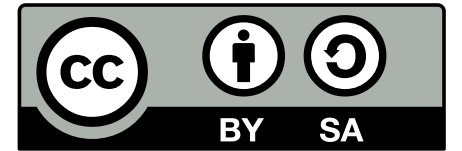
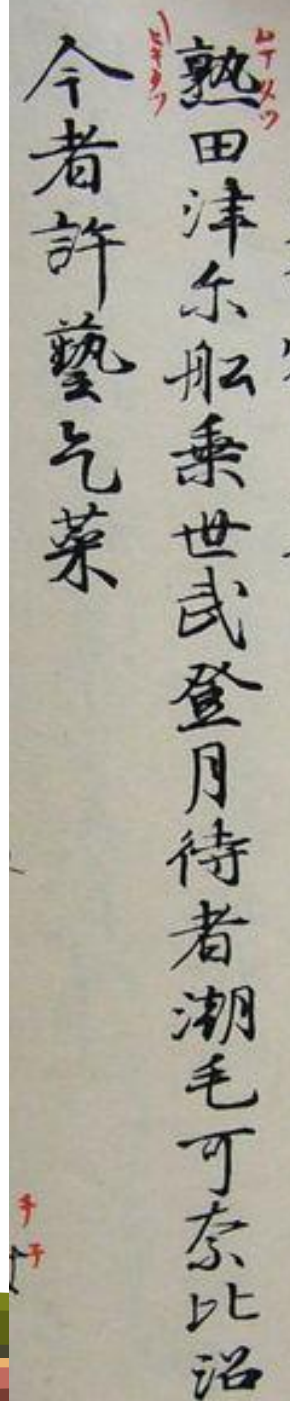


# Manyoushū (万葉集)

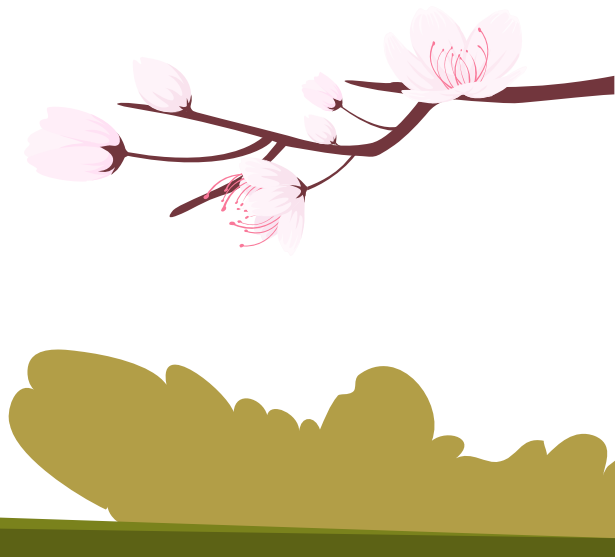
*Jakub Levý*

*[jakub.levy15@gmail.com](mailto:jakub.levy15@gmail.com)*

*Faculty of Mathematics and Physics  
Charles University*



unless otherwise stated



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



无 えん	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 かわ	寸 す	久 く	宇 う
	恵 ゑ	礼 れ		女 め	部 ぶ	祢 ね	天 てん	世 せ	計 け	衣 え
	遠 を	呂 ろ	与 よ	毛 も	保 ほ	乃 の	止 と	曾 そ	己 こ	於 お





n	w-	r-	y-	m-	h-	n-	t-	s-	k-		
 <i>N</i>	 <i>WA</i>	 <i>RA</i>	 <i>YA</i>	 <i>MA</i>	 <i>HA</i>	 <i>NA</i>	 <i>TA</i>	 <i>SA</i>	 <i>KA</i>	 <i>A</i>	-a
	 <i>WI</i>	 <i>RI</i>		 <i>MI</i>	 <i>HI</i>	 <i>NI</i>	 <i>CHI</i>	 <i>SHI</i>	 <i>KI</i>	 <i>I</i>	-i
		 <i>RU</i>	 <i>YU</i>	 <i>MU</i>	 <i>FU</i>	 <i>NU</i>	 <i>TSU</i>	 <i>SU</i>	 <i>KU</i>	 <i>U</i>	-u
	 <i>WE</i>	 <i>RE</i>		 <i>ME</i>	 <i>HE</i>	 <i>NE</i>	 <i>TE</i>	 <i>SE</i>	 <i>KE</i>	 <i>E</i>	-e
	 <i>WO</i>	 <i>RO</i>	 <i>YO</i>	 <i>MO</i>	 <i>HO</i>	 <i>NO</i>	 <i>TO</i>	 <i>SO</i>	 <i>KO</i>	 <i>O</i>	-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.

