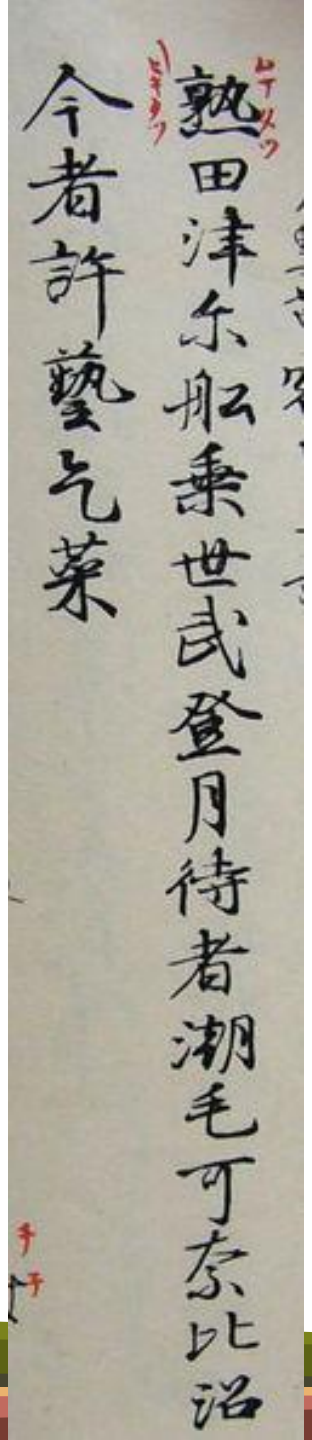
Manyoushuu (万葉集)

