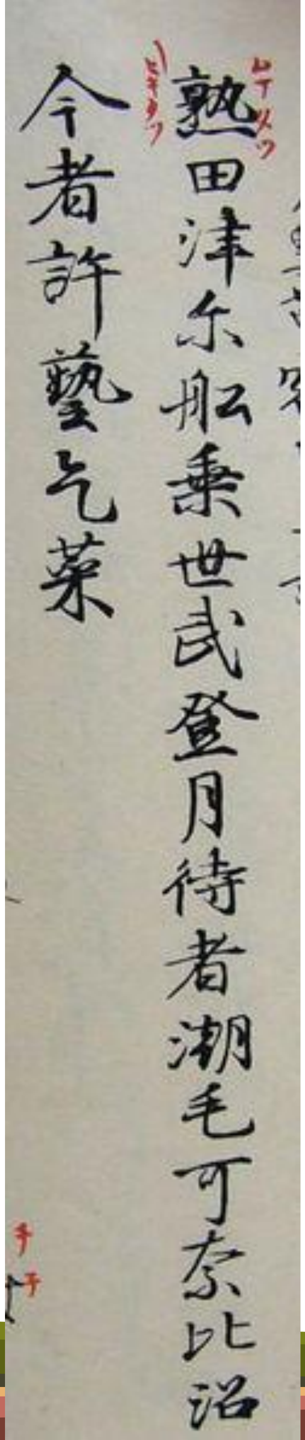
Manyoushuu (万葉集)

Jakub Levý
jakub.levy15@gmail.com

Faculty of Mathematics and Physics
Charles University



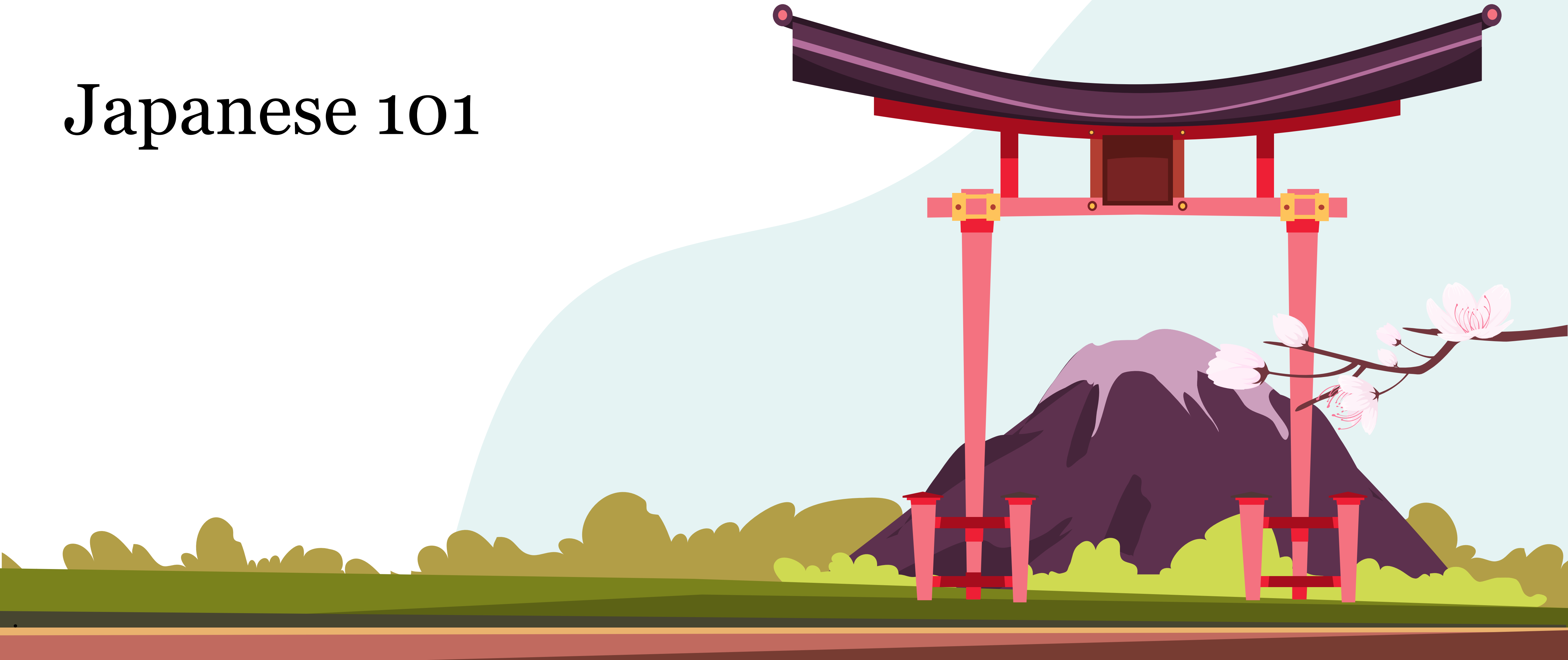
熟田津尓船乗世武登月待者潮毛可奈比沼
今者許藝乞菜

Content

- Japanese 101
- Writing System
  - Mora
  - Waka
  - Haiku
- Manyoushuu
  - About
  - Example in detail

- Project
  - Visualization Goals
  - Overview
  - Technologies
  - Look at the Code
  - (Live) Examples
  - Conclusion
- References

Japanese 101

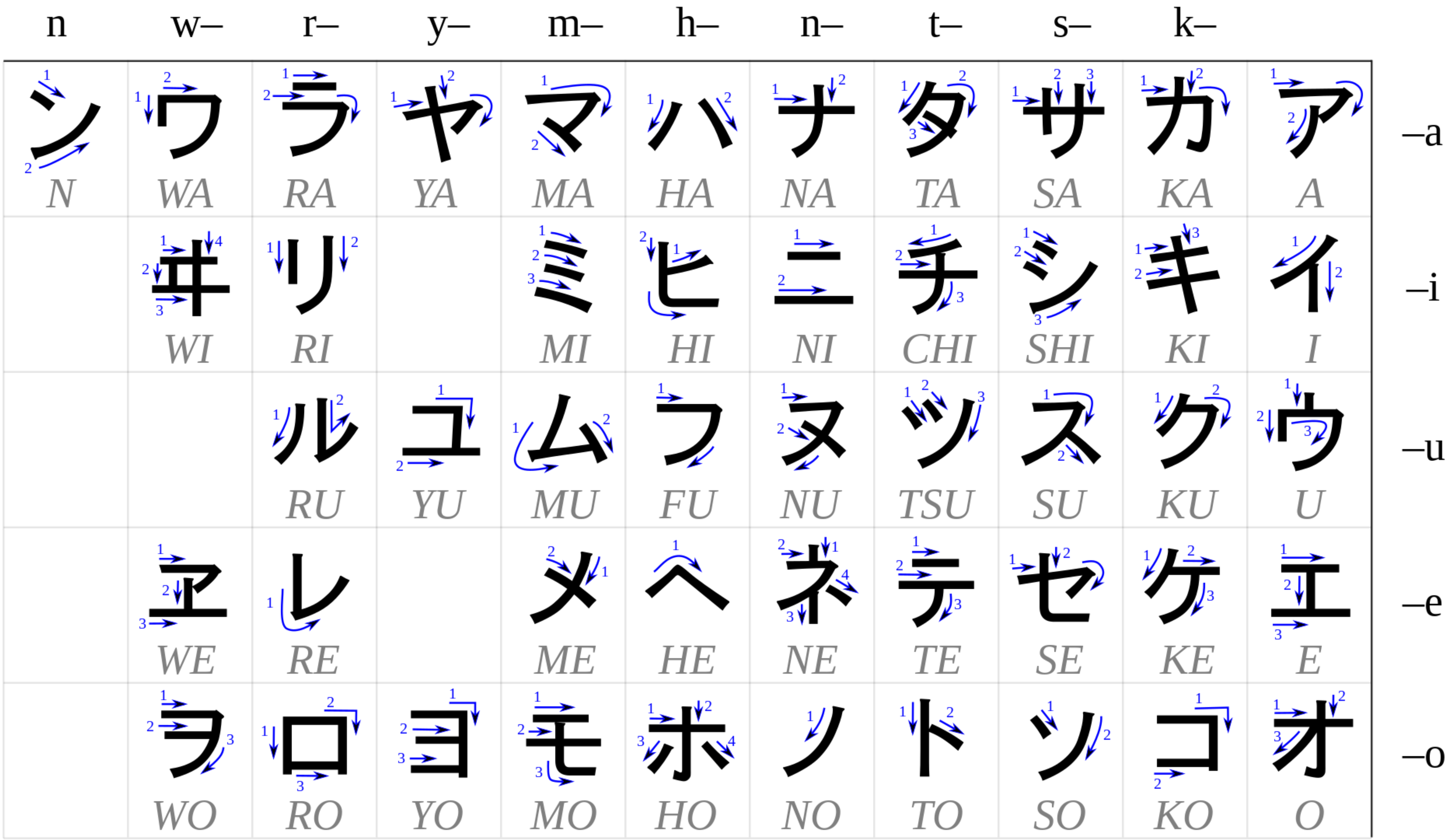


Japanese Writing System (1)

