RESEARCH PAPER:

SURVEY ON THE USER PERCEPTION OF THE HOMOGRAPHIC CHARACTER SET SPECIFIED BY JGP

September 28, 2020

Tatsuya Mori, Waseda University

1. BACKGROUND AND OBJECTIVES OF THE STUDY

The IETF standardized internationalized domain names (IDNs) in 2003, more than 15 years ago. As a countermeasure against the potential threats of IDN homograph attacks, ICANN established the "Guidelines for the Implementation of Internationalized Domain Names" in 2005. TLD registries that offer IDN registrations are supposed to register rules that define the languages and character ranges to be accepted (henceforth referred to as the IDN Table). However, the rules are not fully enforced in gTLDs and gTLD-like ccTLDs, which have many registered domain names. This situation allows IDNs to be misused for phishing (i.e., homograph attacks) and hinders the healthy spread of IDNs. In recent years, it has been pointed out that visually similar characters exist not only between different scripts but also within the same script. Without any legitimate rules, the same situation may happen on IDN TLDs, and it is the reason why the Root LGR is being developed. Mixed script was strictly limited from the beginning of Root LGR development, but avoiding visually similar characters was raised later. Therefore, avoiding the mixing of scripts is not sufficient to prevent homograph attacks. A universal mechanism is needed to define a set of possible visually similar characters within the same script, and to notify Internet users of their use somehow.

This survey targets the Root LGR Japanese Generation Panel's (JGP) candidate character sets (8 pairs, see Table 1) and investigates user perceptions of homoglyphs and homographs (words). Based on the survey's results, we aim to develop preventive measures against IDN homograph attacks.

2. SURVEY SPECIFICATIONS

2.1 CHARACTER SET

In this study, we use a set of characters shown in Table 1, specified by the JGP as a candidate, which JGP says are recommended as confusable characters in Japanese script (Hiragana, Katakana and Kanji) by Unicode Consortium.

Table 1: JGN's candidate similar character pairs (8 pairs)

Hiragana	Katakana	Chinese Letters
へ	へ	
	ニ	二
	ハ	八
	カ	力
	ト	卜
	ロ	口
	タ	夕
	エ	工

In this exercise, in addition to the study of visually similar characters, we will also study words that contain visually similar characters.For this purpose, we will use the words shown in Table 2. These are the words with a high frequency of occurrence, including the visually similar characters shown in Table 1. In extracting the words, we used the lexicon table of the "Modern Japanese Written Language Equilibrium Corpus." We extracted the words that frequently appear in the corpus, including each character.

https://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

Table 2. Frequent words containing similar characters

ヘリコプター　(helicopter)
コミュニケーション　(communication)
シャンハイ　(Shanghai)
ホッカイドウ　(Hokkaido)
インターネット　(internet)
プロジェクト　(project)
コンピューター　(computer)
ダイエット　(diet)

2.2 Evaluation conditions

This study investigates the influence of font type, font size, and user's linguistic background on the perception of homoglyphs/homographs. The nine typical font families listed in Table 3 were used. These are the standard used to display Japanese characters in major web browsers and operating systems on both desktop and mobile environment. As a result of preliminary experiments, three font sizes were selected: 18 px, 24 px, and 36 px. As for the linguistic background, we experimented with non-Japanese users and Japanese users.

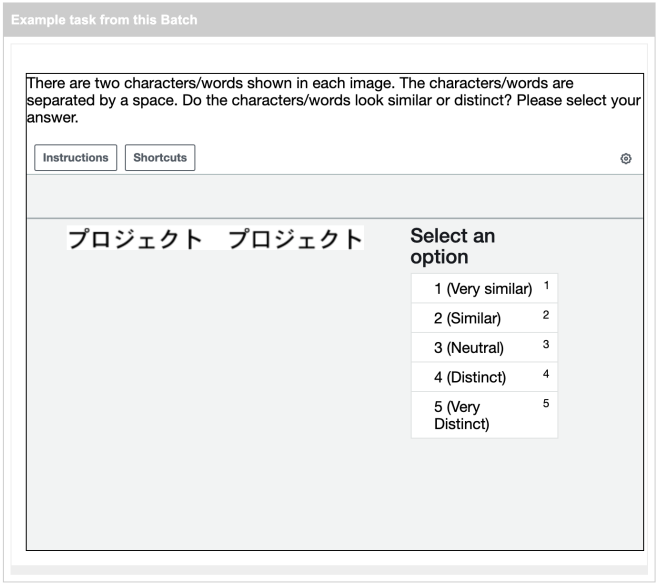
Table 3. List of fonts used in the experiment

