Data Science in ESL

Using EFL blog writing patterns to predict L1 backgrounds

Acronyms

EFL (English as a foreign language)

ESL (English as a second language)

L1 (First language)

L2 (second language)

Background

1. Observations
2. Hypothesis
3. Goals

Research

4. Data Source
5. Data Exploration
6. Language Samples
7. Models
8. Areas of Improvement

1. Background

Observations

- ESL students acquire different English skills at different speeds
- ESL acquire English quicker when there are similarities between their L1 and L2
- Students from the same backgrounds use similar English expressions

Hypothesis

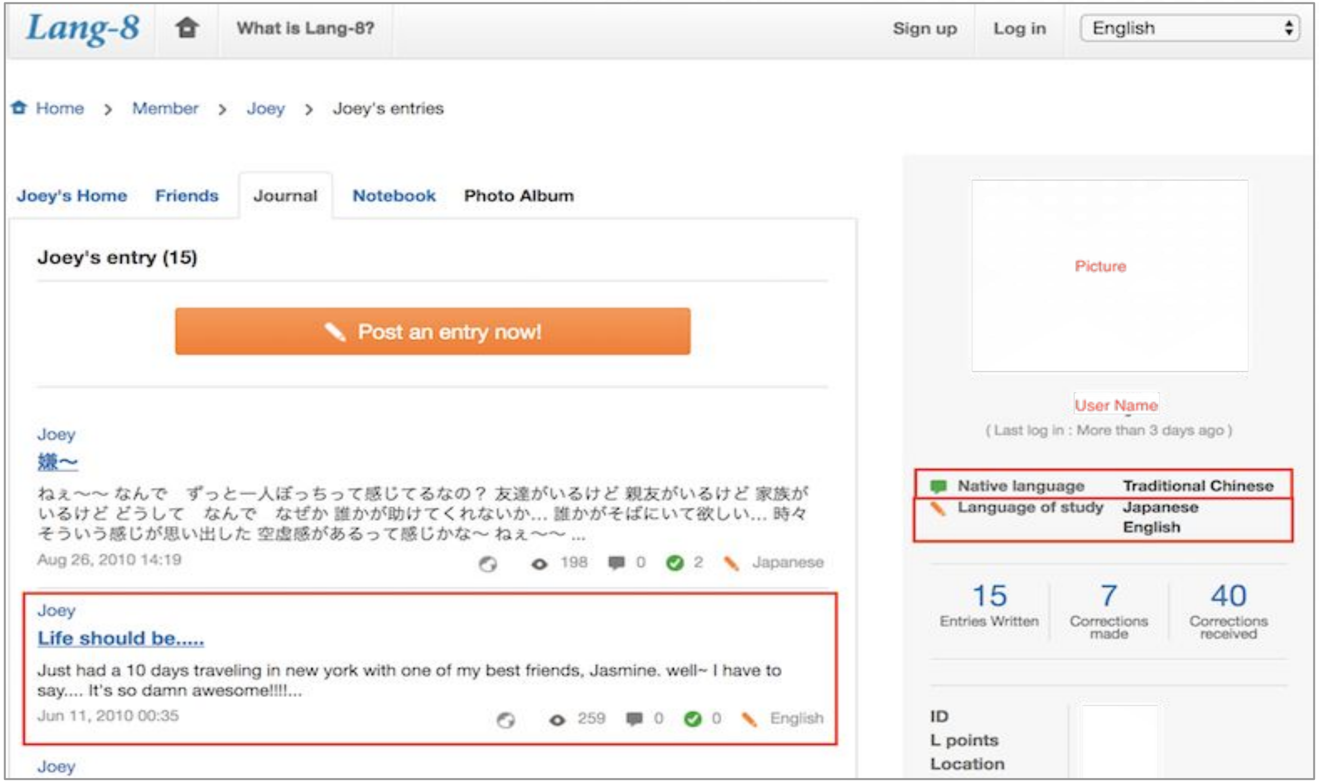
Students from other countries who are studying English will have unique writing patterns reflective of their L1.

Goals

- Use models to predict the L1 from blog writing style
- Use observations from data exploration to help ESL teachers develop material for students