Jakub Levý
jakub.levy15@gmail.com

NDBI042 2021/2022

Faculty of Mathematics and Physics
Charles University

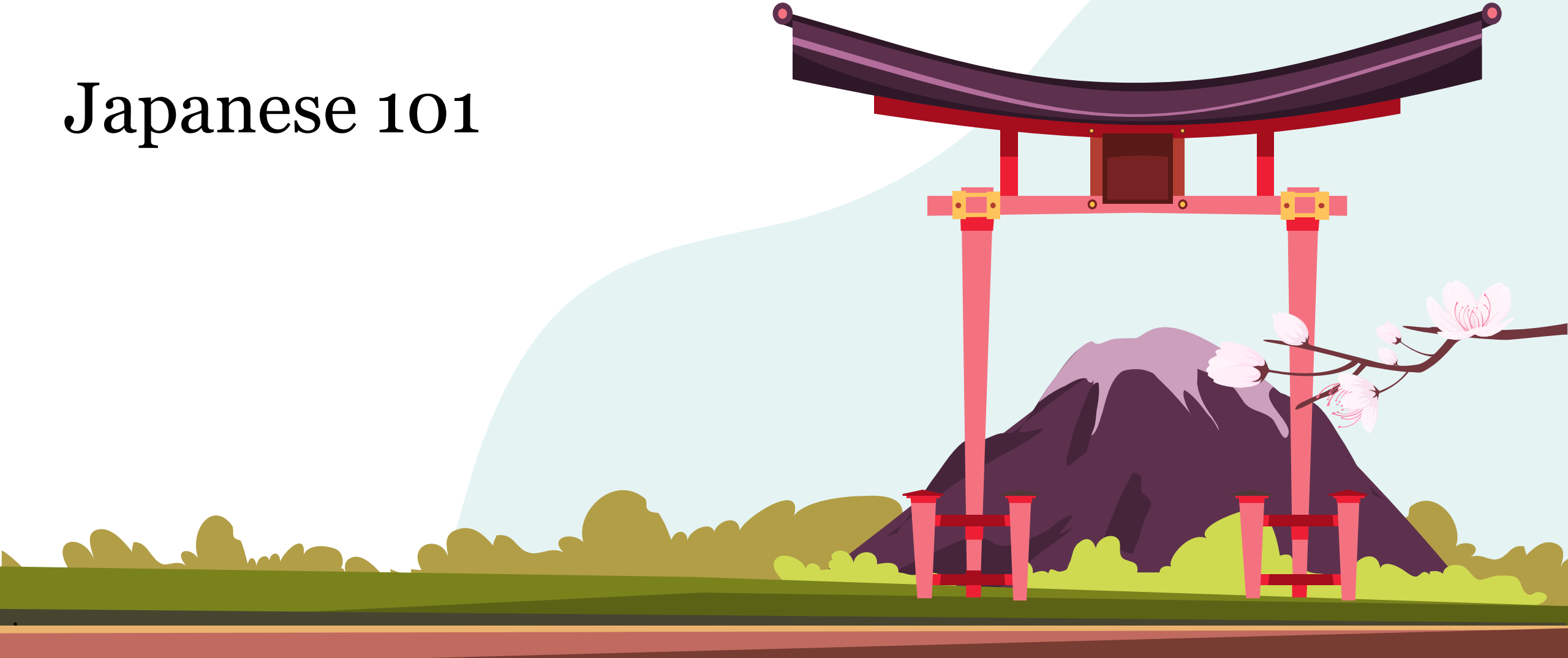


CC BY SA
unless otherwise stated

Contents

- Japanese 101
  - Writing System
  - Mora
  - Waka
  - Haiku
- Manyoushuu
  - About
  - Poem in detail
- Project
  - Visualization Goals
  - Overview
  - Technologies
  - Look at the Code
  - (Live) Examples
- Conclusion
- References

Japanese 101

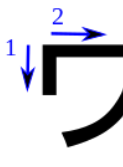
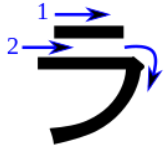
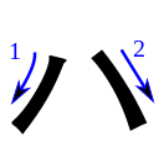
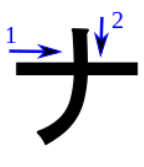
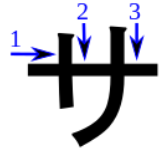
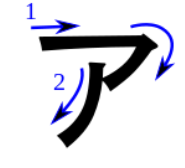
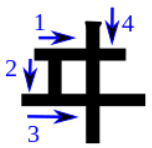
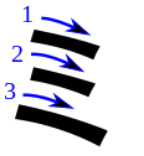
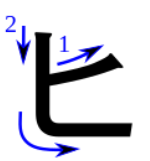
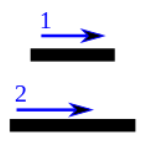
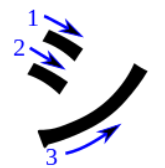
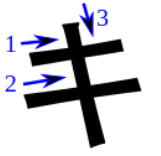
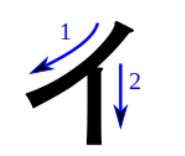
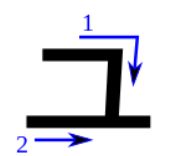
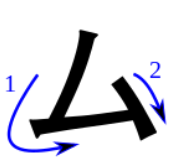
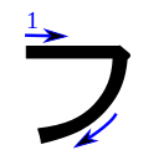
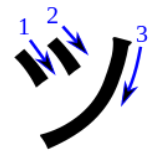
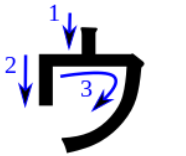
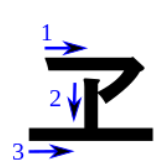
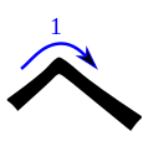
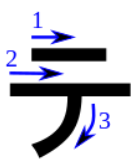
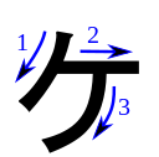
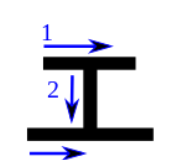
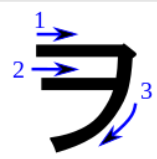
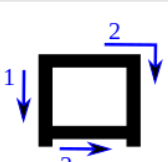
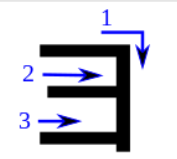
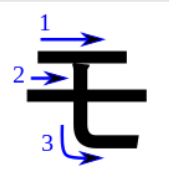
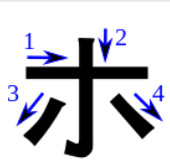
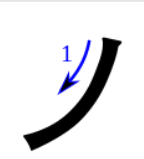
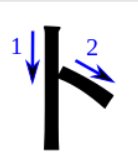
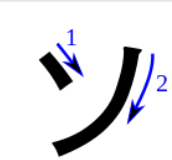
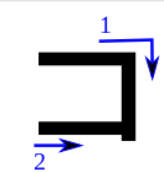
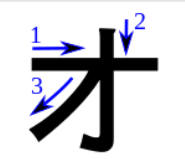


Japanese Writing System (1)

- 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)