Font
MS-P Gothic
MS-P Mincho
MS Gothic
Meiryo
Yu-Mincho
Hiragino Kaku Gothic
Hiragino Mincho ProN W3
Hiragino Kaku Gothic W3
Osaka

2.3 Evaluation method

The similarity judgment for each letter and word is based on a 5-point Likert scale. Figure 1 shows an image of the screen used for our user study. The characters and words are displayed as images so that they are not affected by the browser environment.

Fig. 1. Image of the response screen in MTurk



Instructional manipulation check (IMC) was introduced to take into account the user's attention, i.e., the quality of the responses in the user survey, by presenting apparently different letter/word pairs and checking whether they were judged correctly. If a participant made a wrong answer, all the answers made by the participant were removed.

2.4 Experimental setup

There were 16 pairs of target letters/words, eight pairs of homoglyphs, and eight pairs of homographs. Aside from these, four dummy pairs were also used to check if the participant fully understand specifically what (s)he is asked, while paying attention to the assignments. A total of 20 different pairs were created. In the user survey, 20 participants responded to each pair. Since there are nine font types and three font size types, all pair combinations are $20 \times 9 \times 3 = 540$. This is the case. From these 540 pairs, 300 randomly sampled pairs were tested. Each participant was presented with shuffled results. The 300 pairs were also divided into six sets. Each set contains a set of 50 pairs. Each participant is presented with 50 sets of results to minimize the load on the user. Participants can work on more than one set if they wish. In the end, we get 1000 responses per set (50 sets x 20 participants). Total: $6 \text{ sets} \times 1000 = 6000$ responses. From this, we remove the user responses that are excluded as a result of the IMC.

3. Tools used in the survey

The crowdsourcing platforms used for user evaluation were as follows:

- Non-Japanese-speaking users: Amazon Mechanical Turk (MTurk) https://www.mturk.com/
- Japanese-speaking users: Lancers https://www.lancers.jp/

For MTurk, we used the crowd image classifier API; for Lancers, we used Google Forms. In this case, we used Google App Script to generate the forms automatically. The data collected by the respective crowdsourcing platforms were processed using a scripting language (we used Ruby).

4. EXPERIMENTAL RESULTS

4.1 NON-JAPANESE USERS

When we conducted a preliminary experiment with MTurk, we could not get any Japanese participants. Therefore, we assume that the following results of the MTurk experiment are all for non-Japanese users.

Tables 5, 6, 7, and 8 summarize the evaluation results by character type, word type, font type, and character size type, respectively.

Table 5: Results of character-by-character similarity ratings. Number of ratings and mean score (Avg)

	Very similar	similar	neutral	distinct	Very distinct		
Chars	1	2	3	4	5	Total	Avg
ヘ	207	43	1	2	0	253	1.2
ニ	117	98	8	11	1	235	1.6
ハ	5	90	37	77	21	230	3.1
カ	169	102	4	8	0	283	1.5
ト	104	83	2	5	0	194	1.5
ロ	38	130	14	42	14	238	2.4
タ	149	99	5	15	1	269	1.6
エ	54	138	16	21	3	232	2.1

Table 6: Results of word-by-word similarity ratings. Number of ratings and mean score (Avg)

	Very similar	similar	neutral	distinct	Very distinct		
Words	1	2	3	4	5	Total	Avg
ヘリコプター	260	7	0	1	1	269	1.1
コミュニケーション	207	70	1	4	0	282	1.3
シャンハイ	65	188	25	76	3	357	2.3
ホッカイドウ	139	21	0	0	0	160	1.1
インターネット	236	29	0	1	1	267	1.1
プロジェクト	195	73	4	15	0	287	1.4
コンピューター	247	39	1	1	0	288	1.2
ダイエット	138	109	4	1	0	252	1.5

