An Analysis of Japanese Loanwords

LING 1340 | Lindsey Rojtas

Terminology

- Japanese has three writing systems:
    - Hiragana (ひらがな) - a Japanese syllabic script (meaning each character = one syllable) typically used for grammatical function words and words with no Kanji equivalent
    - Kanji (漢字) - logographic Chinese characters that are used in the Japanese writing system; different combinations can have different meanings or pronunciations (天 may be pronounced “ten” or “ama”)
    - Katakana (カタカナ) - a Japanese syllabic script used for onomatopoeia and loan words from languages such as English, French, Portuguese, etc.

Gairaigo (外来語) is the word for “loan word”/”borrowed word” specifically - this is what my project is on

A list of Hiragana and Katakana

ア行 /	あ ア A	い イ I	う ウ U	え エ E	お オ O
カ行 K	か カ KA	き キ KI	く ク KU	け ケ KE	こ コ KO
サ行 S	さ サ SA	し シ SI	す ス SU	せ セ SE	そ ソ SO
タ行 T	た タ TA	ち チ CHI	つ ツ TSU	て テ TE	と ト TO
ナ行 N	な ナ NA	に ニ NI	ぬ ヌ NU	ね ネ NE	の ノ NO
ハ行 H	は ハ HA	ひ ヒ HI	ふ フ HU	へ ヘ HE	ほ ホ HO
マ行 M	ま マ MA	み ミ MI	む ム MU	め メ ME	も モ MO
ヤ行 Y	や ヤ YA		ゆ ユ YU		よ ヨ YO
ラ行 R	ら ラ RA	り リ RI	る ル RU	れ レ RE	ろ ロ RO
ワ行 W	わ ワ WA				を ヲ WO
	ん ン NN				
ガ行 G	が ガ GA	ぎ ギ GI	ぐ グ GU	げ ゲ GE	ご ゴ GO
ザ行 Z	ざ ザ ZA	じ ジ JI	ず ズ ZU	ぜ ゼ ZE	ぞ ゾ ZO
ダ行 D	だ ダ DA	ぢ ヂ DI	づ ヅ DU	で デ DE	ど ド DO
バ行 B	ば バ BA	び ビ BI	ぶ ブ BU	べ ベ BE	ぼ ボ BO
パ行 P	ぱ パ PA	ぴ ピ PI	ぷ プ PU	ぺ ペ PE	ぽ ポ PO

Motivation

- Language borrowings was fun to learn about in LING1000
- I watched a lot of anime as a middle schooler
    - The interest in the language stuck with me so I started teaching myself to read hiragana/katakana
- I wanted to do something in non-English, but I can only read hiragana and katakana characters
- Loanwords are easy to find and translate
    - Many borrowings from English just sound like the English word in a Japanese accent
        - トイレ = toire = toilet
- English was introduced into Japanese society relatively recently, but some Japanese people will use English words instead of their Japanese equivalences
    - レッド (reddo) vs 赤 (aka) - both mean red, but the former is borrowed from English
    - Is the usage of one over the other related to age?
- Some loanwords are shortened if they're a bit long
    - プロレス (puroresu) = professional wrestling - we don't say "pro-res" in English
    - What determines whether or not a word is shortened?

Big Questions and Hypotheses

1. Are shortened versions of katakana words more likely to be used in casual Japanese conversation?
2. Are age and the amount of katakana words used correlated in any way?
    a. If so, can we use a machine learning model to predict an age based off katakana words used?

HYPOTHESIS 1: Shorter versions of katakana words will be used more often than longer words, since these shortened words are likely easier to say than their longer counterparts.

HYPOTHESIS 2: Age and the amount of katakana words used are related; younger Japanese speakers will use these katakana words more than older speakers. This correlation may not be horribly strong, so a machine learning model may not be effective at age prediction

My Data

I used two corpora in my project:

- The Nagoya University Conversation Corpus
    - 129 unstructured conversations with several different participants of varying age groups
    - Ages range from late teens to early nineties
    - Most participants are female
- Balanced Corpus of Contemporary Written Japanese - Word List
    - A list of words and their web-based frequencies, as well as some other arbitrary data
    - Reddit user u/Alphyn provided a cleaned version that made my life way easier!
    - List of words used because Japanese has no word boundaries - no way to tokenize
        - Also got the idea to compare web frequencies and conversational frequencies later on from this!