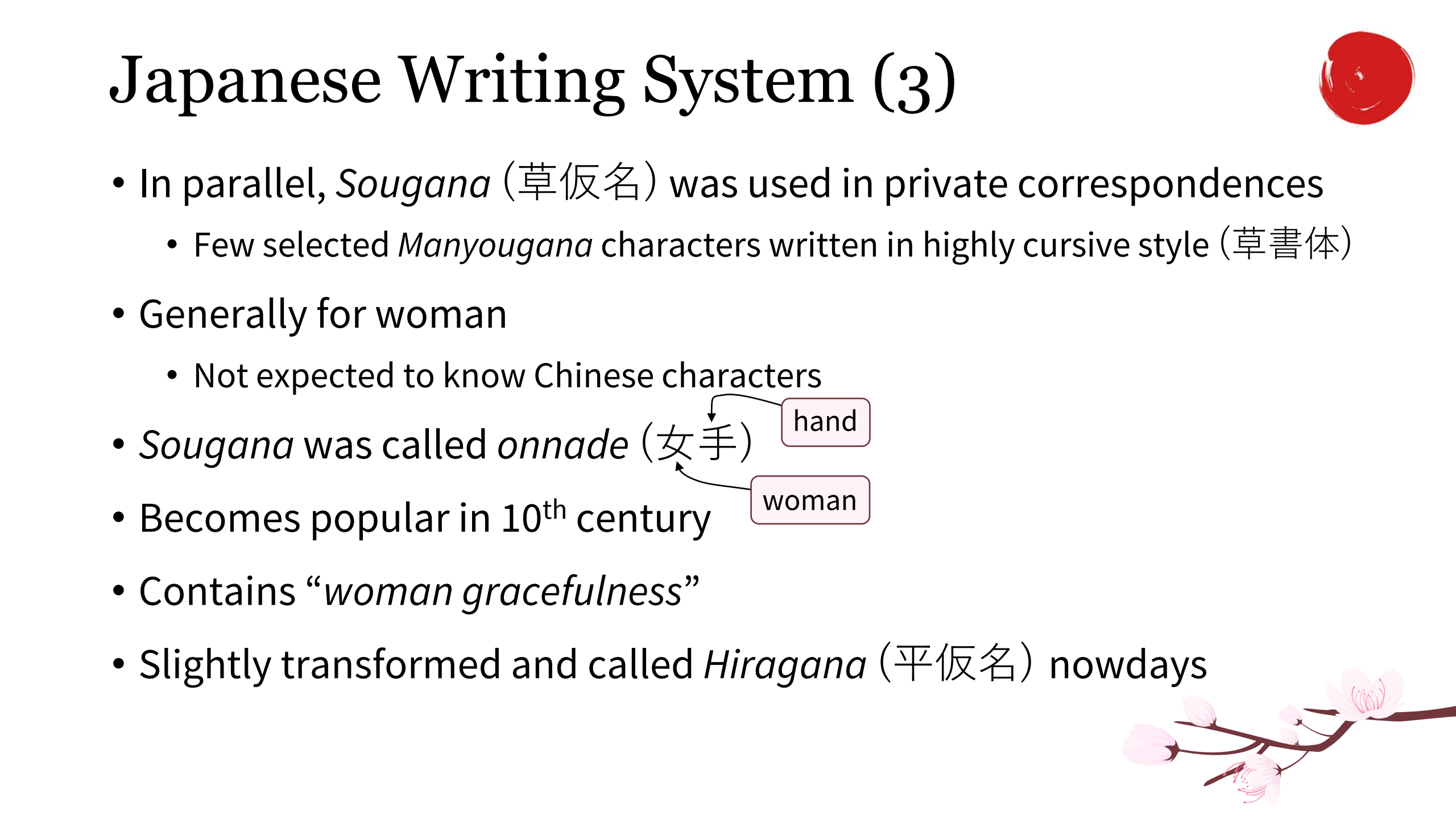
Japanese Writing System (2)

- *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–	
N	 *WA*	 *RA*	*YA*	*MA*	 *HA*	 *NA*	*TA*	 *SA*	*KA*	 *A*	–a
	 *WI*	*RI*		 *MI*	 *HI*	 *NI*	*CHI*	 *SHI*	 *KI*	 *I*	–i
		RU	 *YU*	 *MU*	 *FU*	*NU*	 *TSU*	*SU*	*KU*	 *U*	–u
	 *WE*	*RE*		*ME*	 *HE*	*NE*	 *TE*	*SE*	 *KE*	 *E*	–e
	 *WO*	 *RO*	 *YO*	 *MO*	 *HO*	 *NO*	 *TO*	 *SO*	 *KO*	 *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)
- Becomes popular in 10th century
- Contains "*woman gracefulness*"
- Slightly transformed and called *Hiragana* (平仮名) nowdays



无 ゑん	和 わ	良 ら	也 や	末 ま	波 は	奈 な	太 た	左 さ	加 か	安 あ
	爲 ゐる	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

Chinese character

Sougana

Hiragana

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

CD 1 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.