2. Research

Data Source

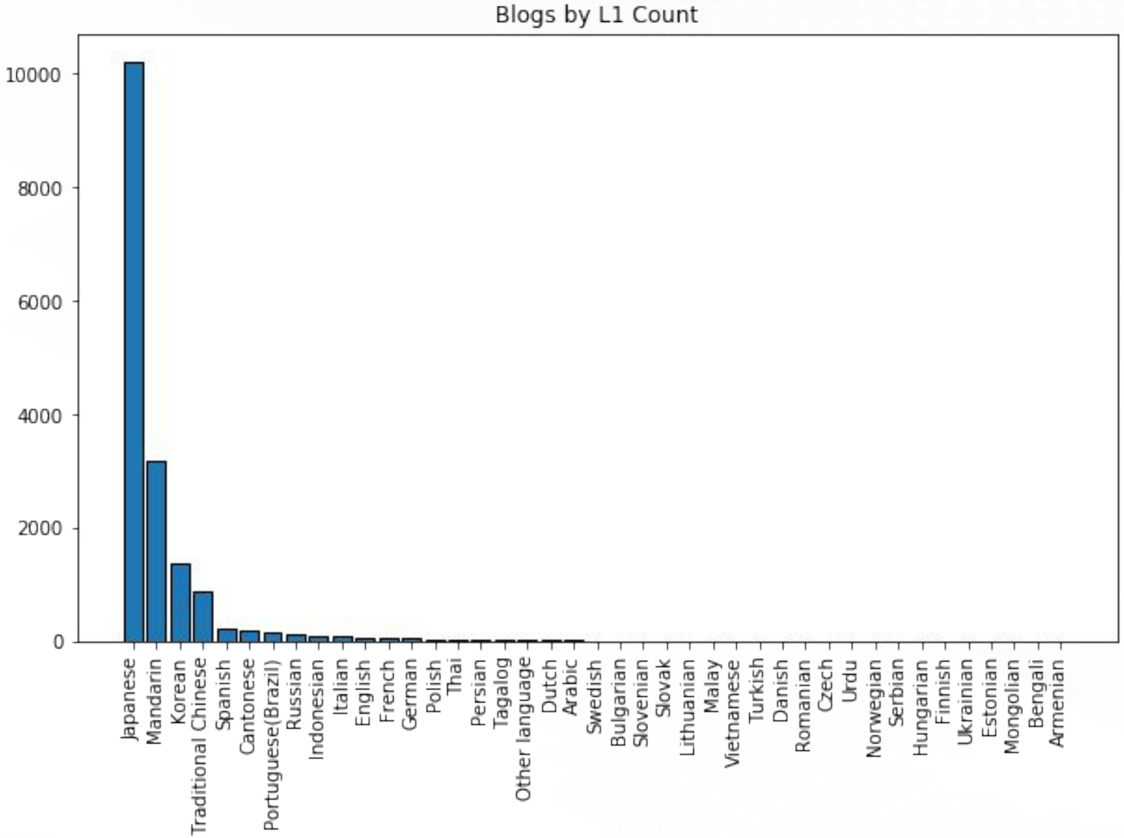


Lang-8
What is Lang-8?
Sign up
Log in
English
Home > Member > Joey > Joey's entries
Joey's Home
Friends
Journal
Notebook
Photo Album
Joey's entry (15)
Post an entry now!
Joey
嫌～
ねぇ～～ なんで　ずっと一人ぼっちって感じてるなの？ 友達がいるけど 親友がいるけど 家族がいるけど どうして　なんで　なぜか 誰かが助けてくれないか... 誰かがそばにいて欲しい... 時々そういう感じが思い出した 空虚感があるって感じかな～ ねぇ～～ ...
Aug 26, 2010 14:19
198
0
2
Japanese
Joey
Life should be.....
Just had a 10 days traveling in new york with one of my best friends, Jasmine. well~ I have to say.... It's so damn awesome!!!!...
Jun 11, 2010 00:35
259
0
0
English
Joey
Picture
User Name
(Last log in : More than 3 days ago)
Native language
Traditional Chinese
Language of study
Japanese
English
15
Entries Written
7
Corrections made
40
Corrections received
ID
L points
Location

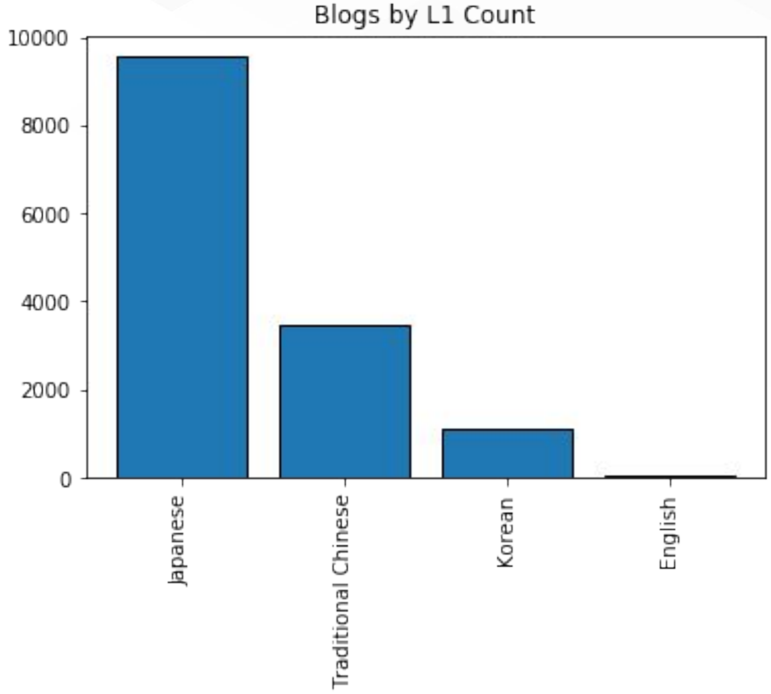
Scraped Data

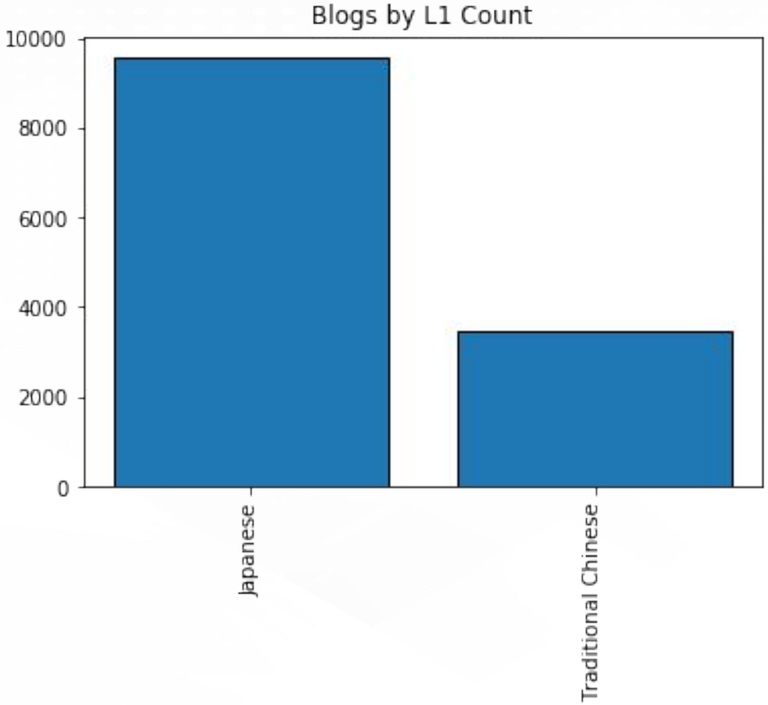
	id	time	title	content	language
15	3	2018-03-16 00:01:59.746874+00:00	I NEED SOMEBODY'S HELP ASAP!!	I want to apply for an internship but I don't...	Japanese
16	4	2018-03-16 00:05:54.394085+00:00	As they get older…	The other day, a man who is around 70 mention...	Japanese
17	4	2018-03-16 00:05:58.308874+00:00	I went to Nagoya port to welcome my foreign gu...	I went to Nagoya port to welcome my foreign g...	Japanese
18	5	2018-03-16 00:07:34.015592+00:00	英単語 (no need to correct)	sooner the better rugged bank on loyal fan ...	Japanese
19	5	2018-03-16 00:07:34.758156+00:00	英単語 (no need to correct)	millage TaxPropertyTax rate levy: an act of l...	Japanese