- 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

馬鹿 → <sup>BA KA</sup>  
ばか

- 2 K, 2 M, 2 S

大学 → <sup>DA I GA KU</sup>  
だいがく

- 4 K, 4 M, 3 S

結婚 → <sup>KE K KO N</sup>  
けっこん

- 4 K, 4 M, 2 S

東京 → <sup>TO U KYO U</sup>  
とうきょう

- 5 K, 4 M, 2 S

- Generally  $\#moraes \neq \#kana \neq \#syllables$



# Waka (和歌 or 倭歌)



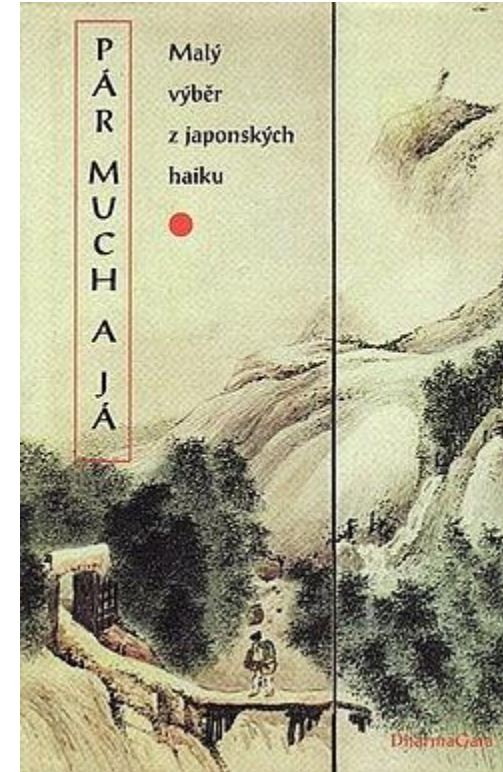
- Poetry in classical Japanese literature
- Many forms
  - Differ by the number of morae
- *Tanka* (短歌)
  - Major genre
  - 5-7-5-7-7
- *Katauta* (片歌)
  - Shortest form of *Waka*
  - 5-7-7
- *Sedouka* (旋頭歌)
  - *Katauta* twice
  - 5-7-7-5-7-7
- *Chouka* (長歌)
  - 5-7 repeated at least twice
  - Concluded 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 17<sup>th</sup> 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)



# Manyoushū (万葉集) (1)



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



# Manyoushū (万葉集) (2)



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

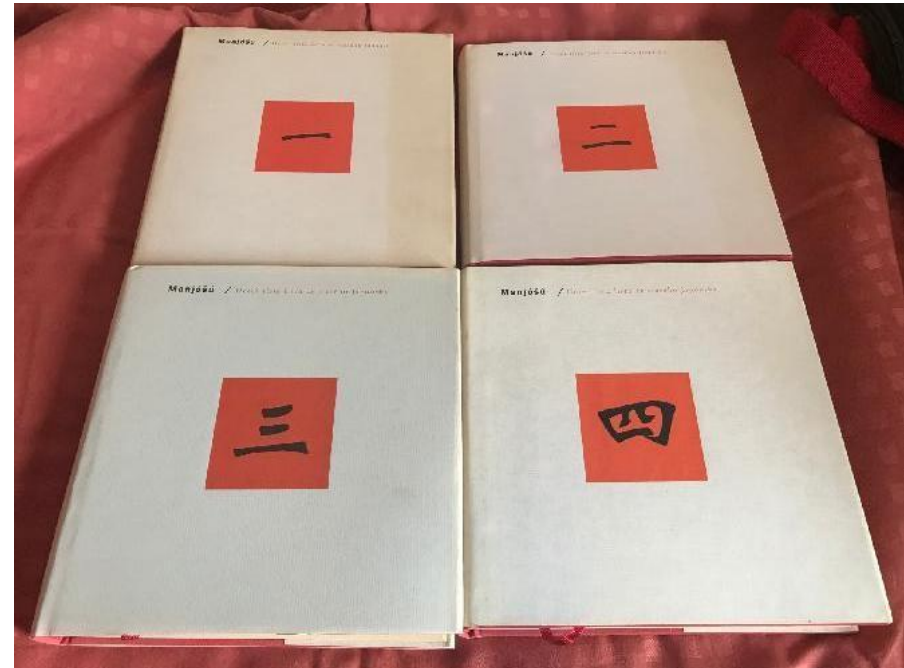




# Manyoushū (万葉集) (3)



- 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	熟田津尔	NIGI TA TSU NI 熟田津に	NI GI TATSU 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