IP PO DE MO UGO I TA RI MYOU NA MA NE O SU RE BA KO NO GA KI NO KUBI O KA KI KI RU

一歩でも動いたり 妙 なマネをすればこのガキの 首 を掻き切る。

Take even one step or make a funny move and I'll slit this brat's throat.

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
$k = 5, m = 4, s = 2$

kana　morae　syllables

- 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

古池や
蛙飛び込む
水の音

Manyoushuu (万葉集)

Manyoushuu (万葉集) (1)

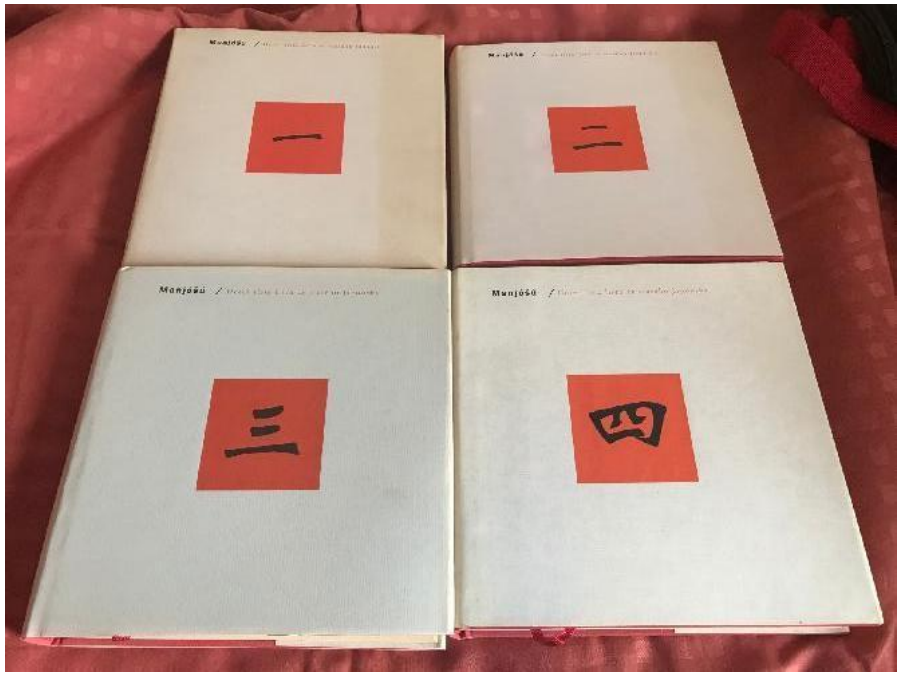
- 万葉集 ~ “collection of ten thousand leaves”
- Oldest collection of *Waka* poetry
- Written in *Manyougana*
- More than 4000 *Waka* poems
- 561 authors
  - 70 woman
  - People of various statuses: emperors, aristocrats, peasants, street performers, …
- Over 2100 poems by unknown authors

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)