Table 7: Results of similarity ratings for each font Number of ratings and mean score (Avg)

	Very similar	similar	neutral	distinct	Very distinct		
Font	1	2	3	4	5	Total	Avg
MS-P Gothic	405	151	7	29	10	602	1.5
MS-P Mincho	270	101	13	17	2	403	1.5
MS Gothic	210	165	15	37	2	429	1.7
Meiryo	248	151	13	27	3	442	1.6
Yu-Mincho	175	140	16	38	7	376	1.8
Hiragino Kaku Gothic	245	132	7	25	4	413	1.6
Hiragino Mincho ProN W3	179	165	20	41	5	410	1.8
Hiragino Kaku Gothic W3	204	180	13	37	11	445	1.8
Osaka	394	134	18	29	1	576	1.5

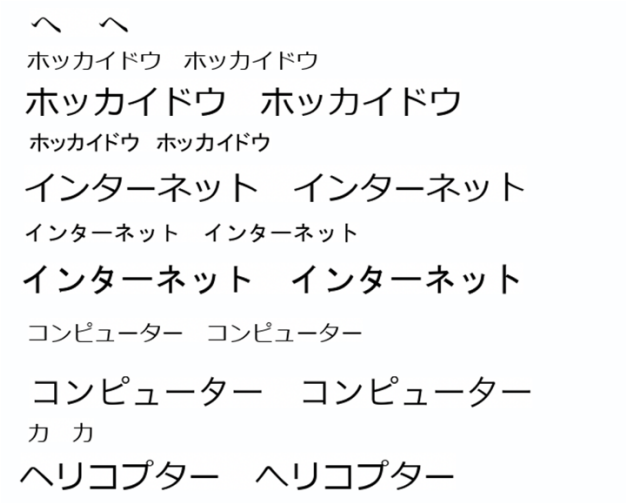
Table 8: Results of similarity ratings for each font size. Number of ratings and mean score (Avg)

	Very similar	similar	neutral	distinct	Very distinct		
Size	1	2	3	4	5	Total	Avg
18px	718	387	33	70	10	1218	1.6
24px	735	454	41	98	16	1344	1.7
36px	877	478	48	112	19	1534	1.6

Fig. 2: Most distinguished letters and words

Parameter	Score	
2-4-36	3.61	ハ　八
2-6-18	3.39	ハ 八
2-0-36	3.29	八　八
2-8-24	3.17	八　八
2-0-24	3.17	八　八

Fig. 3: Most indistinguishable letters and words (all scored 1)



Highlights of the experiment for non-Japanese users are as follows:

- As a stand-alone character, the“へ”is the hardest to distinguish.
- It is hard to tell them apart in general when it comes to words.
- Only the“ハ”and“八”are reasonably recognizable.
- There are no significant differences between the fonts.
- There are no significant differences between sizes.

A SUMMARY OF THE FINDINGS

- The more a word consists of multiple letters, the harder it is to distinguish between them.
- The“へ”is incredibly difficult to distinguish by itself, and the“ハ”is easy to distinguish.
- The overall trend is not dependent on font or font size.

4.2 JAPANESE USERS

The Lancers survey is conducted for Japanese users. All the descriptions of the experiment are given in Japanese. Therefore, we assume that all the participants in the survey were users fluent in Japanese.

Tables 9, 10, 11, and 12 summarize the study results by character type, word type, font type, and character size type, respectively.

Table 9: Results of character-by-character similarity ratings. Number of ratings and average score (Avg) for Japanese users

	Very similar	similar	neutral	distinct	Very distinct		
Chars	1	2	3	4	5	Total	Avg Score
ヘ	175	73	0	7	0	255	1.4
ニ	55	108	8	67	9	247	2.5
ハ	0	95	22	100	25	242	3.2
カ	48	167	5	70	5	295	2.4
ト	37	127	1	37	3	205	2.2
ロ	11	142	8	78	7	246	2.7
タ	45	162	6	65	4	282	2.4
エ	12	110	8	96	25	251	3

Table 10: Results of word-by-word similarity ratings. Number of ratings and mean score (Avg) for Japanese users