Language samples



Reduced samples





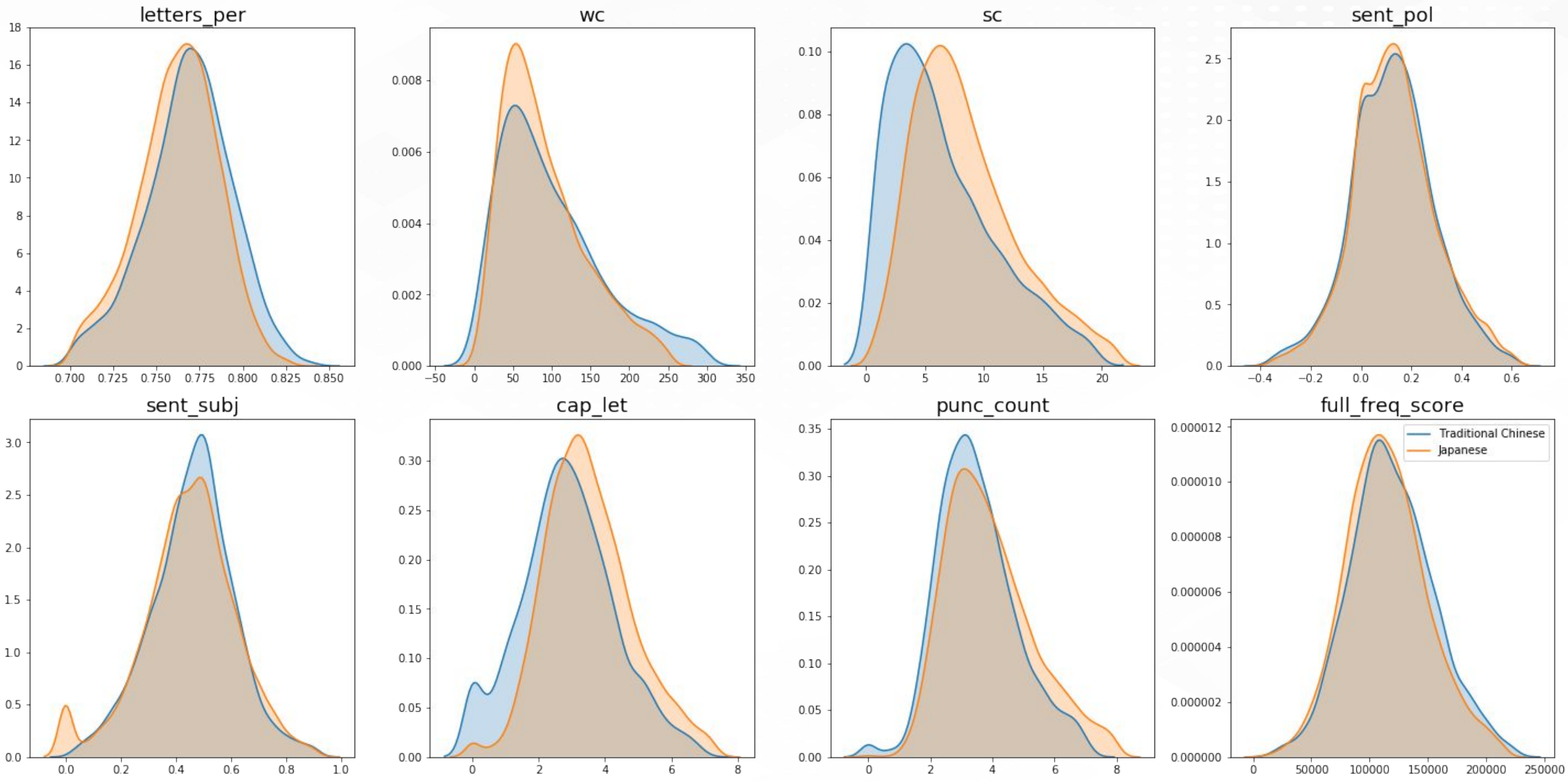
Continuous Features

Word Count
Sentence Count
Punctuation %
Capital Letter %
Frequency Score
Subjectivity (textblob)
Polarity (textblob)