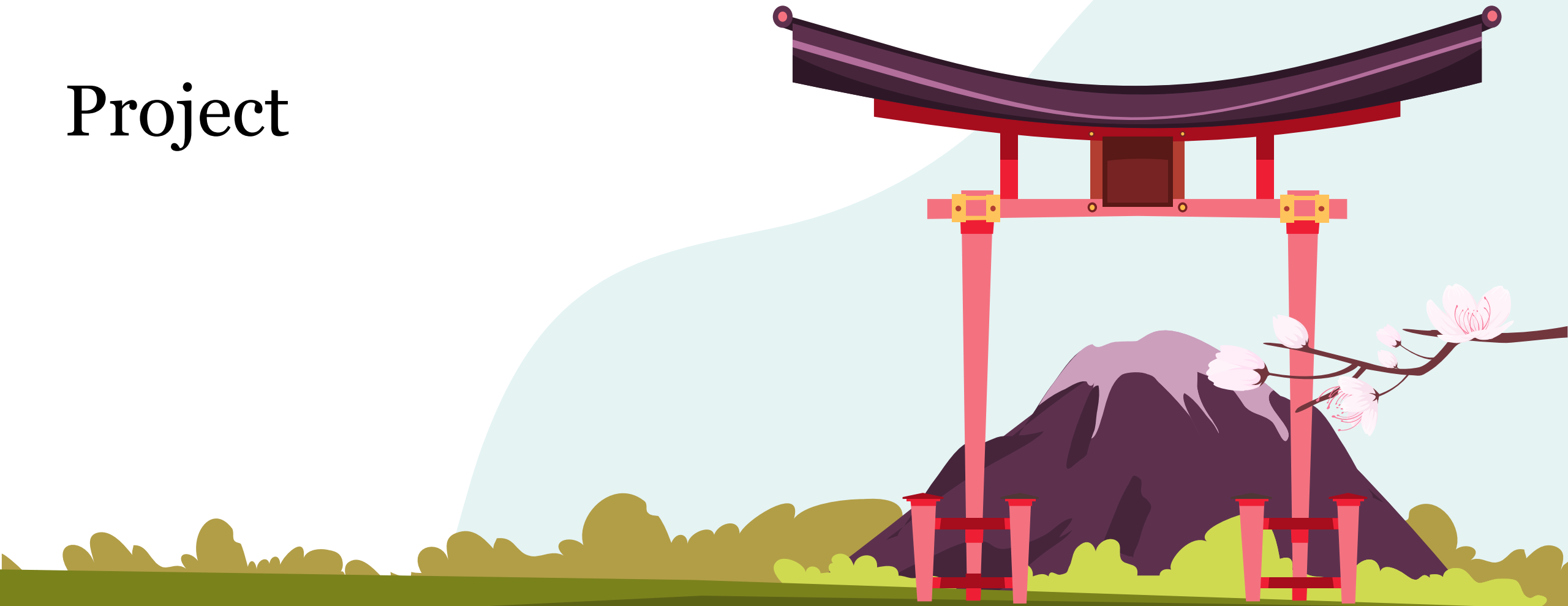
U lod'ky v zálivu Nikita  
čekáme na měsíc...  
Příliv nám přeje,  
nuže, k veslům, k veslům!

(Translated by Antonín Líman)