- No script existed until ~ AD 285
- Chinese symbols not perfect due to language differences
  - Sino-Japanese （音読） texts unintelligible
- Japanese reading of Chinese characters invented （訓読）
- How to capture the exact sound (of Japanese names)?
  - Use an arbitrary Chinese symbol with the correct 音読 reading
- Ex. *no* could be written as 能, 乃, 農, 濃, 迺, …
- The list was then reduced and partially defined
- *Manyougana* （万葉仮名）

Japanese Writing System (2)

- *Manyoushuu* is written in *Manyougana*
- Grammatical forms are captured by certain *Manyougana* characters
- Characters are complicated, but often used
  - Simplification
- In 10th century, shortcuts defined as *Katakana*（片仮名）
  - First Japanese syllabic alphabet
  - Considered less aesthetic than *Manyougana*



n	w–	r–	y–	m–	h–	n–	t–	s–	k–		
ン N	ワ WA	ラ RA	ヤ YA	マ MA	ハ HA	ナ NA	タ TA	サ SA	カ KA	ア A	–a
	ヰ WI	リ RI		ミ MI	ヒ HI	ニ NI	チ CHI	シ SHI	キ KI	イ I	–i
		ル RU	ユ YU	ム MU	フ FU	ヌ NU	ツ TSU	ス SU	ク KU	ウ U	–u
	ヱ WE	レ RE		メ ME	ヘ HE	ネ NE	テ TE	セ SE	ケ KE	エ E	–e
	ヲ WO	ロ RO	ヨ YO	モ MO	ホ HO	ノ NO	ト TO	ソ SO	コ KO	オ O	–o

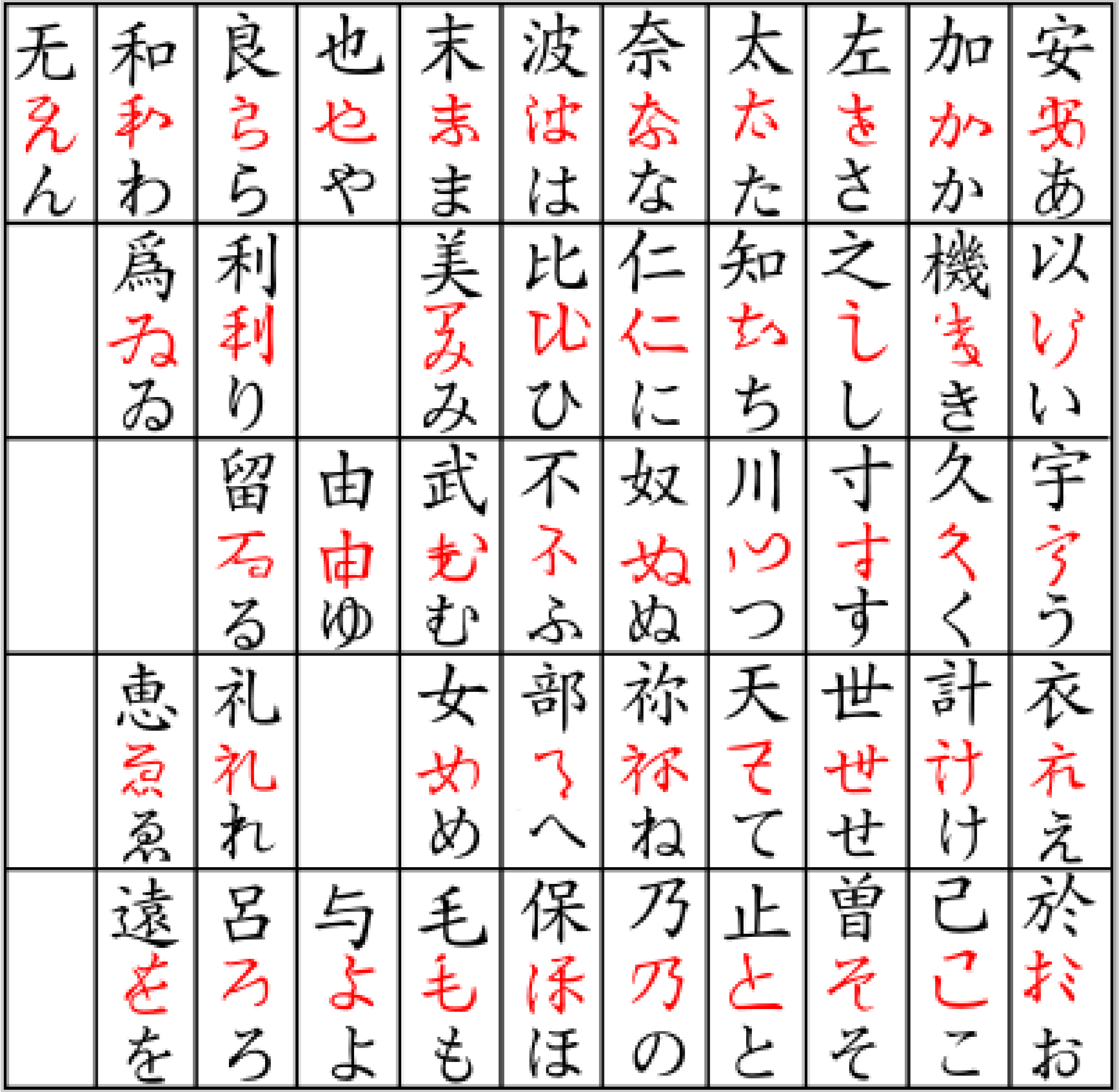
Japanese Writing System (3)

- In parallel, *Sougana* (草仮名) was used in private correspondences
  - Few selected *Manyougana* characters written in highly cursive style (草書体)
- Generally for woman
  - Not expected to know Chinese characters
- *Sougana* was called *onnade* (女手)

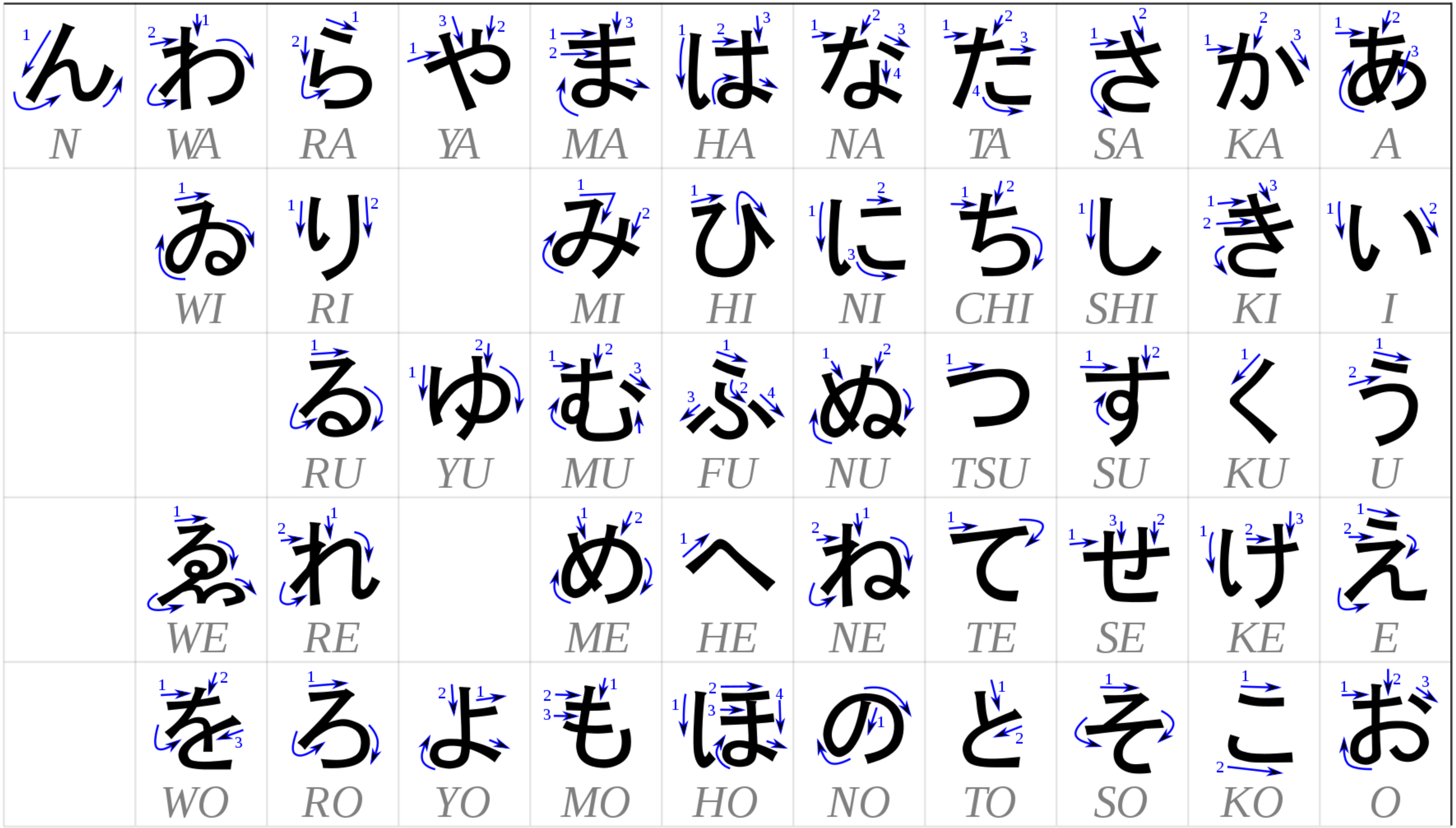
hand

woman

- Becomes popular in 10th century
- Contains “*woman gracefulness*”
- Slightly transformed and called *Hiragana* (平仮名) nowdays