Japanese vs Chinese

p-value: 3.851288704175285e-25

KDeplot (Histograms) of Continuous Features

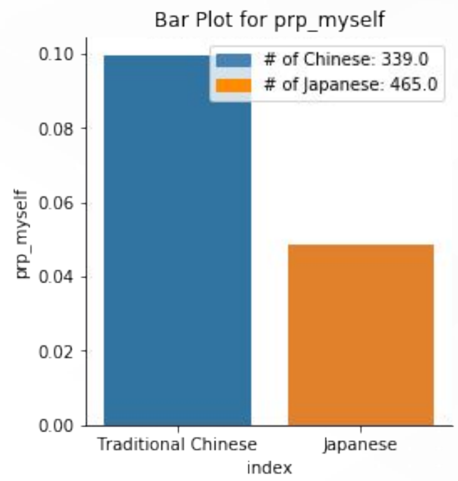


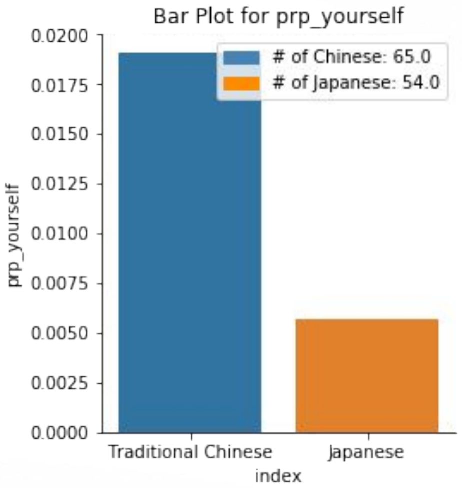
p-value:3.2764552529309356e-38

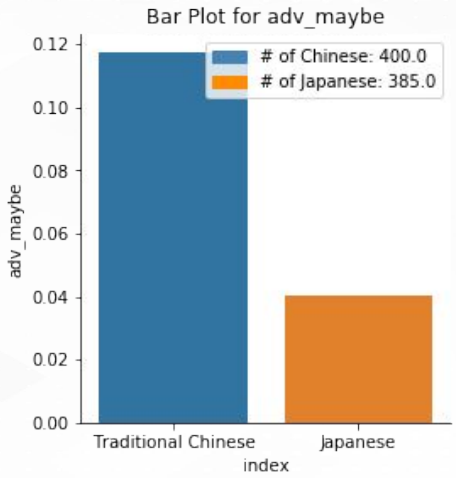
p-value:1.401298464324817e-45

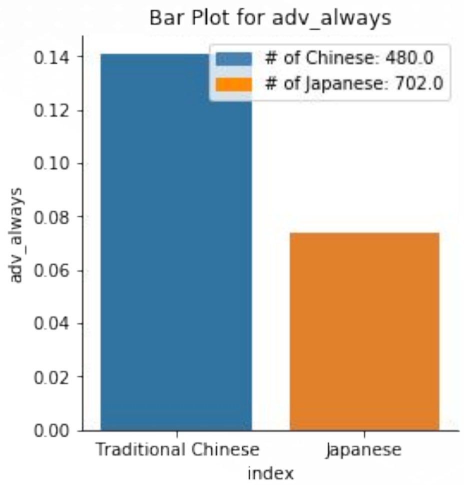
Discrete Features [~15000]

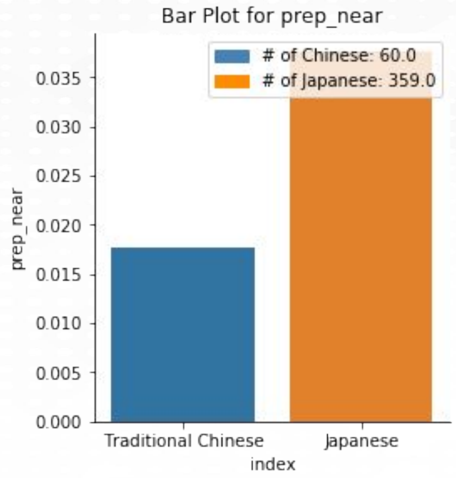
Adverbs **[50]**
Prepositions **[83]**
Pronouns **[29]**
POS (unigram) **[36]**
POS (bigrams) **[1003]**
POS (trigrams) **[13266]**
Letters (unigram) **[26]**
Letters (bigrams) **[542]**

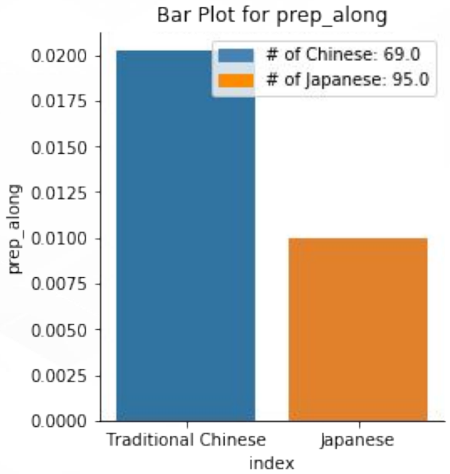

Bar Plot for prp_myself
# of Chinese: 339.0
# of Japanese: 465.0
prp_myself
0.10
0.08
0.06
0.04
0.02
0.00
Traditional Chinese
Japanese
index


Bar Plot for prp_yourself
# of Chinese: 65.0
# of Japanese: 54.0
prp_yourself
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Traditional Chinese
Japanese
index


Bar Plot for adv_maybe
# of Chinese: 400.0
# of Japanese: 385.0
adv_maybe
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Traditional Chinese
Japanese
index


Bar Plot for adv_always
# of Chinese: 480.0
# of Japanese: 702.0
adv_always
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Traditional Chinese
Japanese
index


Bar Plot for prep_near
# of Chinese: 60.0
# of Japanese: 359.0
prep_near
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Traditional Chinese
Japanese
index


Bar Plot for prep_along
# of Chinese: 69.0
# of Japanese: 95.0
prep_along
0.0200
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
Traditional Chinese
Japanese
index

Models

Basic scores for all features with an imbalanced train/test set

model	cpu time	overall score
Logistic Regression	1.74 s	0.730389
K Nearest Neighbors	3.4 s	0.792239
Naive Bayes Bernoulli	168 ms	0.702347
Decision Tree	79.7 ms	0.742943
Random Forest	13.4 s	0.80828

All Features

No Transformation

	name	time	total	prec: | JA | CH |	rec: | JA | CH |	f1: | JA | CH |
0	Logistic Regression	0.19	0.631667	[0.89, 0.58]	[0.30, 0.96]	[0.45, 0.72]
1	K Nearest Neighbor	8.06	0.500000	[0.50, 0.50]	[0.02, 0.98]	[0.04, 0.66]
2	Naive Bayes - Bernoulli	0.89	0.635000	[0.71, 0.60]	[0.45, 0.82]	[0.55, 0.69]
3	Decision Tree	0.17	0.636667	[0.87, 0.58]	[0.32, 0.95]	[0.47, 0.72]
4	Random Forest	0.21	0.505000	[1.00, 0.50]	[0.01, 1.00]	[0.02, 0.67]