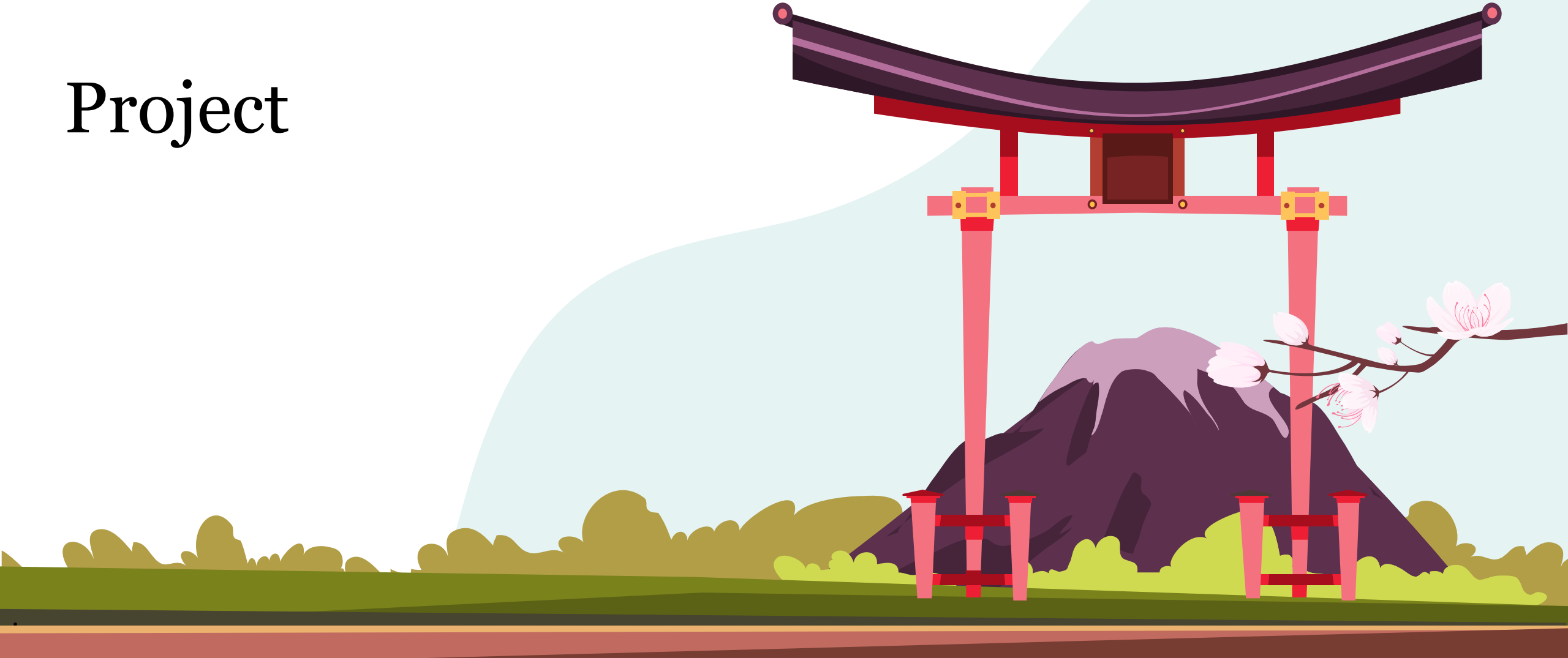
Uta 8 (Princess Nutaka's Poem)

