

Confusing String Similarity from the aspect of Japanese scripts

22 February 2023
APTLD

Hiro Hotta

Characters in Japanese Scripts

- Script and Language
 - Words of daily-use contain characters in 4 scripts : Kanji, Hiragana, Katakana, and ASCII alpha-numeric
 - Characters can be mingled in any order in words
 - E.g., “400gのラム肉” numeral – alphabet – Hiragana – Katakana - Kanji
 - Characters defined in JIS (Japanese Industrial Standard) level-1 and level-2 are used in daily life
 - 6,000+ characters
 - Kanji is used also in Chinese language and Korean language
- Background of variants in LGR (Language Generation Rules)
 - Characters with the same meaning and pronunciation that can be used interchangeably
 - E.g., “学” and “學” in Chinese language
 - Characters with visual identicalness



focus of this session

Visual Similarity vs. Visual Identicalness

- Among 6,000 Kanji characters, no two characters have an identical shape. However, there are characters with small differences from each other.
 - E.g., personal names 渡邊 and 渡邊
- As Japanese daily life is in such an environment, daily readers of Japanese words/sentences naturally can identify small differences between Kanji characters.
- However, across scripts, there may be characters with identical shapes – especially depending on fonts.
 - One-stroke symbols/marks can look easily identical
 - — and —, \ and \
 - For characters, we conducted 2 kinds of field research to find character sets that look identical

Field Research 1

- Online human perception experiment
 - Unicode Consortium recommends pairs of characters with Japanese scripts to be handled as “confusable”
 - <https://www.unicode.org/Public/security/latest/confusables.txt>
 - They were checked if they are truly confusable or not by field-researching with human subject, under multiple environment
- Environment
 - a. pairs of isolated confusable characters + pairs of confusable strings
 - b. 9 popular fonts with 3 font sizes
 - c. 40 examinee – 20 read Japanese well, while 20 don't
 - d. every experiment (=every combination of a. b. c.) gives rate 1-5
 - 1(very similar), 2(similar), 3(neutral), 4(distinct), 5(very distinct)
- result
 - All pairs are rated less than 3.2
- Conclusion
 - All pairs look identical enough to be confused

Confusable characters sets

- Based on the Unicode list regarding confusable characters, following 10 pairs were candidates to be visually identical

	Hiragana	Katakana	Kanji
1	へ	ヘ	
2	べ	ベ	
3	ぺ	ペ	
4		ニ	二
5		ハ	八
6		カ	力
7		ト	卜
8		ロ	口
9		タ	夕
10		エ	工

Words containing these characters

ヘリコプター
 シャーペン
 シャンハイ
 インターネット
 コンピューター

ショベル
 コミュニケーション
 ホッカイドウ
 プロジェクト
 ダイエット

Field Research 2

- Online survey question
 - other than those characters pre-listed in Field Research 1, do you have experiences of confusion caused by character identicalness?
- Result (73 respondents)
 - no pairs other than pre-listed ones confused more than 2 respondents due to visual identicalness.
 - other than pre-listed pairs, 5 character pairs confused 2 respondents due to visual identicalness.
 - other than pre-listed pairs, 8 character pairs confused 1 respondent due to visual identicalness.
- Summary
 - Less than 3% (2 out of 73) of the respondents were confused by character pairs other than pre-listed 10 pairs.
- Conclusion
 - It is appropriate NOT to regard other pairs as identical than pre-listed 10 pairs in Field Research 1.

Summary

- Japanese situation
 - Japanese words use 4 kinds of scripts – in a mingled way
 - Characters within one script have different shapes
 - Japanese readers are accustomed to differentiate characters within one script
 - Some characters stemmed from different scripts may be visually identical - especially on computer/smart-phone screens
- How to define visual identicalness
 - 2 kinds of field research were conducted
 - (1) Online human perception experiment
 - (2) Online survey to ask respondents' experience on confusion due to visual identicalness
- Result
 - 12 pairs of characters and symbols/marks from different scripts are regarded identical and defined as variants in LGR

Thank you

Hiro Hotta
hotta@jprs.co.jp