Truncated SVD Transformation

	name	time	total	prec: | JA | CH |	rec: | JA | CH |	f1: | JA | CH |
0	Logistic Regression	0.01	0.540000	[0.77, 0.52]	[0.11, 0.97]	[0.20, 0.68]
1	K Nearest Neighbor	0.05	0.505000	[0.64, 0.50]	[0.02, 0.99]	[0.05, 0.67]
2	Naive Bayes - Bernoulli	0.01	0.530000	[0.78, 0.52]	[0.08, 0.98]	[0.15, 0.68]
3	Decision Tree	0.01	0.551667	[0.72, 0.53]	[0.17, 0.94]	[0.27, 0.68]
4	Random Forest	0.07	0.511667	[0.82, 0.51]	[0.03, 0.99]	[0.06, 0.67]

Selected Features

No Transformation

	name	time	total	prec: | JA | CH |	rec: | JA | CH |	f1: | JA | CH |
0	Logistic Regression	0.05	0.501667	[1.00, 0.50]	[0.00, 1.00]	[0.01, 0.67]
1	K Nearest Neighbor	1.66	0.505000	[0.67, 0.50]	[0.02, 0.99]	[0.04, 0.67]
2	Naive Bayes - Bernoulli	0.15	0.703333	[0.84, 0.65]	[0.51, 0.90]	[0.63, 0.75]
3	Decision Tree	0.04	0.608333	[0.87, 0.56]	[0.25, 0.96]	[0.39, 0.71]

Truncated SVD Transformation

	name	time	total	prec: | JA | CH |	rec: | JA | CH |	f1: | JA | CH |
0	Logistic Regression	0.01	0.501667	[1.00, 0.50]	[0.00, 1.00]	[0.01, 0.67]
1	K Nearest Neighbor	0.03	0.505000	[0.67, 0.50]	[0.02, 0.99]	[0.04, 0.67]
2	Naive Bayes - Bernoulli	0.01	0.648333	[0.79, 0.60]	[0.40, 0.89]	[0.53, 0.72]
3	Decision Tree	0.01	0.696667	[0.85, 0.64]	[0.48, 0.92]	[0.61, 0.75]
4	Random Forest	0.06	0.640000	[0.90, 0.59]	[0.32, 0.96]	[0.47, 0.73]

3. Next Steps

Plan to address

- A control variable (English samples)
- Sample distributions
- Deeper look at POS and syntax structure
- Improved feature selection through PCA/SVD and models’ ranked features
- More data cleaning - blog posts with multiple languages

Language Samples

- Get more Korean samples
- Find more english blog posts to establish control variable
- Work with samples from L1s with more linguistic diversity

More Features?

- Cognates
- Word stemming
- Noun categories
- Syntax structures:
    - imperatives
    - interrogatives
    - passive vs active sentences