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

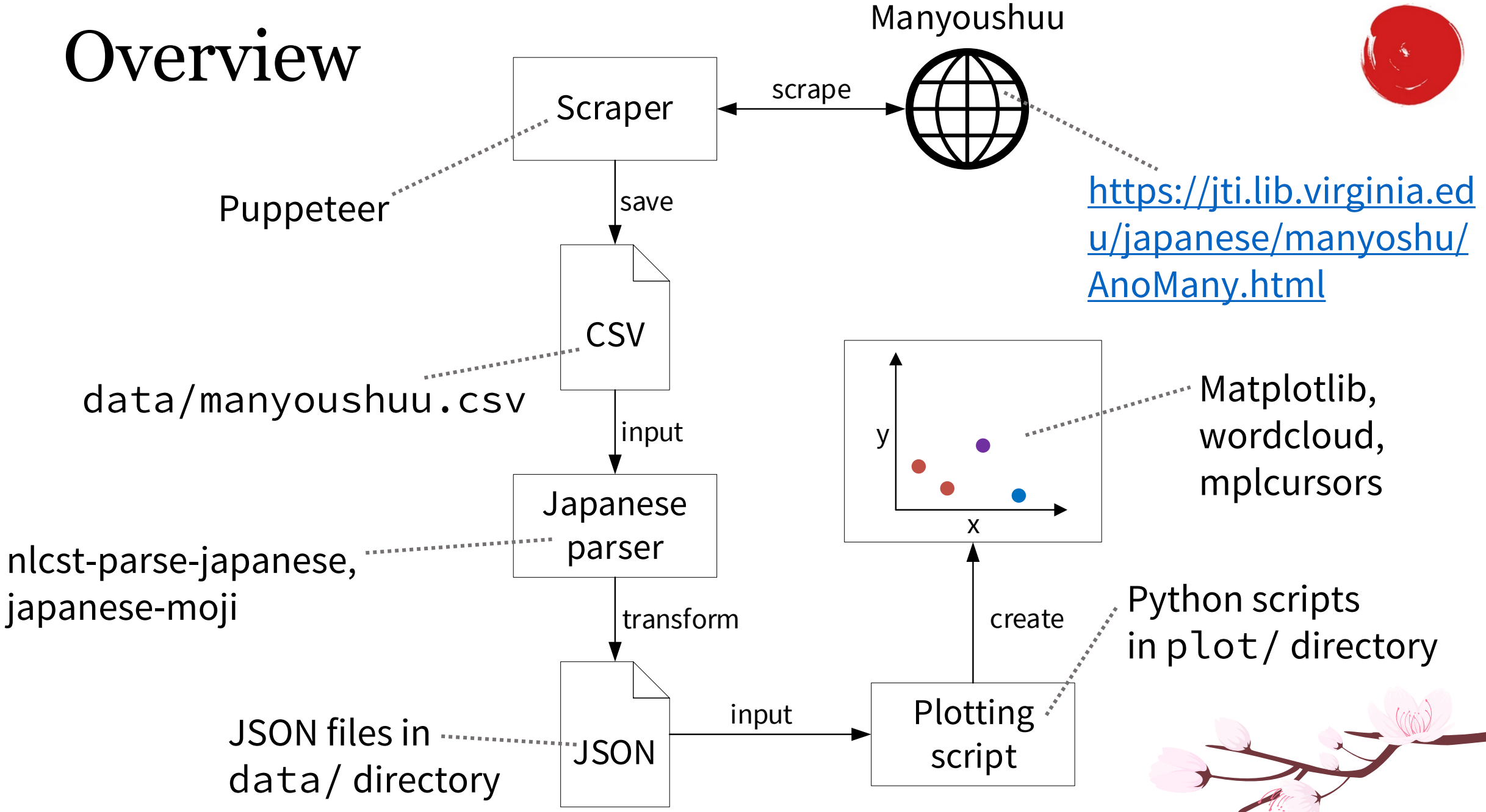


Visualization Goals

- 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

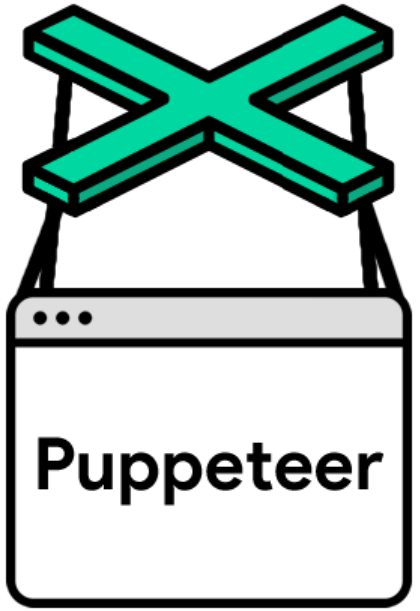
Different forms of a word must be handled correctly
持たず and 持ち are forms of 持つ

Overview


Manyoushuu
Scraper
scrape
Puppeteer
save
https://jti.lib.virginia.edu/japanese/manyoshu/AnoMany.html
CSV
data/manyoushuu.csv
input
y
x
Matplotlib,
wordcloud,
mplcursors
Japanese
parser
nlcst-parse-japanese,
japanese-moji
transform
create
Python scripts
in plot/ directory
JSON files in
data/ directory
JSON
input
Plotting
script

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)

Parser output for 持ち (MO CHI) (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

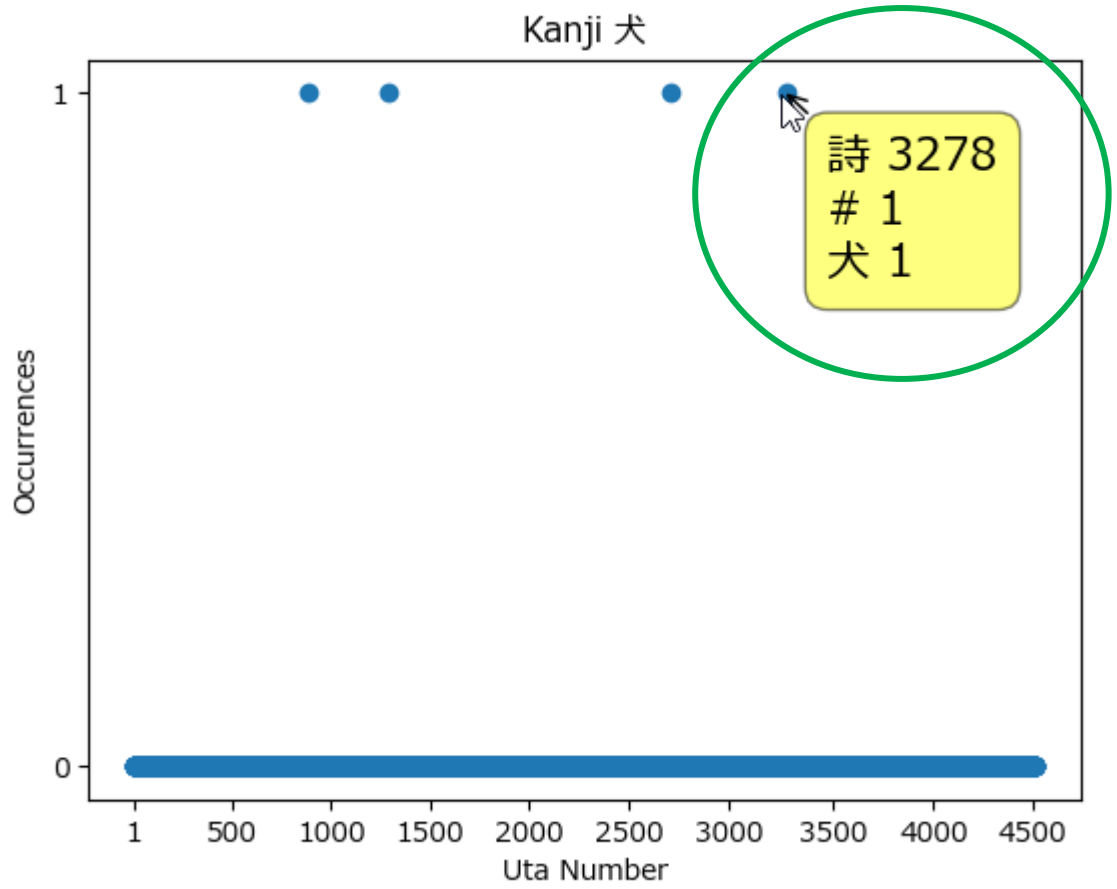
- Handy functions for lazy people
  - `isValidKanji`
- Supports creation of custom validators