无 ん	和 わ	良 ら	也 や	末 ま	波 は	奈 な	太 た	左 さ	加 か	安 あ
	爲 ゐ	利 り		美 み	比 ひ	仁 に	知 ち	之 し	機 き	以 い
		留 る	由 ゆ	武 む	不 ふ	奴 ぬ	川 つ	寸 す	久 く	宇 う
	恵 ゑ	礼 れ		女 め	部 へ	祢 ね	天 て	世 せ	計 け	衣 え
	遠 を	呂 ろ	与 よ	毛 も	保 ほ	乃 の	止 と	曽 そ	己 こ	於 お



n	w–	r–	y–	m–	h–	n–	t–	s–	k–		
ん *N*	わ *WA*	ら *RA*	や *YA*	ま *MA*	は *HA*	な *NA*	た *TA*	さ *SA*	か *KA*	あ *A*	–a
	ゐ *WI*	り *RI*		み *MI*	ひ *HI*	に *NI*	ち *CHI*	し *SHI*	き *KI*	い *I*	–i
		る *RU*	ゆ *YU*	む *MU*	ふ *FU*	ぬ *NU*	つ *TSU*	す *SU*	く *KU*	う *U*	–u
	ゑ *WE*	れ *RE*		め *ME*	へ *HE*	ね *NE*	て *TE*	せ *SE*	け *KE*	え *E*	–e
	を *WO*	ろ *RO*	よ *YO*	も *MO*	ほ *HO*	の *NO*	と *TO*	そ *SO*	こ *KO*	お *O*	–o

Japanese Writing System (4)

- Modern Japanese use
  - Hiragana, Katakana, Kanji (Chinese characters), Latin, Arabic numerals

SHII DII ICHI MAI TI SHA TSU NI MAI RI N GO SAN KO O KA I TA I TO OMO I MA SU

CD１枚、Tシャツ２枚、リンゴ３個を買いたいと思います。

I want to buy 1 CD, 2 T-shirts, and 3 apples so much.

- Notice SOV (subject-object-verb) word order

Mora

- One mora is the length of time it takes to pronounce one *kana* (Hiragana or Katakana) character

BA KA

馬鹿 ⟶ ばか

$k = m = s = 2$

KE K KO N

結婚 ⟶ けっこん

$k = m = 4, s = 2$

DA I GA KU

大学 ⟶ だいがく

$k = m = 4, s = 3$

TO U KYO U

東京 ⟶ とうきょう

$\underset{\boxed{\text{kana}}}{k} = 5, \underset{\boxed{\text{morae}}}{m} = 4, \underset{\boxed{\text{syllables}}}{s} = 2$

- Generally $k \neq m \neq s$

Waka (和歌 or 倭歌)

- Poetry in classical Japanese literature
- Many forms
    - Differ by the number of morae

- *Tanka* (短歌)
    - Major genre
    - 5-7-5-7-7

- *Katauta* (片歌)
    - Shortest form of *Waka*
    - 5-7-7

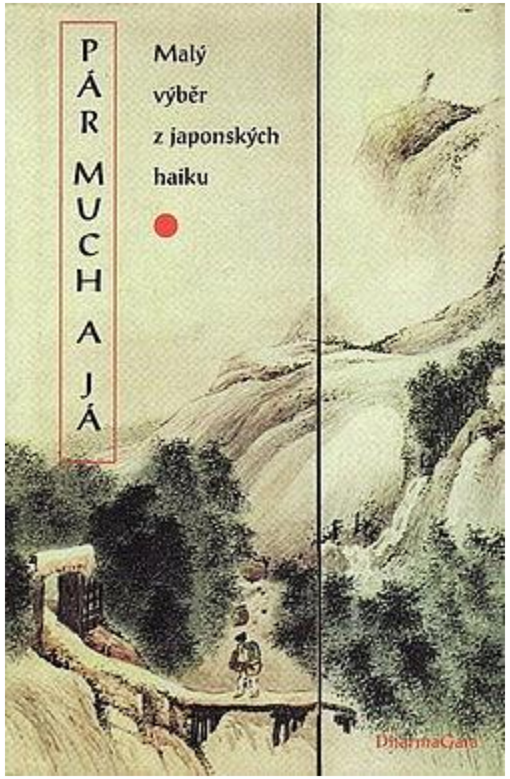
- *Sedouka* (旋頭歌)
    - *Katauta* twice
    - 5-7-7-5-7-7

- *Chouka* (長歌)
    - 5-7 repeated at least twice
    - Concluded with 5-7-7 ending

- *Bussokusekika* (仏足石歌)
    - *Tanka* with an extra 7 phrase at the end
    - 5-7-5-7-7-7

Why is Haiku (俳句) missing?

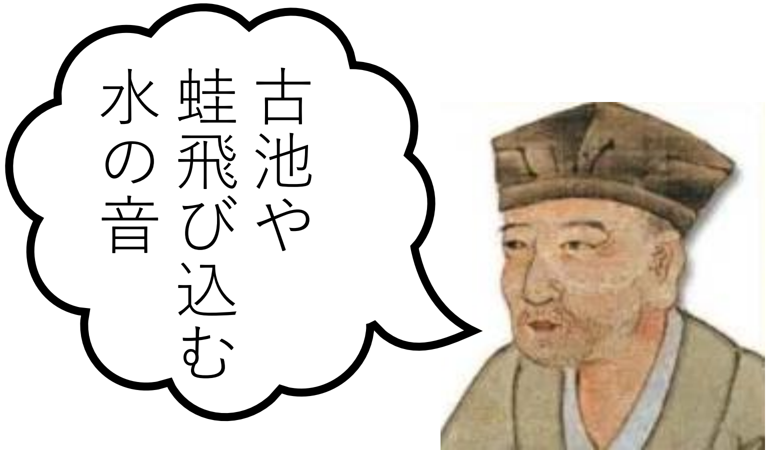
- First emerged in 17th century
- Much younger than all forms of *Waka*
- 5-7-5
- Not relevant for us

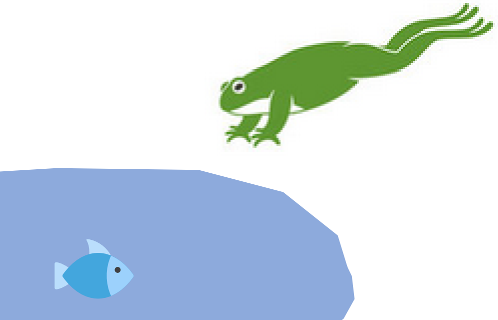


PÁR MUCHA A JÁ

Malý
výběr
z japonských
haiku

DharmaGaia





Manyoushuu （万葉集）

Manyoushuu (万葉集) (1)