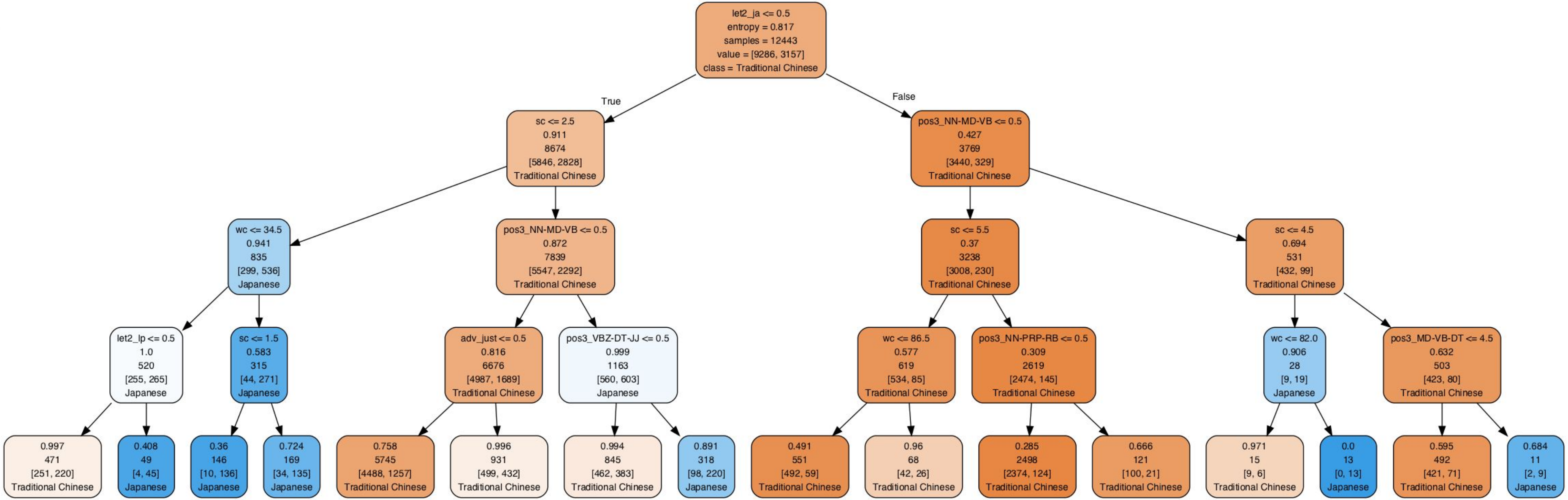
For ESL

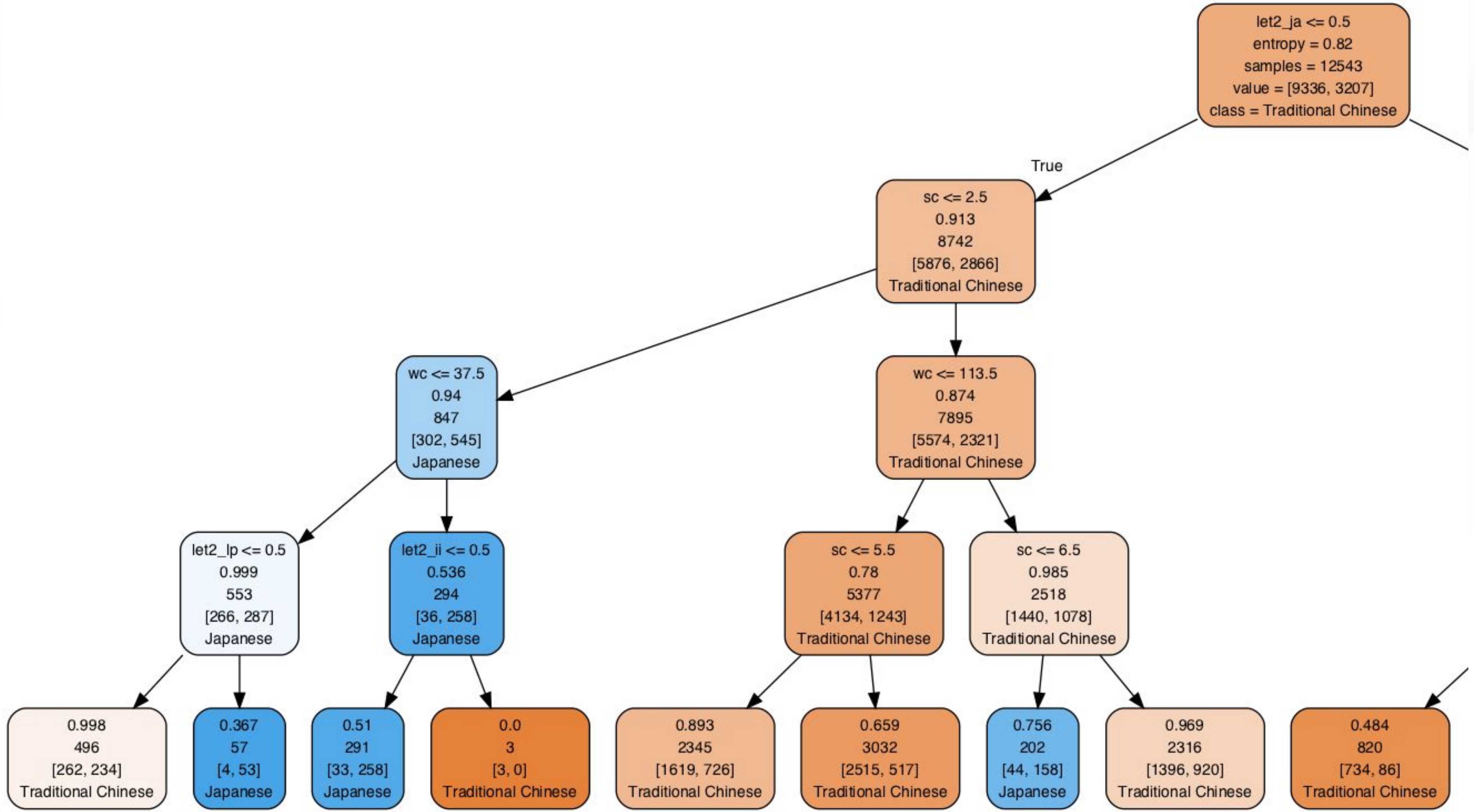
- Add a lesson on the use of reflexive pronouns in English for Japanese students