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

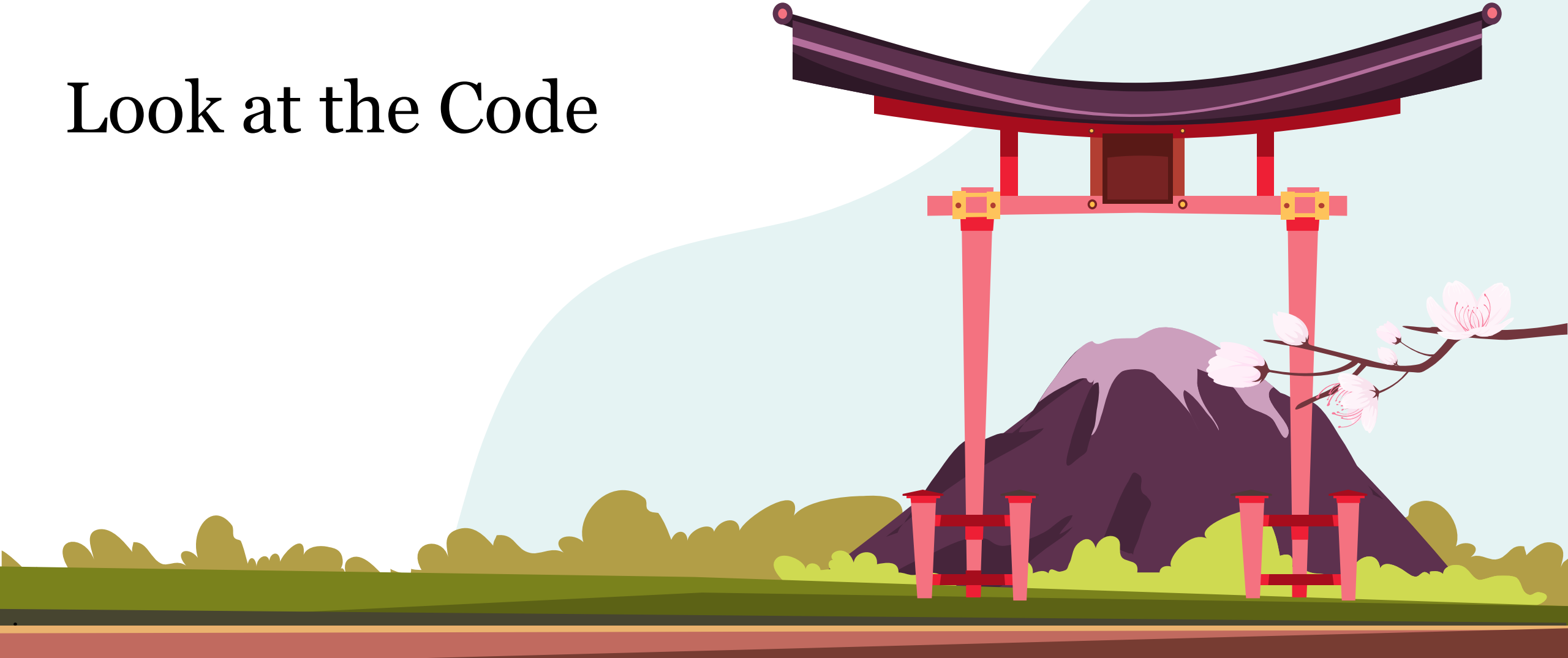
文字
Japanese Moji
A toolkit to validate Japanese characters