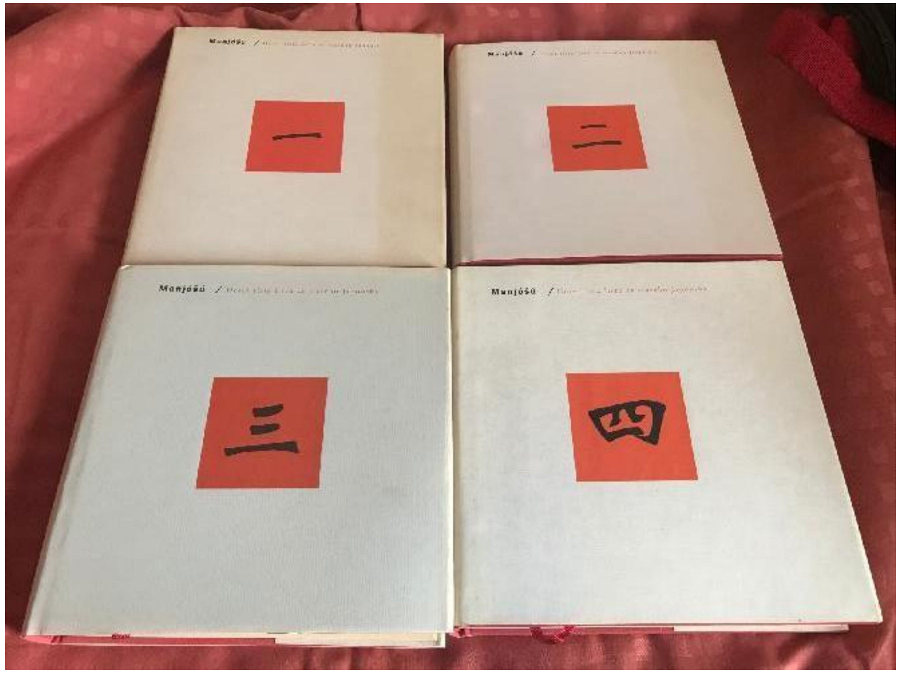
- 万葉集 ~ “collection of ten thousand leaves”
- Oldest collection of *Waka* poetry
- Written in *Manyougana*
- More than 4000 *Waka* poems
- 561 authors
  - 70 woman
  - People of various statuses: emperors, aristocrats, peasants, street performers, …
- Over 2100 poems by unknown authors

Manyoushuu（万葉集）(2)

- First poem by emperor *Nintoku* (313-399)
- Last poem from AD 759
- Most poems probably from 7th and 8th century
- Probably compiled by *Ootomo no Yakamochi*（大伴 家持）in 8th century

Manyoushuu（万葉集）(3)

- Complete translation of the whole anthology available only in
  - French
  - **Czech** (by Antonín Líman)



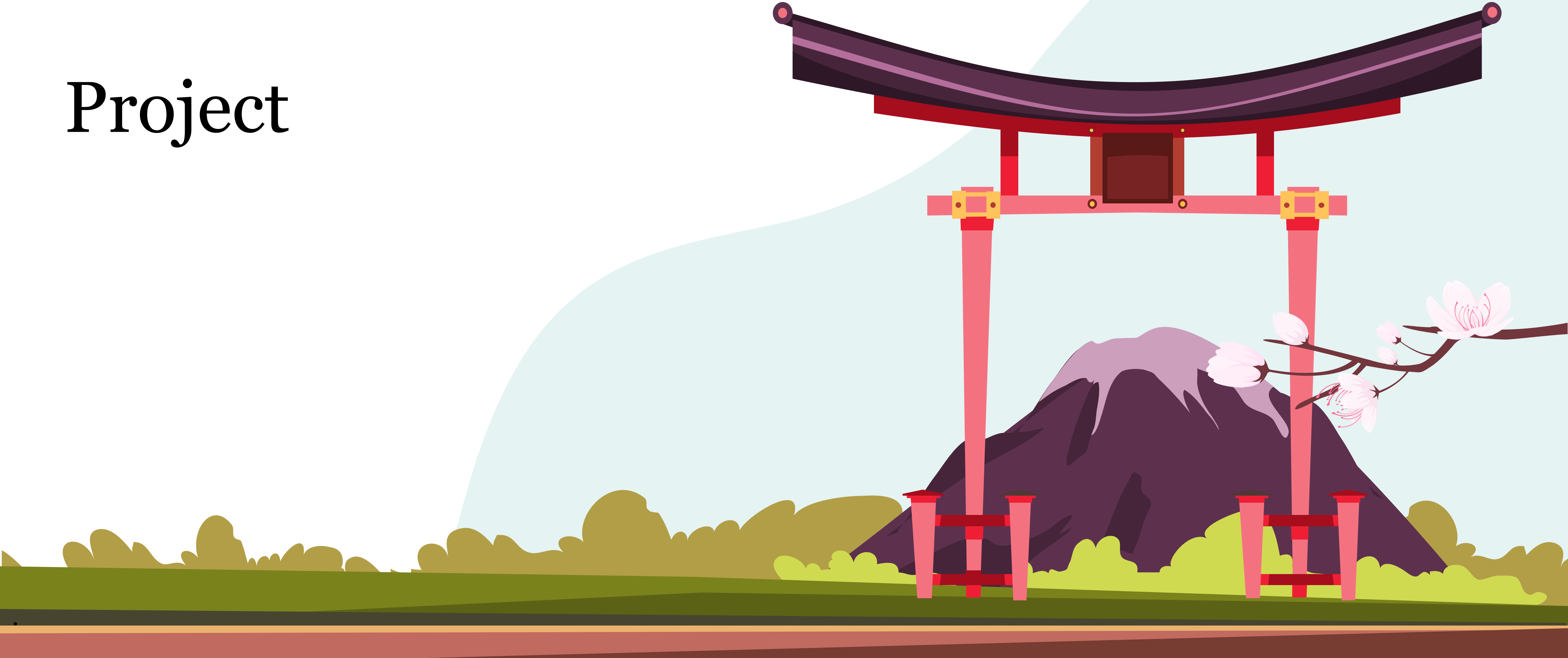
Uta 8 (Princess Nutaka’s Poem)

	Original	Modern	Hiragana only	
5	熟田津尓	NIGI TA TSU NI 熟田津に	NI GI TA TSU NI にぎたつに	5
7	船乗世武登	FUNA NO RI SE MU TO 船乗りせむと	FU NA NO RI SE MU TO ふなのりせむと	7
5	月待者	TSUKI MA TE BA 月待てば	TSU KI MA TE BA つきまてば	5
7	潮毛可奈比沼	SHIHO MO KA NA HI NU 潮もかなひぬ	SHI HO MO KA NA HI NU しほもかなひぬ	7
7	今者許藝乞菜	IMA WA KO GI I DE NA 今は漕ぎ出でな	I MA WA KO GI I DE NA いまはこぎいでな	**8**

U loďky v zálivu Nikita
čekáme na měsíc…
Příliv nám přeje,
nuže, k veslům, k veslům!

(Translated by Antonín Líman)