Data Cleaning and Reorganization

- Word list was relatively easy to work with
- Dropped irrelevant columns
- Renamed columns that were relevant
- Dropped words without any English equivalent (names, Japanese cities)
- Dropped words with a web frequency of less than 75
- Ended up w/ approx. 5000 words

	katakana	translation	frequency
0	パーセント	percent	63392
1	アメリカ	America	28243
2	ページ	page	24642
3	センター	center	20664
4	サービス	service	16630

Data Cleaning and Reorganization

- Conversational data was much more of a process
- Created my own .csv files
    - My conversation corpus was only text files; I wanted to organize those contents by which file they were in and which participants they were spoken by
    - Tedious, but way worth it
- Imported in text data, but ran into many issues trying to clean it …

In [25]:

```
def readtxt(fn):
    f = open(glob.glob('../privdata/nucc/' + fn)[0], encoding="utf8")
    text = f.read()
    f.close()
    return text

byfile['content'] = byfile['file'].apply(readtxt)

byfile.head()
```

Out[25]:

	file	participants	content
0	data001.txt	F107 F023 M023 F128	@データ１（約３５分）\n@収集年月日：２００１年１０月１６日\n@場所：ファミリーレストラ...
1	data002.txt	F107 F023 F128	@データ２（６０分）\n@収集年月日：２００１年１０月１６日\n@場所：ファミリーレストラン...
2	data003.txt	F033 F056	@データ０３（４３分）\n@収集年月日：２００１年１０月２３日\n@場所：車中（某大から所属...
3	data004.txt	M018 F128	@データ０４（３５分）\n@収集年月日：２００１年１０月２３日\n@場所：車中（知立駅より西...
4	data005.txt	M023 F128 F116 M026	@データ０５（５５分）\n@収集年月日：２００１年１０月２３日\n@場所：M023の自宅\n...

Data Cleaning and Reorganization

- Lots of trial and error - created toy dataframes to test before doing the work on the whole file
- Documentation removal
- Tokenized by new lines, but new lines didn't always mean a new speaker
- F100 wasn't even participating in anything… must've just been listening

In [53]:
```
byfile.head()
```

Out[53]:

	file	participants	content
0	data001.txt	[F107, F023, M023, F128]	[F107：＊＊＊の町というのはちいちゃくって、城壁がこう町全体をぐるっと回ってて、それが城...
1	data002.txt	[F107, F023, F128]	[F107：今度はーイギリスにもアメリカと同様のテロが起こるだろうって言ったんだってよ。, ...
2	data003.txt	[F033, F056]	[F033：倒れちゃう。, F056：いきなり倒れた。, F033：どうしよう。あっ、この間...
3	data004.txt	[M018, F128]	[F128：いや、別にいいよ。ローソンでいいやろ。ちょっと倒していい、これ。どうよ、調子は。...
4	data005.txt	[M023, F128, F116, M026]	[F128：来てたときによく貸してもらったやつだ。, M023：そう、そんな感じのとこ。, ...

In [54]:
```
byfile['content'][4][:5]
```

Out[54]:
```
['F128：来てたときによく貸してもらったやつだ。', 'M023：そう、そんな感じのとこ。', 'F128：わーい。サンキュー。ちょっと待って。', 'M023：会話って、何を会話するや。', 'F128：いや、別に。ていうか早く決めよう。あんね、まず、あの、１１月４日の話。']
```

Data Cleaning and Reorganization

- Eventually was able to filter dialogue into speaker entries in other dataframe!
    - Had to use a nested for-loop, unfortunately :(luckily didn’t take too long

```
In [69]: byparticipant.head()
```