mplcursors

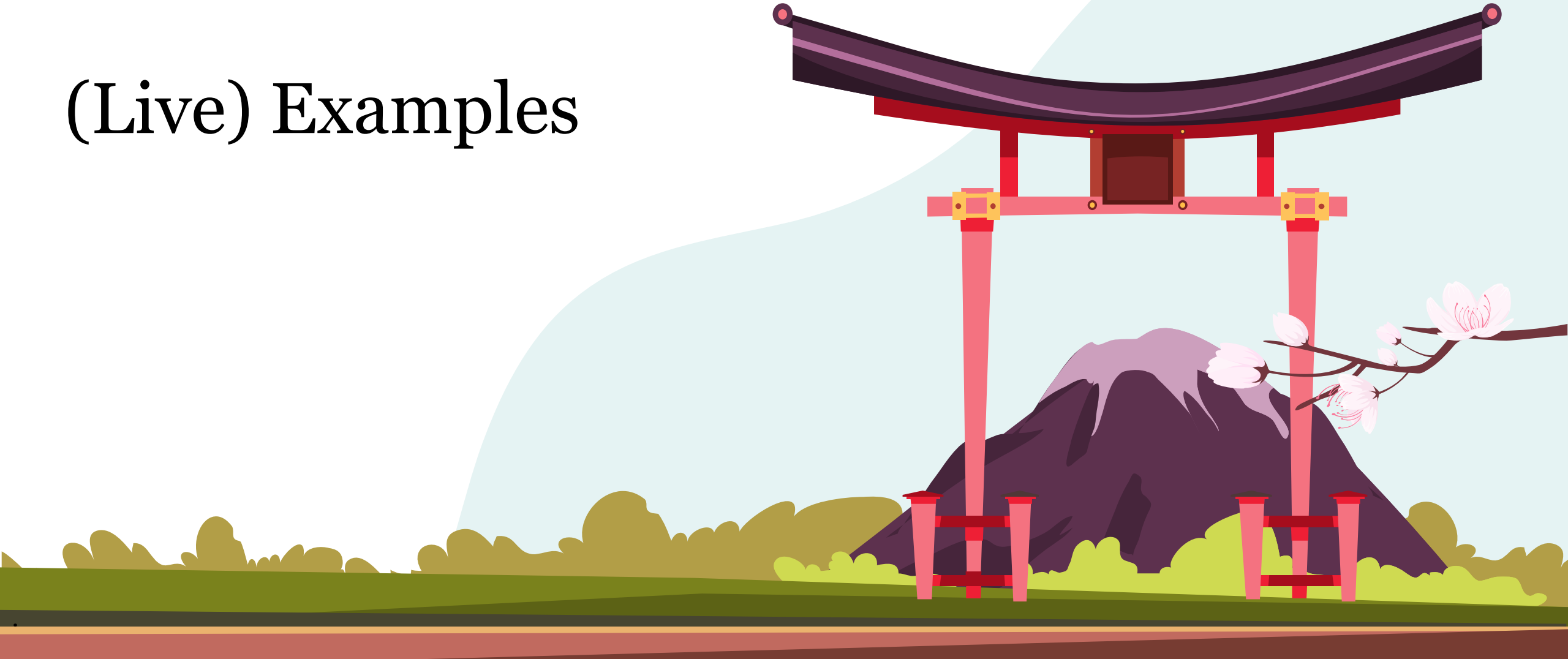
- Interactive data selection tooltips for Matplotlib



Look at the Code



(Live) Examples



嘆く 見える
咲き 過ぎる
馬忘れる

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 WARE 我 or WA GA 我が
- 君 is virtually always used in KIMI 君 or OO KIMI 大君
  - 君 used to mean either “*you*” or “*ruler*”

large, big

- *Kanji* connected with nature are generally used in many different words

YAMA 山, KI 木, TA 田, HANA 花, KUSA 草, NO 野, TORI 鳥, HARU 春, YORU 夜,…

night

mountain, tree, rice field, flower, grass, field, bird, spring

Words containing 山

#occurrences

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)

Common word

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"
```

god

dog

snow

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