Project

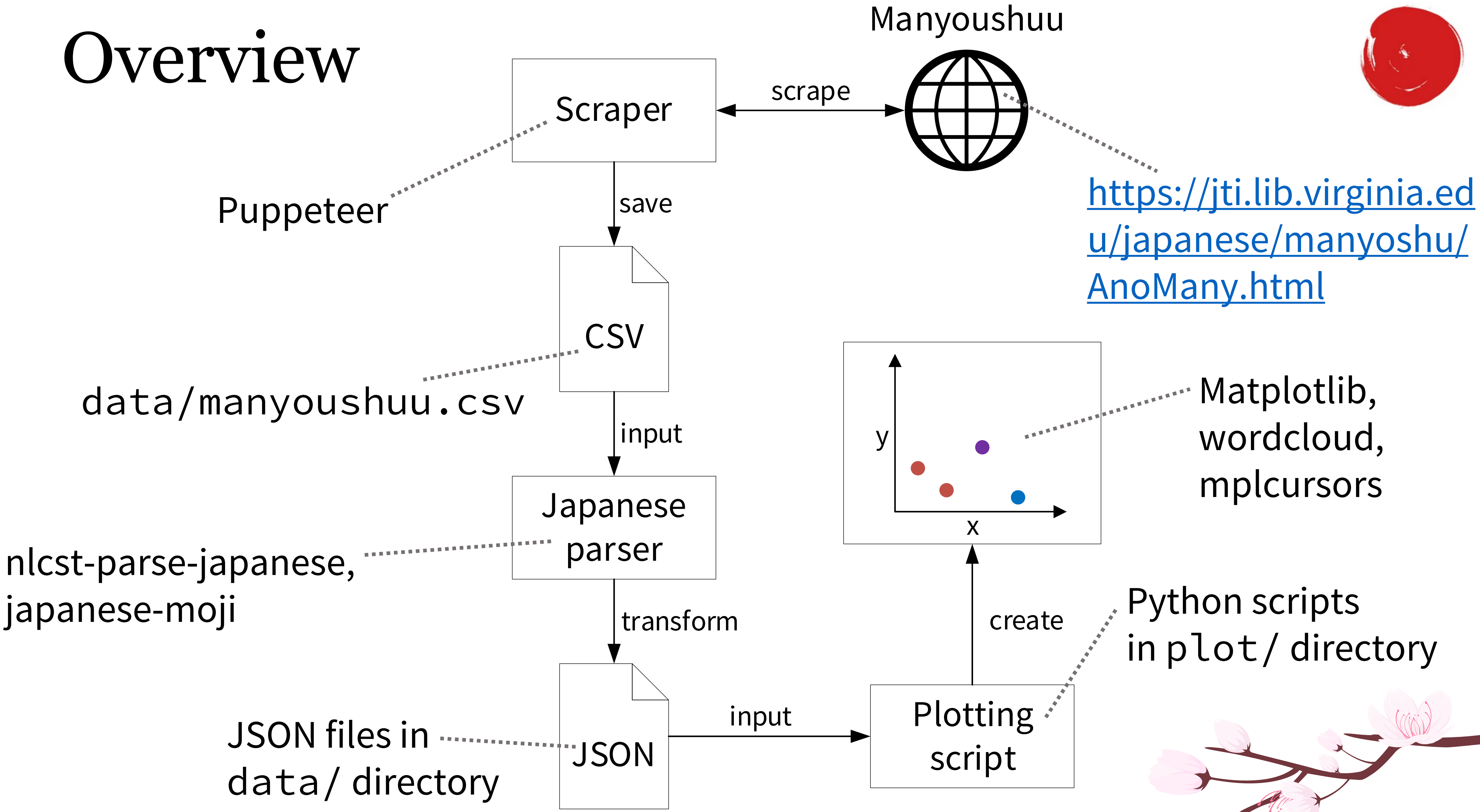


Visualization Goals

- Word cloud of the most used words
- Frequency of a *kanji* vs the number of different words the *kanji* is used in
  - And be able to somehow quickly see JLPT level of the *kanji*
- Calculate the number of poems for every *Waka* form
- Check the compliance of the form of a poem in modern Japanese
- Show the occurrence of a *kanji*/word across all poems

Different forms of a word must be handled correctly
持たず and 持ち are a form of 持つ

Overview



Manyoushuu

https://jti.lib.virginia.edu/japanese/manyoshu/AnoMany.html

Scraper

scrape

Puppeteer

save

CSV

`data/manyoushuu.csv`

input

Japanese parser

nlcst-parse-japanese, japanese-moji

transform

JSON

JSON files in `data/` directory

input

Plotting script

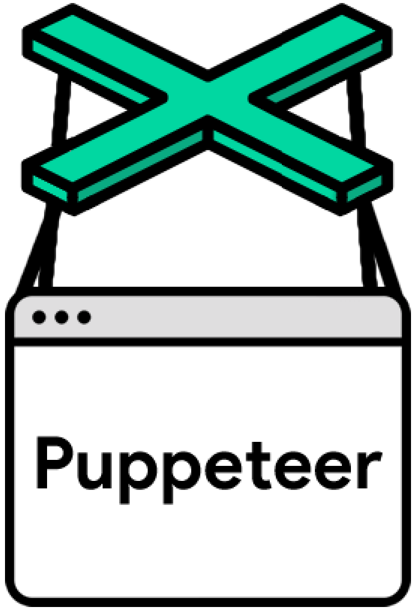
Python scripts in `plot/` directory

create

Matplotlib, wordcloud, mplcursors

Puppeteer

- Node.js library providing API to control Chrome/Chromium
- Uses DevTools Protocol



nlcst-parse-japanese (1)

- Parser for Japanese text
- Output format is *NLCST*
  - **N**atural **L**anguage **C**oncrete **S**yntax **T**ree format

nlcst

nlcst-parse-japanese (2)

MO CHI
Parser output for 持ち (boilerplate nodes omitted)

```
{
  "type": "TextNode",
  "value": "持ち",
  "position": {
    "start": {
      "line": 1,
      "column": 8,
      "offset": 7
    },
    "end": {
      "line": 1,
      "column": 10,
      "offset": 9
    }
  },
```

```
  "data": {
    "word_id": 3144310,
    "word_type": "KNOWN",
    "surface_form": "持ち",
    "pos": "動詞",
    "pos_detail_1": "自立",
    "pos_detail_2": "*",
    "pos_detail_3": "*",
    "conjugated_type": "五段・タ行",
    "conjugated_form": "連用形",
    "basic_form": "持つ",
    "reading": "モチ",
    "pronunciation": "モチ"
  }
}
```

japanese-moji

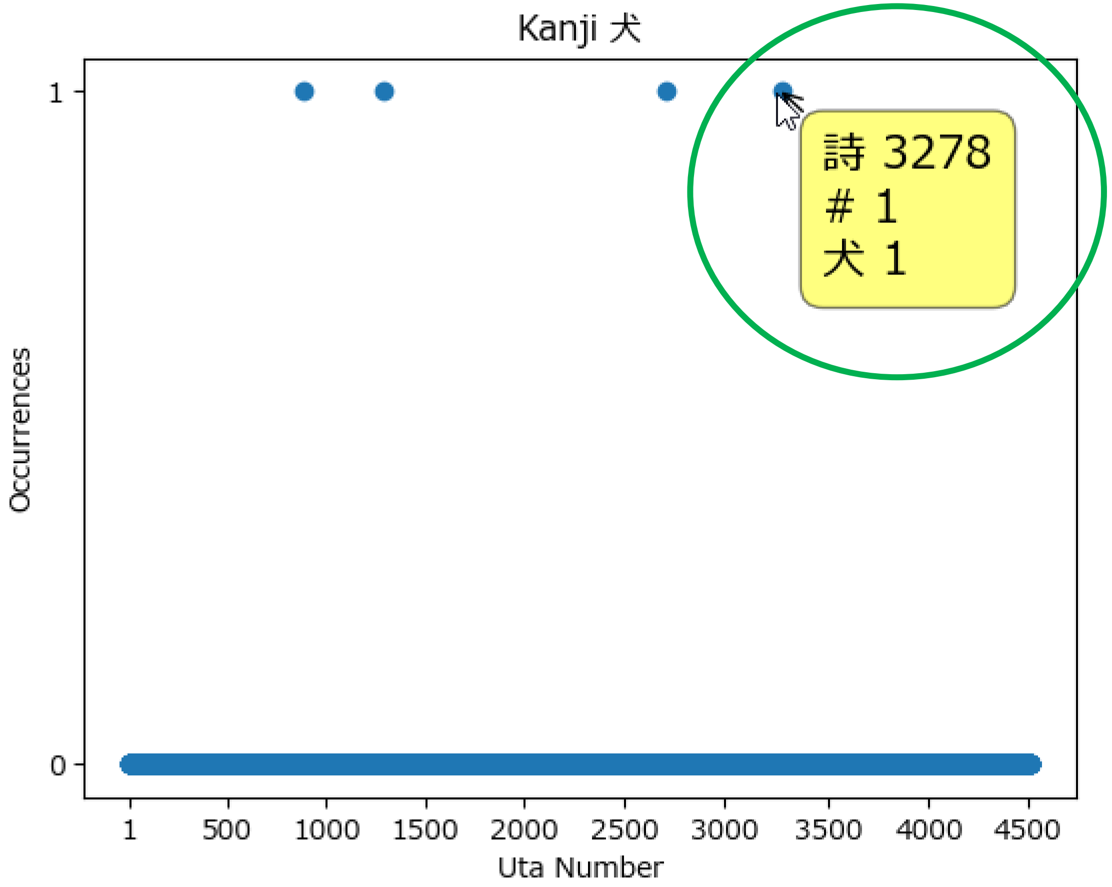
- Handy functions for lazy people
  - `isValidKanji`
- Supports creation of custom validators

```
kanaValidatorOptions = {
    characterSets : [
         CharacterSet.Katakana,
         CharacterSet.KatakanaPhoneticExtension,
         CharacterSet.HalfWidthKatakana,
         CharacterSet.Hiragana
    ]
}
isValidKana = createStrictValidator(kanaValidatorOptions)
```