- Add a lesson about hedging for Japanese students

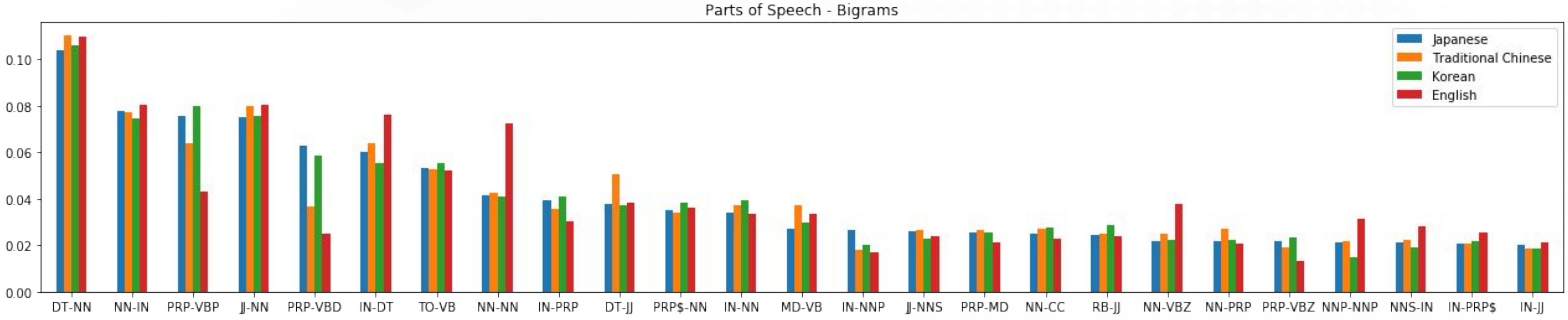
4. Extra Visuals

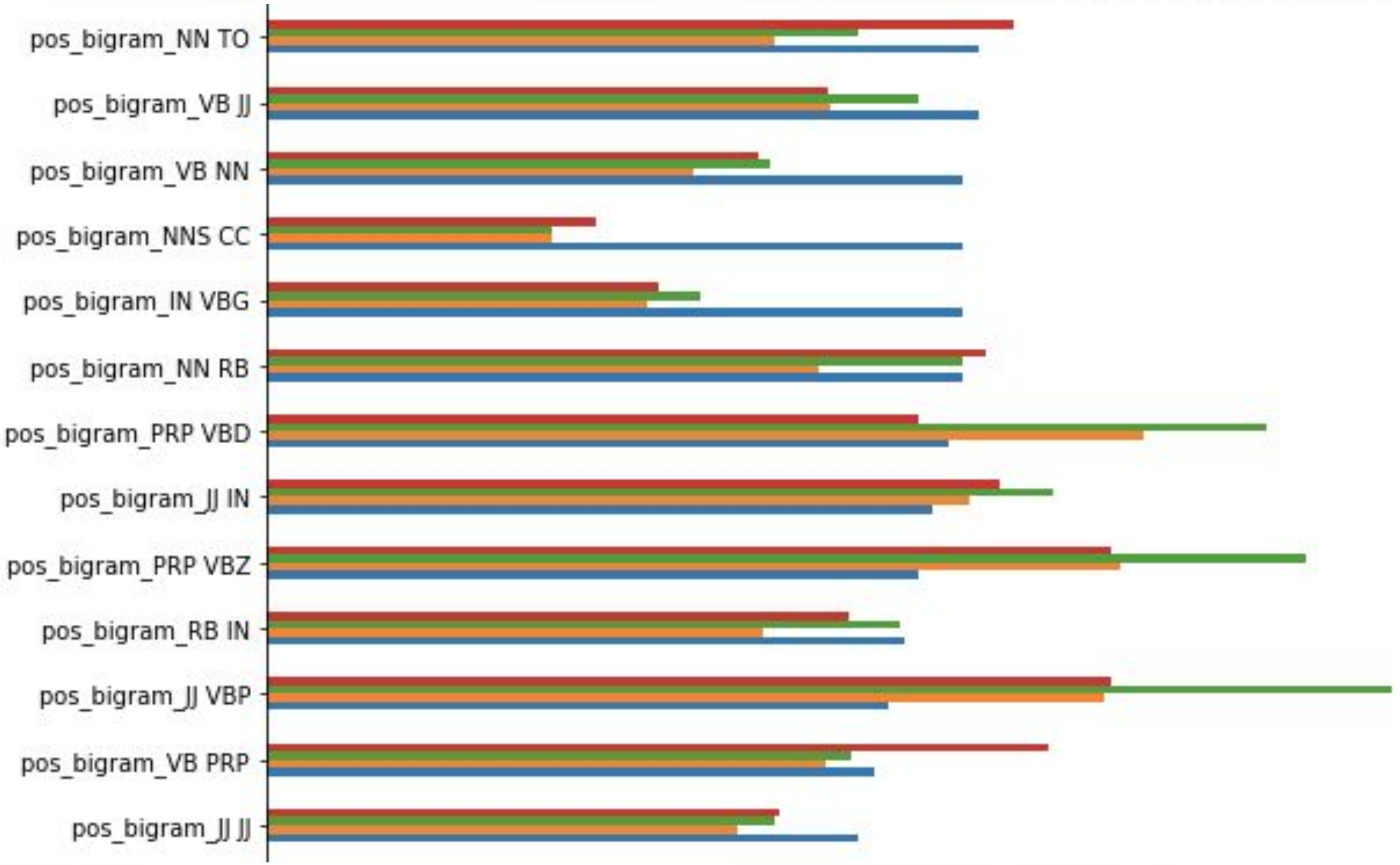
Decision Tree

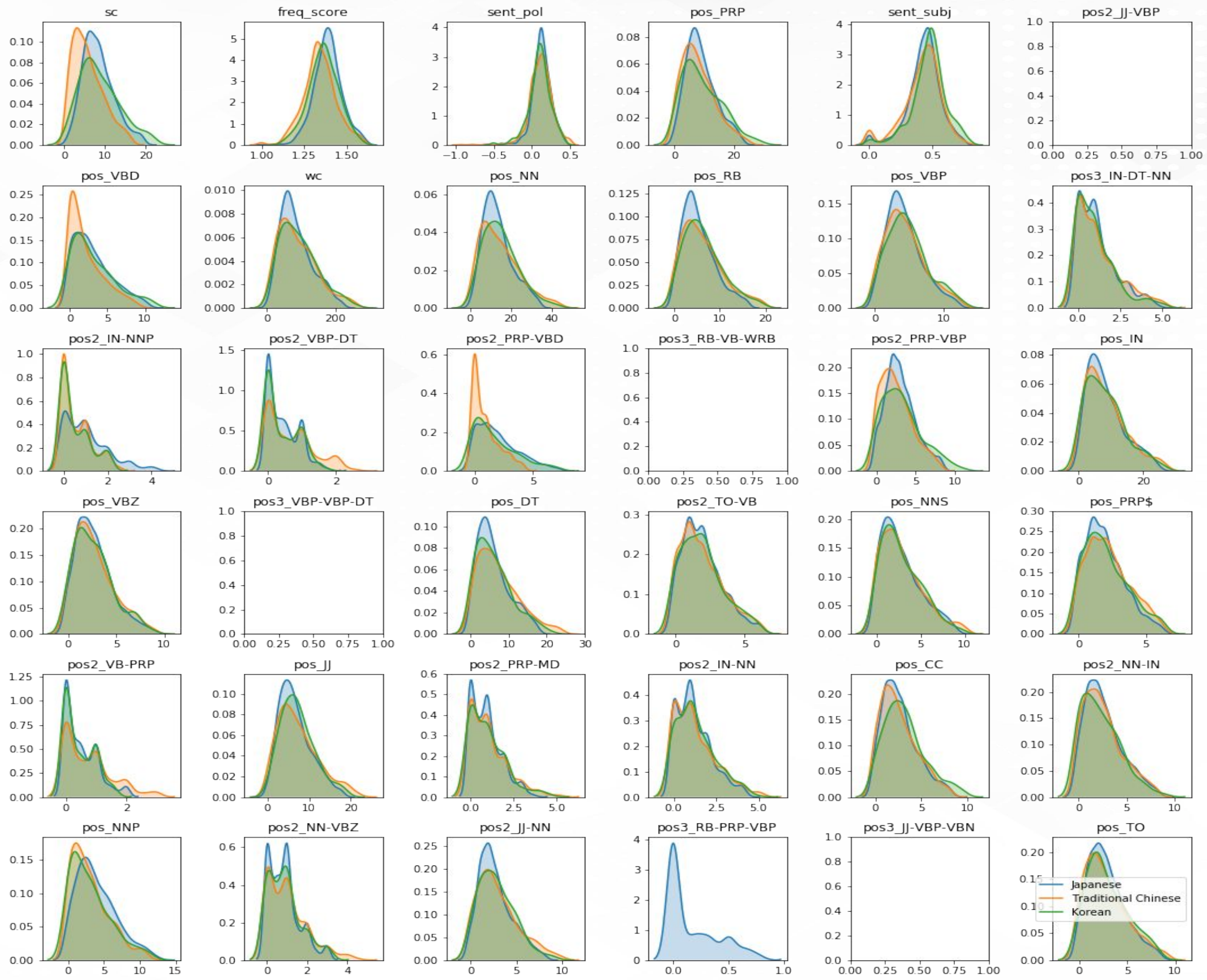




let2_ja <= 0.5
entropy = 0.82
samples = 12543
value = [9336, 3207]
class = Traditional Chinese
True
sc <= 2.5
0.913
8742
[5876, 2866]
Traditional Chinese
wc <= 37.5
0.94
847
[302, 545]
Japanese
wc <= 113.5
0.874
7895
[5574, 2321]
Traditional Chinese
let2_lp <= 0.5
0.999
553
[266, 287]
Japanese
let2_ii <= 0.5
0.536
294
[36, 258]
Japanese
sc <= 5.5
0.78
5377
[4134, 1243]
Traditional Chinese
sc <= 6.5
0.985
2518
[1440, 1078]
Traditional Chinese
0.998
496
[262, 234]
Traditional Chinese
0.367
57
[4, 53]
Japanese
0.51
291
[33, 258]
Japanese
0.0
3
[3, 0]
Traditional Chinese
0.893
2345
[1619, 726]
Traditional Chinese
0.659
3032
[2515, 517]
Traditional Chinese
0.756
202
[44, 158]
Japanese
0.969
2316
[1396, 920]
Traditional Chinese
0.484
820
[734, 86]
Traditional Chinese


Parts of Speech - Bigrams
Japanese
Traditional Chinese
Korean
English
0.00
0.02
0.04
0.06
0.08
0.10
DT-NN
NN-IN
PRP-VBP
JJ-NN
PRP-VBD
IN-DT
TO-VB
NN-NN
IN-PRP
DT-JJ
PRP$-NN
IN-NN
MD-VB
IN-NNP
JJ-NNS
PRP-MD
NN-CC
RB-JJ
NN-VBZ
NN-PRP
PRP-VBZ
NNP-NNP
NNS-IN
IN-PRP$
IN-JJ


pos_bigram_NN TO
pos_bigram_VB JJ
pos_bigram_VB NN
pos_bigram_NNS CC
pos_bigram_IN VBG
pos_bigram_NN RB
pos_bigram_PRP VBD
pos_bigram_JJ IN
pos_bigram_PRP VBZ
pos_bigram_RB IN
pos_bigram_JJ VBP
pos_bigram_VB PRP
pos_bigram_JJ JJ


sc
freq_score
sent_pol
pos_PRP
sent_subj
pos2_JJ-VBP
pos_VBD
wc
pos_NN
pos_RB
pos_VBP
pos3_IN-DT-NN
pos2_IN-NNP
pos2_VBP-DT
pos2_PRP-VBD
pos3_RB-VB-WRB
pos2_PRP-VBP
pos_IN
pos_VBZ
pos3_VBP-VBP-DT
pos_DT
pos2_TO-VB
pos_NNS
pos_PRP$
pos2_VB-PRP
pos_JJ
pos2_PRP-MD
pos2_IN-NN
pos_CC
pos2_NN-IN