Out[69]:

	participant	age	appears_count	appears_in	content
0	F001	Early 20s	5	[data105.txt, data086.txt, data076.txt, data07...	うーん、わかんない。そういうこと言わないで。うるさいな。うるさい。うるさいって言ってるの。う...
1	F002	Late 60's	3	[data033.txt, data032.txt, data031.txt]	２７歳から現在まで東京都に居住。南仏へいらしたそうだけど、（うん）どうでしたか。みんな、太っ...
2	F003	Late 80's	1	[data129.txt]	そうねえ。＜笑い＞先生はね、師範卒業したのがねえ、１９歳だったのよ。だーから、若い先生でね。...
3	F004	Late 20's	14	[data096.txt, data094.txt, data092.txt, data08...	うん、まあね。はい、もう始まってますからね。よろしくね。ちょっとちょっと、ちゃんとさ、つなぐ...
4	F005	Late 20's	3	[data052.txt, data023.txt, data015.txt]	はーい。いや、F034さんってー、やっぱりー、ハンバーガーとか好きですよねー。＊はなしを＊。...

Clean Data Deets

- List contained 5192 words
- 197 total participants
- 129 different conversations

```
In [6]: len(wordlist)
Out[6]: 5192
```

So there are around 5,000 entries in our list of Katakana words....

```
In [7]: len(bypar)
Out[7]: 197
```

... 197 participants...

```
In [8]: len(byfile)
Out[8]: 129
```

... who spoke in 129 files.

We aren't tooootally done yet, though...

Speeding up the process....

- No point in trying to find words that don't show up in the conversation, right?
- 2-in-1 process: removing words that don't appear in the Nagoya conversations while also gathering frequency of the words used to compare web usage to conversational usage
    - See how the highest-ranked word as far as web frequency goes isn't that occurrent in casual conversation at all?

```
In [17]: wordlist['conv_freq'] = ''
```

```
In [18]: for i in range(len(wordlist)):
             word =  wordlist['katakana'][i]
             ct = 0
             for j in range(len(bypar)):
                 if word in bypar['content'][j]:
                     ct += 1
             if ct == 0:
                 wordlist['conv_freq'][i] = None
             else:
                 wordlist['conv_freq'][i] = ct
```

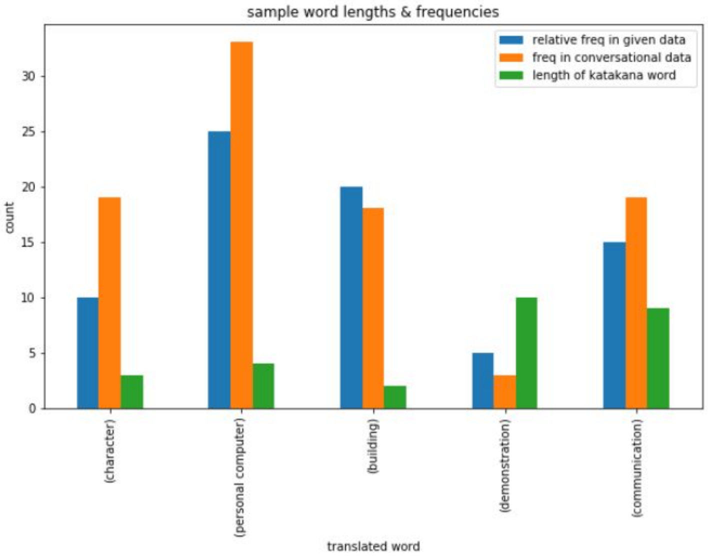
```
In [19]: wordlist.head(15)
```

Out[19]:

	katakana	translation	frequency	conv_freq
0	パーセント	percent	63392	3
1	アメリカ	America	28243	69
2	ページ	page	24642	27
3	センター	center	20664	33
4	サービス	service	16630	21
5	システム	system	16458	17
6	メートル	metre	15960	17
7	テレビ	television	15644	76
8	メール	mail	15589	72
9	データ	data	13210	28
10	フランス	France	10957	54
11	ポイント	point	10919	15
12	ホーム	home	10790	38
13	ホテル	hotel	10503	45
14	ブログ	blog	10205	None

A Peek at Length vs. Frequency

- Took a sample of two varieties of words:
    - Lengthier katakana words that did not shorten from the English version
        - デモンストレーション (demonsutoreeshon = demonstration)
        - コミュニケーション (komyunikeeshon = communication)
    - Shorter katakana words that did shorten from the English version
        - ビル (biru = building)
        - キャラ (kyara = character)
        - パソコン (pasukon = personal computer)
- Graphed their length in Katakana, conversational frequency and web frequency
    - Had to just rank them within those five words for web frequency - those numbers were too big to fit on the graph!