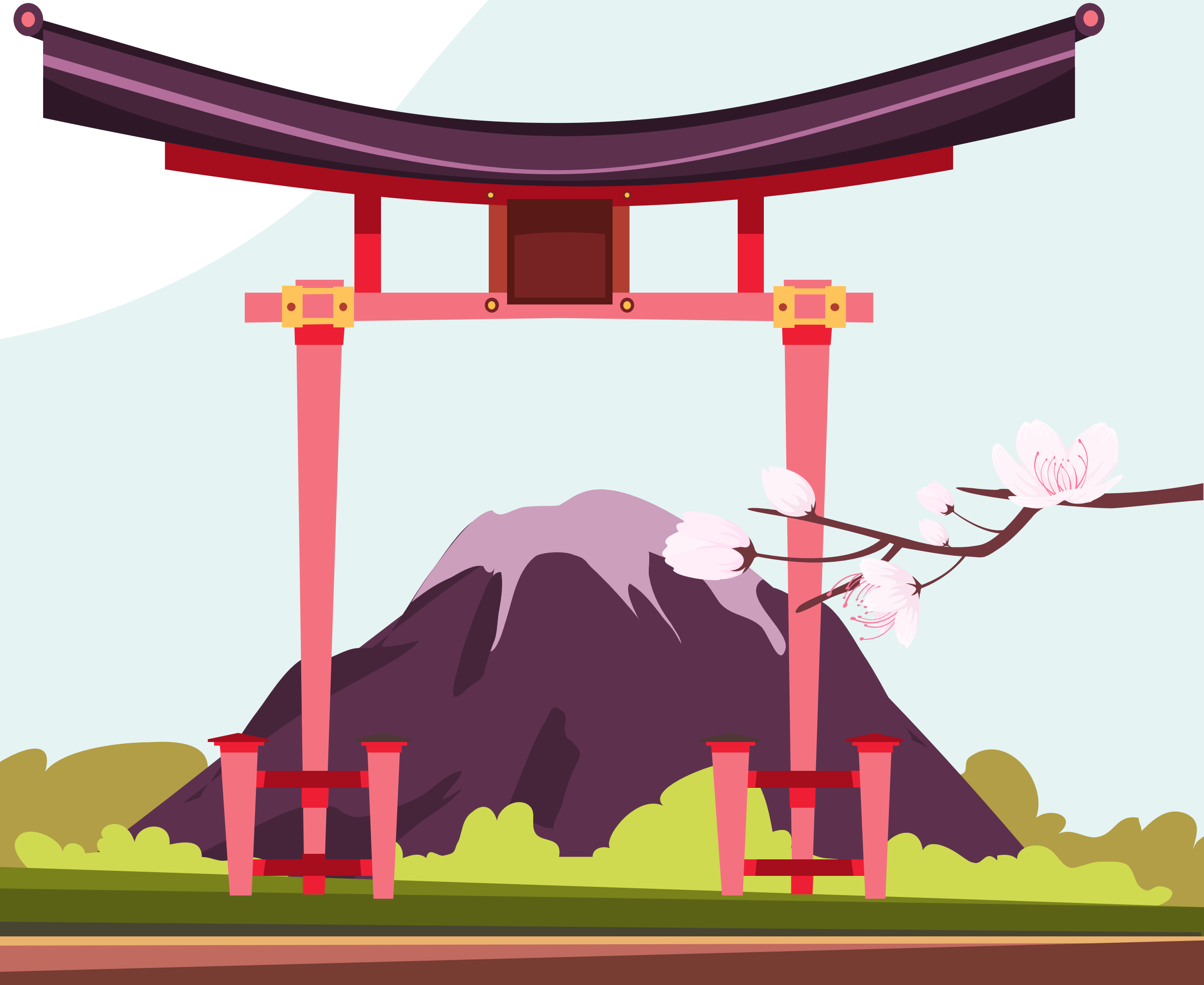
文字
Japanese Moji
A toolkit to validate Japanese characters

mplcursors

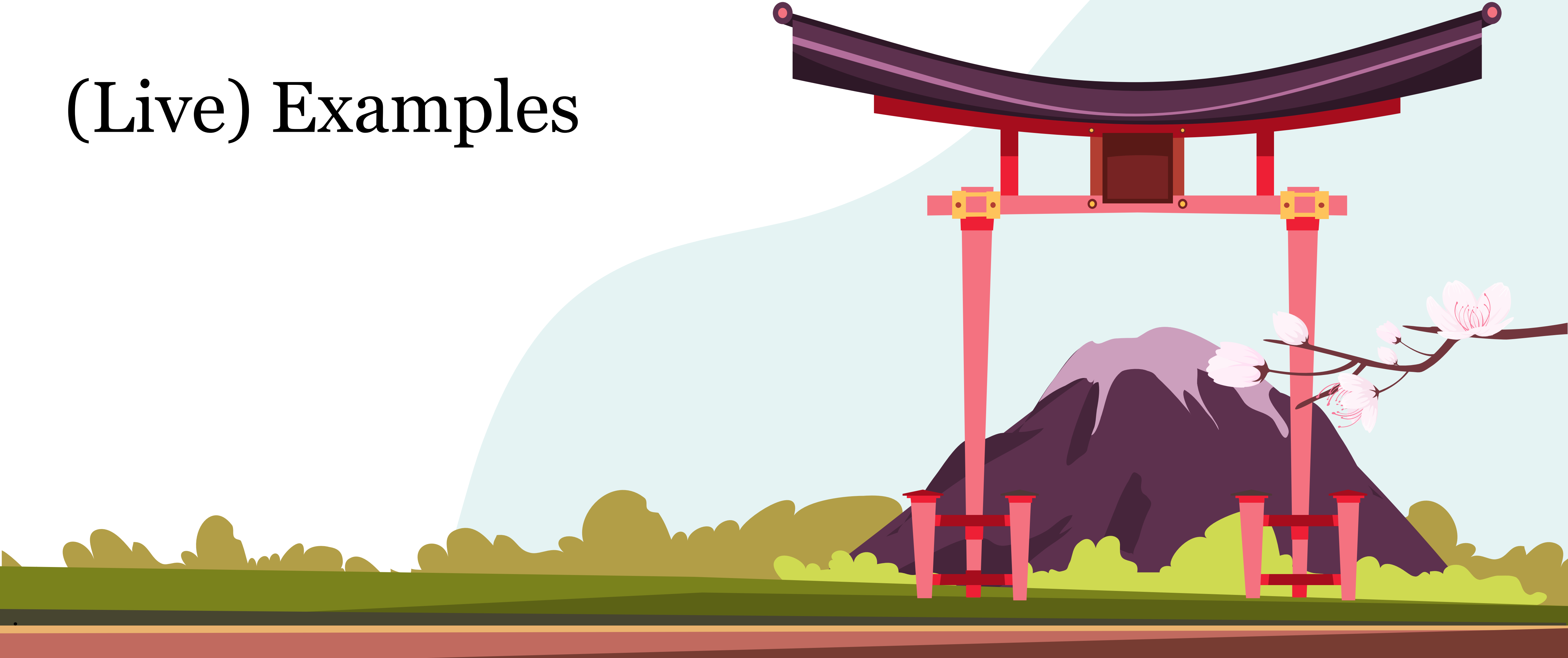
- Interactive data selection tooltips for Matplotlib

