Predictors of script choice in Japanese: data driven study

Yo Sato
Satoama Language Services & Cerence Ltd.
satoama@gmail.com

Kevin Heffernan
Kwansei Gakuin University
kevin.heffernan@kwansei.ac.jp

Target phenomenon: orthographic alternation in Japanese
- options for the same word: kana (phonemic) and kanji (ideographic)
- where the choice seems to be up to the writer's preference 例: りんご/リンゴ/林檎 for ringo (apple)

Question: Is this preference discoverable from the observable factors in data?
- Such as gender, age, genre (metadata) or
- Parts of speech, character complexity (stroke count) and frequency

Method: mixed effect logistic regression

Known facts and previous claims

- Japanese has three main scripts and the vast majority of words can be written in any of them
  - Kana: phonemic, hiragana and katakana subvarieties
  - Kanji: ideographic
- Sino-Japanese words (Chinese origin, *kango*) are predominantly written in kanji
- Grammatical morphemes and particles are predominantly written in kana
- Other Japanese-native words (*wago*) could be written in kana or kanji, with varying proportions from word to word
- For the last group,
  - Seeley (2000) claims for general trends towards kana ('kana-ization') over time
  - Shibatani and Kageyama (2016): function words more in kana, content words more in kanji
  - Kaiho and Nomura (1983): frequency and complexity of kanji influences the choice
  - Smith and Schmidt (1996): possible effects of gender, genre and age

This work attempts to find evidence for these claims, if any, from data

Data

- Balanced Corpus of Contemporary Written Japanese (NINJAL, Maekawa et al. 2014)
  - comes with metadata, gender, author age and genre
- Parsed by morphological analyser (MeCab, Kudo et al 2004)
  - analysis focused on nouns, pronouns and adjective stems
  - Set a threshold for frequency and entropy
    - 625 unambiguous word types used
    - Token count per type 122, total 76250

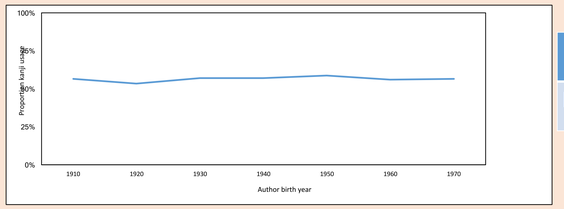
Examples of extracted triples

Top 5 in frequency	Top 5 in entropy
こと/事/コト	きれい/キレイ/綺麗
言葉/ことば/コトバ	オシャレ/おしゃれ/お洒落
だめ/ダメ/駄目	癖/クセ/くせ
タバコ/たばこ/煙草	エサ/餌/えさ
親父/オヤジ/おやじ	ケンカ/けんか/喧嘩

Method

- Mixed effect logistic regression for binary distinction, kana and kanji (no distinction between the subvarieties of kana)
  - Fixed effects:
    - Lexical subcategory
    - Genre
    - Kanji frequency and stroke count (proxy for complexity)
  - Random intercept: document ID (since the same author usually sticks to their preference throughout a single document)
- R package 'lme4'

Lack of correlation for author age and gender



	Male (n=220)	Female (n=193)
Kanji proportion	**0.57532112**	**0.576182989**

Logistic regression for other features

Categorical features
- Subcat: 9 subcategories of nouns
- Genre: 12 genres given by BCCWJ
- Details as in table below

Continuous features
- strokes: stroke count
- freq: average character frequency
- Both highly significant

```
Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.088641   0.155486   7.002 2.53e-12 ***
subcatナイ形容詞語幹 1.140204   0.054227  21.026  < 2e-16 ***
subcat一般         0.022104   0.028836   0.767 0.443341
subcat代名詞       0.630955   0.029943  21.072  < 2e-16 ***
subcat副詞可能     -0.666157   0.036623 -18.190  < 2e-16 ***
subcat形容動詞語幹  -0.052806   0.036623  -1.442 0.149339
subcat接尾         0.325229   0.031213  10.420  < 2e-16 ***
subcat数           1.247585   0.070900  17.597  < 2e-16 ***
subcat非自立       -0.715528   0.029770 -24.035  < 2e-16 ***
genre2_歴史        -0.373573   0.179573  -2.080 0.037494 *
genre3_社会科学     0.133516   0.222187   0.601 0.547895
genre4_自然科学    -0.520215   0.190565  -2.730 0.006336 **
genre5_技術・工学   -0.167756   0.181038  -0.927 0.354116
genre6_産業        -0.319830   0.167054  -1.915 0.055553 .
genre7_芸術・美術   -0.492612   0.161344  -3.053 0.002264 **
genre8_言語        -0.459171   0.159242  -2.883 0.003933 **
genre9_文学        -0.518380   0.156902  -3.304 0.000954 ***
genreOC           -0.055783   0.157529  -0.354 0.723253
genreOT           -0.912319   0.221706  -4.115 3.87e-05 ***
genreOY           -0.253470   0.157272  -1.612 0.107034
genrePM           -0.580828   0.162660  -3.571 0.000356 ***
strokes           -0.227490   0.005915 -38.459  < 2e-16 ***
freq               0.241754   0.007241  33.386  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	kanji-leaning	Ref (at p=0.74)	kana-leaning	Not significant, kanji-leaning	Not significant, kana-leaning
subcat	Numeral (数) Classifier (接尾) Pronoun(代名詞)	Action (サ変接続)	Functional (非自立) Adverbial (副詞可能)	Common (一般)	Adjectival (形容動詞語幹)
genre		Engineering (工学)	Textbook(OT) Magazine(PM) Literature(文学)	social science(社会科学)	online Q&A(OC), history(歴史), natural science(自然科学), arts(芸術), language(言語), industry(産業) and blog(ブログ)

Discussion

- Evidence for kana-ization was not found
- Some lexical subcategories show significant difference, where function items tend to be written in kana
- Frequency and stroke count of kanji are significantly correlated to kana/kanji rendering
  - The more frequent the kanji, and the fewer the strokes (less complex), the more likely the word is written in kanji
- Significant difference in some genres

Conclusion and future work

- Predictors for kana-kanji alternation sought in data-driven manner
- Some previously claimed predictors confirmed, others not
- Analysing other categories, verbs amongst others
- Extending the analysis to a three-way (multinomial) response, i.e. katakana and hiragana as well as kanji
- Combining our results with analyses involving contexts (other words in the same sentence)
- Comparing our results with human introspection