pos_NNP
pos2_NN-VBZ
pos2_JJ-NN
pos3_RB-PRP-VBP
pos3_JJ-VBP-VBN
pos_TO
Japanese
Traditional Chinese
Korean

Chinese vs Japanese

MannWhitney U Test (assumes normality)

sc:stat=536536.5, p=1.7072488700922754e-63
freq_score:stat=636197.5, p=4.515381615578102e-44
sent_pol:stat=886609.0, p=0.013264065652633338
pos_PRP:stat=770243.0, p=5.196098625805676e-08
sent_subj:stat=898769.5, p=0.041557713279235915
pos2_JJ-VBP:stat=0.0, p=0.0
pos_VBD:stat=655521.0, p=2.716598211138757e-24
wc:stat=843355.5, p=0.26494148060009115
pos_NN:stat=837920.5, p=0.1224638429596514
pos_RB:stat=816411.0, p=9.39059114470553e-05
pos_VBP:stat=868216.0, p=0.31286403071741714
pos3_IN-DT-NN:stat=880339.5, p=0.351431874684187
pos2_IN-NNP:stat=646817.5, p=1.9141116383595614e-31
pos2_VBP-DT:stat=753424.5, p=6.693777692873975e-11
pos2_PRP-VBD:stat=535814.5, p=5.280333996839972e-59
pos3_RB-VB-WRB:stat=0.0, p=0.0
pos2_PRP-VBP:stat=746535.0, p=1.7646331654577458e-12

pos_IN:stat=865240.5, p=0.4529202098291991
pos_VBZ:stat=897202.5, p=0.4178459416236008
pos3_VBP-VBP-DT:stat=0.0, p=0.0
pos_DT:stat=759274.5, p=1.677129429910917e-07
pos2_TO-VB:stat=842474.5, p=0.1148711960934507
pos_NNS:stat=798576.5, p=0.2746068949796701
pos_PRP$:stat=780511.5, p=0.00014482613404109238
pos2_VB-PRP:stat=793547.0, p=7.078343140193685e-10
pos_JJ:stat=818322.5, p=0.0038570494555401655
pos2_PRP-MD:stat=909842.5, p=0.13433003038081376
pos2_IN-NN:stat=888407.0, p=0.4481120057426727
pos_CC:stat=828476.5, p=0.0681338959087248
pos2_NN-IN:stat=818886.0, p=0.06117635396174209
pos_NNP:stat=669732.0, p=8.695664716145719e-19
pos2_NN-VBZ:stat=865825.0, p=0.004159327861131947
pos2_JJ-NN:stat=847599.0, p=0.03543135644157996
pos3_RB-PRP-VBP:stat=0.0, p=0.0
pos3_JJ-VBP-VBN:stat=0.0, p=0.0
pos_TO:stat=832738.5, p=0.03411138379372657