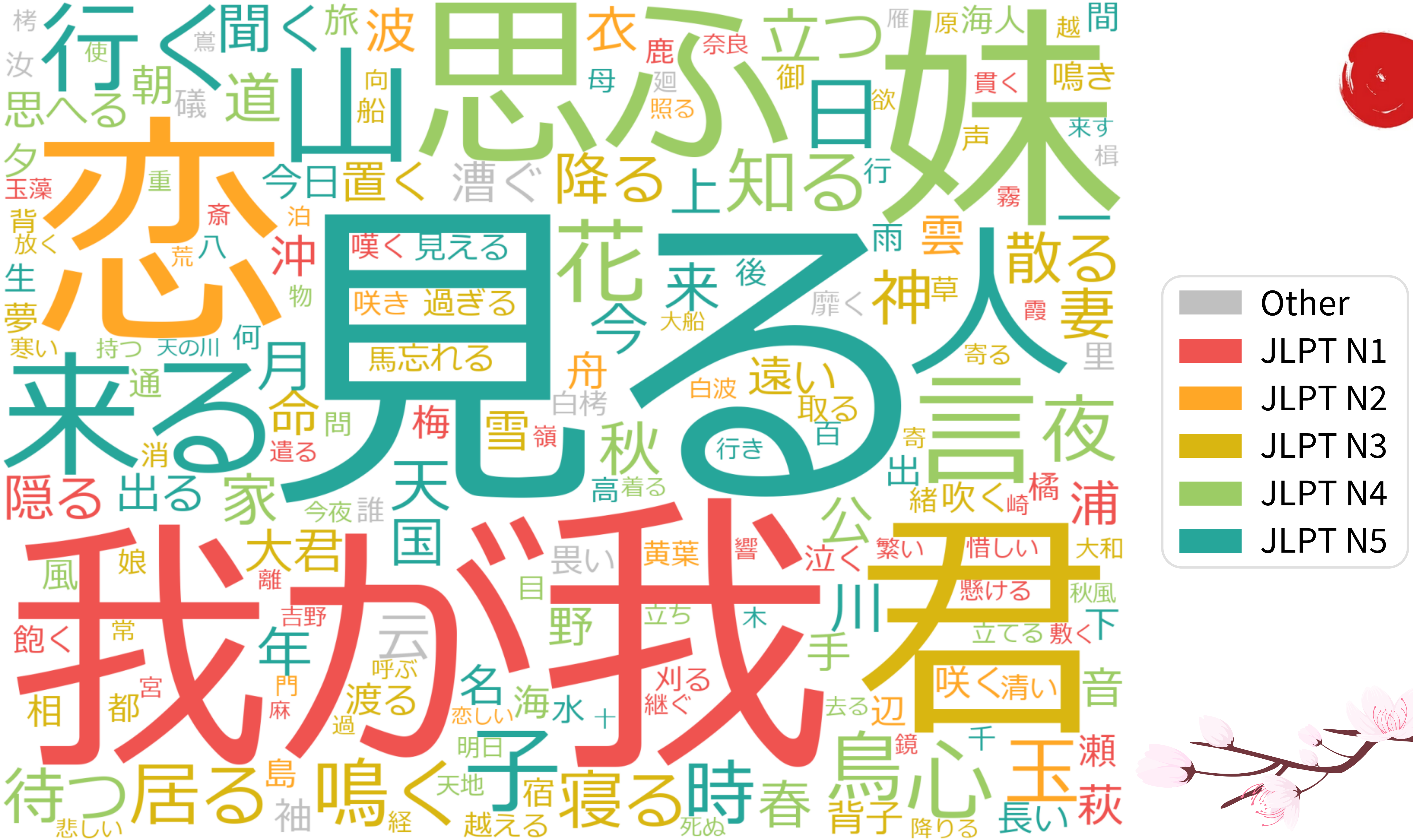


Look at the Code



(Live) Examples




栲
行く
聞く
旅
波
思ふ
衣
鹿
奈良
立つ
雁
原
海人
越
間
使
蔦
汝
山
妹
向
船
母
廻
御
日
欲
貫く
鳴き
朝
思へる
礒
道
照る
来す
声
楫
夕
恋
重
玉藻
今日
置く
漕ぐ
降る
上
知る
行
霧
斎
泊
背
放く
荒
八
沖
見る
嘆く
見える
咲き
過ぎる
馬
忘れる
花
来
後
雨
雲
散る
人
一
生
物
靡く
神
草
霞
妻
夢
大船
今
寒い
持つ
天の川
何
月
寄る
里
来る
通
舟
遠い
白波
白栲
命
問
梅
雪
嶺
取る
百
言
夜
秋
行き
寄
消
遣る
出
隠る
出る
家
天
高
着る
緒
吹く
崎
橘
浦
今夜
誰
公
我
が
我
君
大君
国
畏い
黄葉
響
泣く
繁い
惜しい
大和
風
娘
離
目
懸ける
秋風
川
吉野
野
立ち
木
下
飽く
常
云
年
立てる
敷く
手
呼ぶ
刈る
咲く
清い
宮
門
継ぐ
去る
辺
音
相
都
名
麻
渡る
海
水
十
恋しい
過
子
鳥
心
瀬
明日
鏡
千
玉
島
鳴く
天地
宿
寝る
時
春
萩
待つ
居る
袖
経
越える
死ぬ
背子
降りる
長い
悲しい
Other
JLPT N1
JLPT N2
JLPT N3
JLPT N4
JLPT N5

freq-count.py

Frequency of a *kanji* vs the number of different words the *kanji* is used in.

- In 97% of cases, 我 is used in $\overset{\text{WARE}}{\text{我}}$ or $\overset{\text{WA GA}}{\text{我が}}$
- 君 is virtually always used in $\overset{\text{KIMI}}{\text{君}}$ or $\overset{\text{OO KIMI}}{\text{大君}}$
  - 君 used to mean either “*you*” or “*ruler*”

large, big

- *Kanji* connected with nature are generally used in many different words
  - $\overset{\text{YAMA}}{\text{山}}$, $\overset{\text{KI}}{\text{木}}$, $\overset{\text{TA}}{\text{田}}$, $\overset{\text{HANA}}{\text{花}}$, $\overset{\text{KUSA}}{\text{草}}$, $\overset{\text{NO}}{\text{野}}$, $\overset{\text{TORI}}{\text{鳥}}$, $\overset{\text{HARU}}{\text{春}}$, $\overset{\text{YORU}}{\text{夜}}$, …

Words containing 山

#occurrences

🟩 Common word

YAMA
山 (497)

YAMA BE
山辺 (31)

(SAN SEN)|(YAMA KAWA)
山川 (28)

YAMA MICHI
山道 (23)

AKI YAMA
秋山 (19)

YAMA BUKI
山吹 (18)

OKU YAMA
奥山 (17)

YAMA(SHIGE|SUGA|SUGE)
山菅 (13)

KOU ZAN
高山 (17)

SHIMA YAMA
島山 (11)

(SAN KA)|(YAMA SHITA)|(SAN GE)
山下 (10)

(HARU YAMA)|(SHUN ZAN)
春山 (9)

(NI JOU ZAN)|(FUTA KAMI SAN)|(FUTA GAMI YAMA)
二上山 (9)

YAMA NO HA
山の端 (9)

KASUGA YAMA
春日山 (8)

YAMA GO E
山越え (7)

(SEI ZAN)|(AO YAMA)
青山 (7)

YAMA BIKO
山彦 (7)

and more …
(115 different words in total)

distance-rank.py

- Order the *kanji* by the Euclidian distance
- Associate an integer called *rank* with every *kanji* such that
  - The most distant *kanji* has *rank* 1, the second most 2, …
- Plot the rank and the Euclidian distance

The First Zipf's Law

Let f be the frequency of word, r rank and k an arbitrary constant. Then

$$f \cdot r \approx k$$

Corollary

- A part of lexicon is defined by a small number of frequented words

kanji-word-occ.py

- Plots the occurrence across the whole anthology
- Accepts one positional argument (*word*/*kanji*)
    - In case of a kanji, add an optional parameter –k

```
python kanji-word-occ.py -k "神"
python kanji-word-occ.py "犬"
python kanji-word-occ.py -k "雪"
```

Uta 886

Abych došel do paláce,
tonoucího v záři slunce,
musel jsem opustit náruč matky,
paní plných ňader, a brát se
do krajů, jež jsem v životě neviděl,
překročit sterou hradbu strmých
hor.
Kdy už spatříme hlavní město,
ptám se denně svých druhů,
ale pak bolest láme mé tělo
a já klesám u cesty, která se
táhne do dáli jako kopí z nefritu.
U cesty prostřeli mi narychlo
lože z trávy a zelených větviček
a já vzdychám, jak tu v žalostné
bídě ležím.

Být v rodné zemi, pečoval by o mě otec,
být doma, objímala by mě matka.
Že to tak na světě chodí?
Mám snad zajít u cesty jako toulavý pes?

(Translated by Antonín Líman)

uta-types.py

Calculate the number of poems for every *Waka* form

- Overwhelming majority (92%) of poems are of type *Tanka* (5-7-5-7-7)
- Even a *Bussokusekika* poem is present
  - Poems inscribed on the Buddha's footprints at *Yakushi-ji* (薬師寺)

morae-dist.py

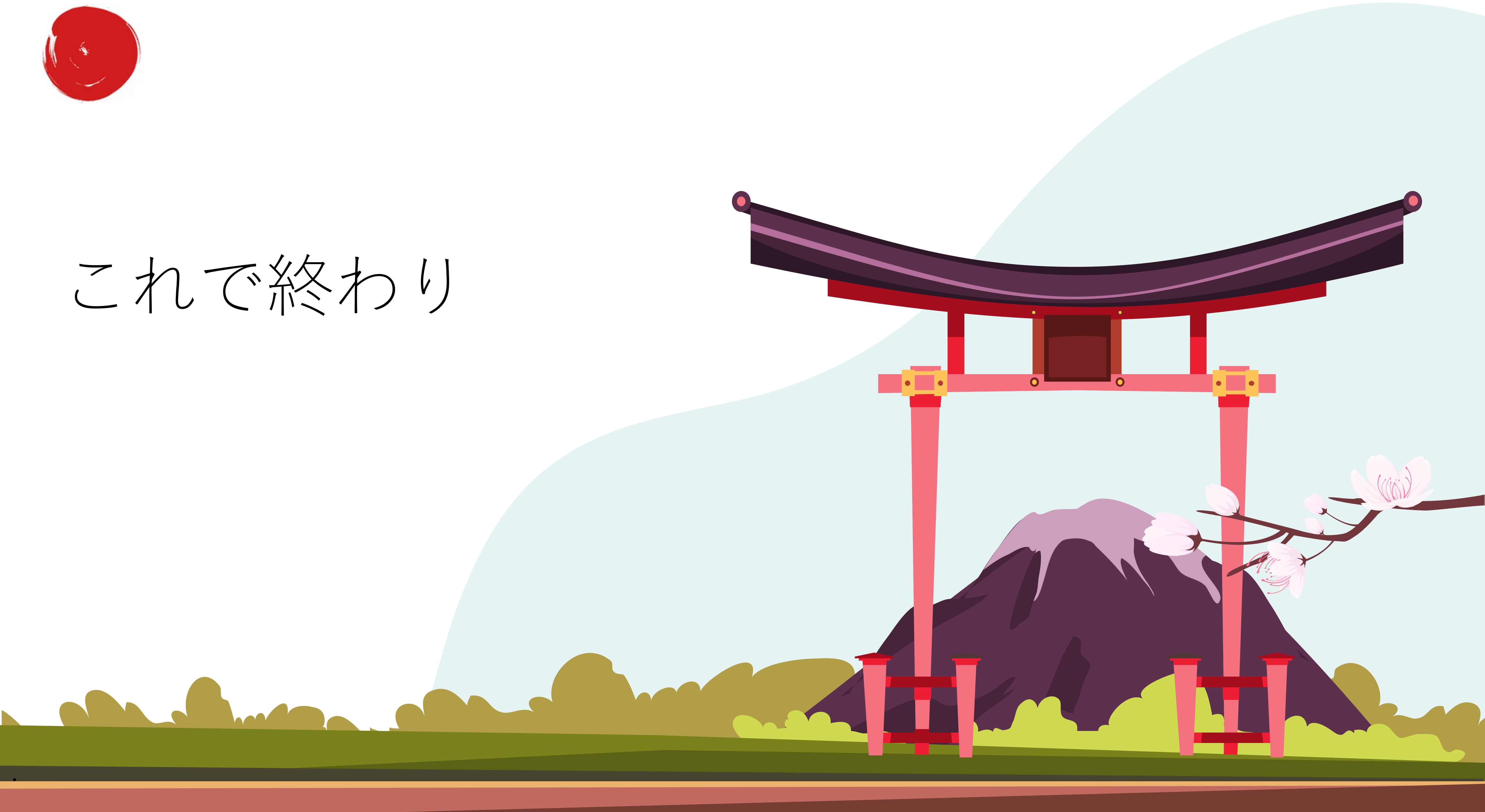
Check the compliance of the form of a poem in modern Japanese.