	Very similar	similar	neutral	distinct	Very distinct		
Words	1	2	3	4	5	Total	Avg Score
ヘリコプター	269	24	0	2	0	295	1.1
コミュニケーション	94	153	3	46	3	299	2
シャンハイ	16	173	25	138	30	382	3
ホッカイドウ	86	65	1	20	0	172	1.7
インターネット	177	103	1	3	0	284	1.4
プロジェクト	106	141	2	48	6	303	2
コンピューター	138	129	4	30	1	302	1.8
ダイエット	21	152	9	73	6	261	2.6

Table 11: Results of similarity ratings for each font. Number of ratings and mean scores (Avg) for Japanese users

Font family	Very similar 1	similar 2	neutral 3	distinct 4	Very distinct 5	Total	Avg Score
MS-P Gothic	270	254	11	91	18	644	2
MS-P Mincho	139	197	8	64	8	416	2.1
MS Gothic	100	190	11	124	22	447	2.5
Meiryo	146	234	11	79	10	480	2.1
Yu-Mincho	76	183	10	111	21	401	2.5
Hiragino Kaku Gothic	117	223	10	80	1	431	2.1
Hiragino Mincho ProN W3	115	183	15	101	14	428	2.3
Hiragino Kaku Gothic W3	102	196	16	141	18	473	2.5
Osaka	225	264	11	89	12	601	2

Table 12: Results of similarity ratings by character size. Number of ratings and mean score (Avg) for Japanese users

Size	Very similar 1	similar 2	neutral 3	distinct 4	Very distinct 5	Total	Avg Score
18px	435	564	19	241	28	1287	2.1
24px	399	625	46	302	48	1420	2.3
36px	456	735	38	337	48	1614	2.2

Fig. 4: Most distinguishable letters and words for Japanese users

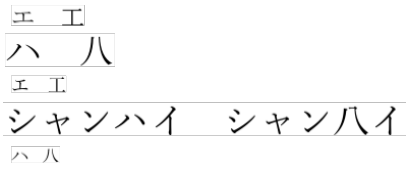
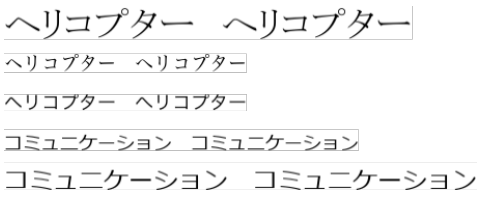


Fig. 5: Most indistinguishable characters and words (all scored 1) for Japanese users



Highlights of the experiment for Japanese users are as follows:

- As a stand-alone character, the “へ” is the hardest to distinguish.
- It is hard to tell them apart in general when it comes to words.
- Only the “ハ” and “八” are reasonably recognizable.
- There are no significant differences between the fonts.
- No significant differences between sizes are observed.

A SUMMARY OF THE FINDINGS

- The more a word consists of multiple letters, the harder it is to distinguish between them.
- By themselves, “へ” and “ニ” are incredibly difficult to distinguish, while “ハ” and “ロ” are relatively easy to distinguish.
- The overall trend is not dependent on font or font size.
- Similar to Mturk’s results ⇒ The results were similar regardless of languages.
- However, Japanese users are better able to distinguish between similar characters.

5. SUMMARY

Our extensive user studies have shown that homographs (words) are more indistinguishable than homoglyphs (letters) and that this tendency is independent of linguistic backgrounds. Because actual domain names consist of words, the result implies that homograph IDNs containing visually similar characters are challenging to distinguish. In general, Japanese users had a higher success rate in distinguishing between similar characters specified by the JGP. It became clear that some visually similar characters specified by the JGP were difficult even for the Japanese participants to distinguish. A surprising result was that the discrimination success rate was not affected by font family and size. This result suggests that the original characters’ glyph structures have a more significant impact on human perception than the differences in the display of the characters in the browser (i.e., font family and size). The results also show that the pairs of eight homographic characters specified by the JGN tend to be difficult to distinguish, but the degree of similarity varies. In particular, it is essential to note that “へ/へ” is a problematic character to distinguish regardless of language. In this study, the eight pairs of similar character sets are evaluated, but this evaluation method is not specific to Japanese and is applicable to other scripts and languages also, we hope this report contributes to the Root LGR development.