Interaction International Conference, Crete, Greece (2014)
9. Gobet, F., Lane, P.C.R., Croker, S., Cheng, P.C.-H., Jones, G., Oliver, I., Pine, J.M.: Chunking Mechanisms in Human Learning. Trends in Cognitive Science 5, 236–243 (2001)
10. Pierrehumbert, J.B.: The Phonetics and Phonology of English Intonation. Unpublished Ph.D. dissertation. MIT (1980)