- *Chouka* is (obviously) the most problematic one
- 2708 poems are compliant to their forms
- 1330 poems have L_1 distance of 1
- 316 poems have L_1 distance of 2
- Together, 4354 poems have L_1 distance of at most 2
    - *Manyoushuu* contains 4516 poems

verse-count.py

- Between poems 3221 and 3346, there are 67 poems of type *Chouka*
  - 25% of all *Chouka* poems
- There are other smaller clusters of *Chouka* poems

これで終わり

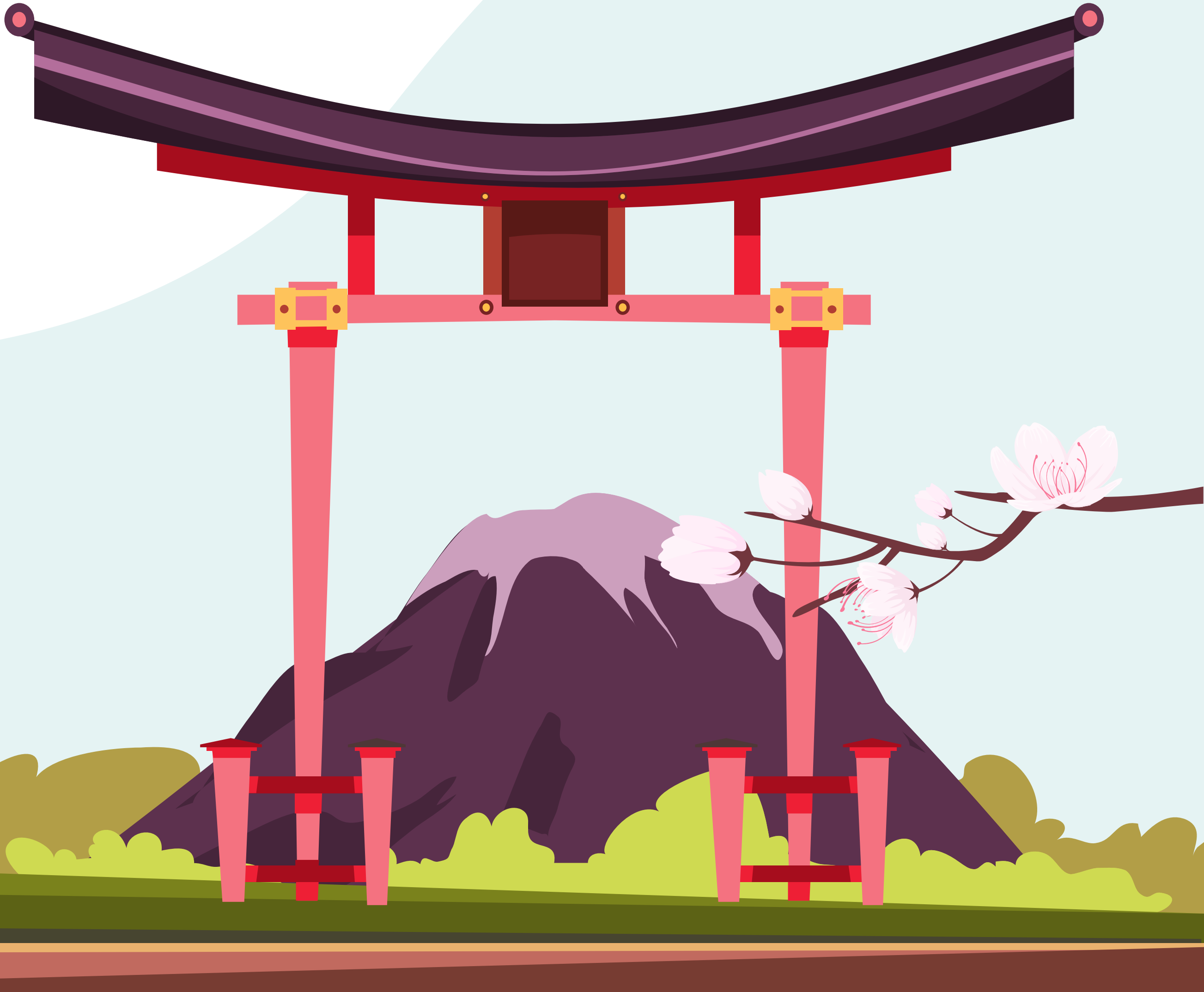


Conclusion

- We have learned a bit
  - About how is/was Japanese written
  - Old Japanese poetry
  - The visualization project
- We have seen
  - Examples (writing, sentences, poetry, code)
  - Visualizations
  - Results

Code & Slides

https://gitlab.mff.cuni.cz/levyjak/visualization-project



References (1)

- Japonsko-český studijní znakový slovník, druhé vydání
- Manjóšú – Deset tisíc listů ze starého Japonska, díl první
- Manjóšú – Deset tisíc listů ze starého Japonska, díl druhý
- Pár much a já: malý výběr z japonských haiku
- https://www.japanesewithanime.com/2019/11/mora.html
- https://www.japanesewithanime.com/2017/12/compound-kana.html
- https://ja.wikipedia.org/wiki/%E5%92%8C%E6%AD%8C
- https://www.britannica.com/art/haiku

References (2)

- https://ja.wikipedia.org/wiki/%E4%B8%87%E8%91%89%E9%9B%86
- https://en.wikipedia.org/wiki/Man%27y%C5%8Dsh%C5%AB
- https://cs.wikipedia.org/wiki/Manj%C3%B3%C5%A1%C3%BA
- https://github.com/puppeteer/puppeteer
- https://github.com/azu/nlp-pattern-match/tree/master/packages/nlcst-parse-japanese
- https://github.com/arjunvegda/japanese-moji
- https://wikisofia.cz/wiki/Zipfovy_z%C3%A1kony

References (Pictures) (1)

- https://www.iconfinder.com/icons/6900234/alligator_crocodile_crocodile_icon_predator_icon
- https://en.wikipedia.org/wiki/Katakana#/media/File:Table_katakana.svg
- https://en.wikipedia.org/wiki/S%C5%8Dgana#/media/File:Sogana.png
- https://en.wikipedia.org/wiki/Hiragana#/media/File:Table_hiragana.svg

References (Pictures) (2)

- https://www.japanpowered.com/japan-culture/the-life-and-impact-of-matsuo-basho
- https://www.databazeknih.cz/img/books/13_/13545/big_par-much-a-ja-yD0-13545.jpg
- https://www.istockphoto.com/cs/vektor/%C5%BEab%C3%AD-sk%C3%A1k%C3%A1n%C3%AD-izolovan%C3%A1-%C5%BE%C3%A1ba-sk%C3%A1%C4%8De-na-b%C3%ADl%C3%A9m-pozad%C3%AD-gm1130502836-299008045
- https://www.stockio.com/free-icon/nature-icons-blue-fish

References (Pictures) (3)

- https://www.historyofcreativity.com/content/6/s_creator.001_1589061380.jpg
- http://www.c-d-f.cz/media/images/authors/antonin-liman.jpg
- https://cdn.aukro.cz/images/sk1622082133201/730x548/manjosu-dil-i-az-iv-komplet-101033023.jpeg
- https://user-images.githubusercontent.com/10379601/29446482-04f7036a-841f-11e7-9872-91d1fc2ea683.png

References (Miscellaneous)

- Template
  - https://www.slideegg.com/japan-theme-presentation
  - https://www.shutterstock.com/cs/image-vector/set-handdrawn-traditional-japanese-symbols-red-418482400
- Speech bubbles
  - https://www.elearningdesigner.co.uk/freebies