A Peek at Length vs. Frequency

- There’s much more to consider here; the word for communication was still more common than one of the shortened words in the conversational data!
    - Could be because of the conversation topics?
- “Demo” vs “demonstration”
    - “Demo” is more commonly used!

```
In [54]: wordlist['conv_freq'][pd.Index(wordlist['katakana']).get_loc('デモンストレーション')] # translation: demonstration

Out[54]: 3
```

```
In [53]: wordlist['conv_freq'][pd.Index(wordlist['katakana']).get_loc('デモ')] # translation: demo

Out[53]: 9
```

Since we have to take double counting into account, the actual conversational frequency of デモ is 6. That is still twice the amount of uses as the longer word for "demonstration."

Some potential issues...

- Double counting!
    - No word boundaries in Japanese, so if one word appears within another, it'll count that instance of a word twice
    - This occurred with ラ = "ra"/"la" - look at all the words it double counted for ラ!
        - This is only about half of them...

```
In [26]: for i in range(len(wordlist)):
             if 'ラ' in wordlist['katakana'][i]:
                 print(wordlist['katakana'][i], end = ', ')
```

```
フランス，クラブ，ライン，カメラ，バランス，クラス，プログラム，ガラス，ドラマ，カラー，ブランド，ボランティア，プ
ラス，ラジオ，レストラン，トラブル，ライブ，ドライブ，プラン，トラック，ラーメン，サラダ，イスラム，イラク，グラ
ス，ラブ，オランダ，グラフ，ラン，オーストラリア，ランキング，ドライバー，ブラック，サラリーマン，ライフ，ランド，
ブラジル，ランチ，キャラクター，ライト，キャラ，イスラエル，イラスト，プラスチック，オペラ，イラン，アラブ，ラッキ
ー，ランプ，リラックス，グランド，コラム，ラベル，ライバル，ブラシ，ランク，ラウンド，ライター，ライオン，クラシッ
ク，グラウンド，ライト，ラスト，プライド，プライバシー，エラー，ベランダ，カメラマン，フライ，ドラゴン，リストラ，
ウラン，マラソン，プライベート，ブラウン，ベテラン，ラ，ライダー，ラップ，ラテン，アラビア，コーラ，ミネラル，ドラ
イ，ポーランド，マフラー，オーケストラ，ドラム，ライス，ピラミッド，フランク，ディーラー，クーラー，フォーラム，フ
ラワー，フラッシュ，レギュラー，テラス，ラリー，ニュージーランド，プラント，ライセンス，ブラウス，サングラス，トラ
ンク，ライフスタイル，カウンセラー，セラー，ラッシュ，トライ，スライス，ストラップ，ミラノ，プラグ，モラル，オー
ラ，スコットランド，ミラー，キャラメル，ライト，フィンランド，ラグビー，カリキュラム，ランニング，マスカラ，フライ
```

Some potential issues...

- Some participants participated way more in the conversations than others
    - ... Way more. One participant participated 14 times
    - That participant had a ton of data, so it's possible that she'll skew everything completely
- Uneven age distribution
    - More younger speakers than older
    - Younger speakers tended to participate in more discussions than older speakers

```
In [9]: bypar['appears_count'].value_counts()

Out[9]: 1      161
        2       16
        3        7
        5        4
        4        4
        7        2
        14       1
        11       1
        6        1
        Name: appears_count, dtype: int64
```

Where to go from here

- I still need to work on age vs. katakana use, but I might have to try a couple different methods that each have their own pros and cons
    - Taking ratio of katakana characters to total characters
        - Pro: avoids outliers
        - Con: could also count onomatopoeia words - not what we're looking for!
    - Counting up how many katakana words each participant uses
        - Pro: sticks to the list of words, excludes onomatopoeia
        - Con: prone to outliers and double counting
- Maybe test out some machine learning methods, if I find a correlation and don't run out of time
    - Multinomial NB? SVC? Gridsearch? We'll see

ありがとうございます！

arigatou gozaimasu!

Be safe, everyone!