# Project



# Visualization Goals

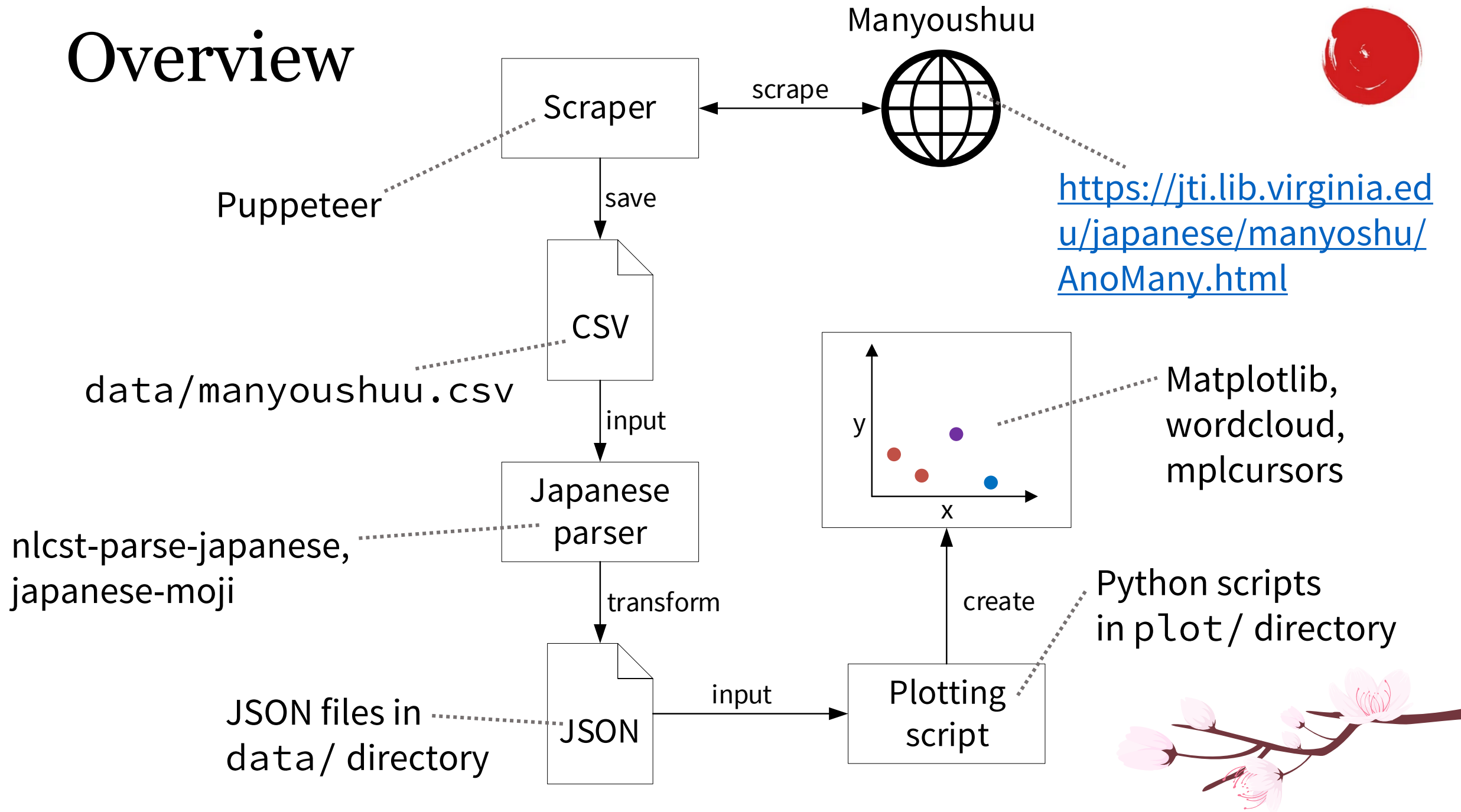


- Word cloud of the most used words
- Frequency of a *kanji* vs the number of different words the *kanji* is used in
  - And be able to somehow quickly see JLPT level of the *kanji*
- Check the compliance of the form of a poem in modern Japanese
- Calculate the number of poems for every *Waka* form
-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



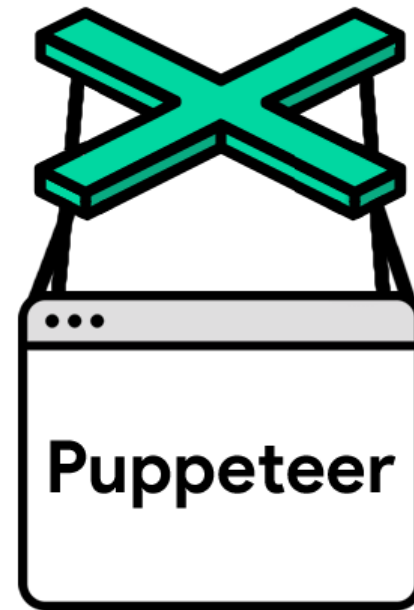
# Overview



# 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

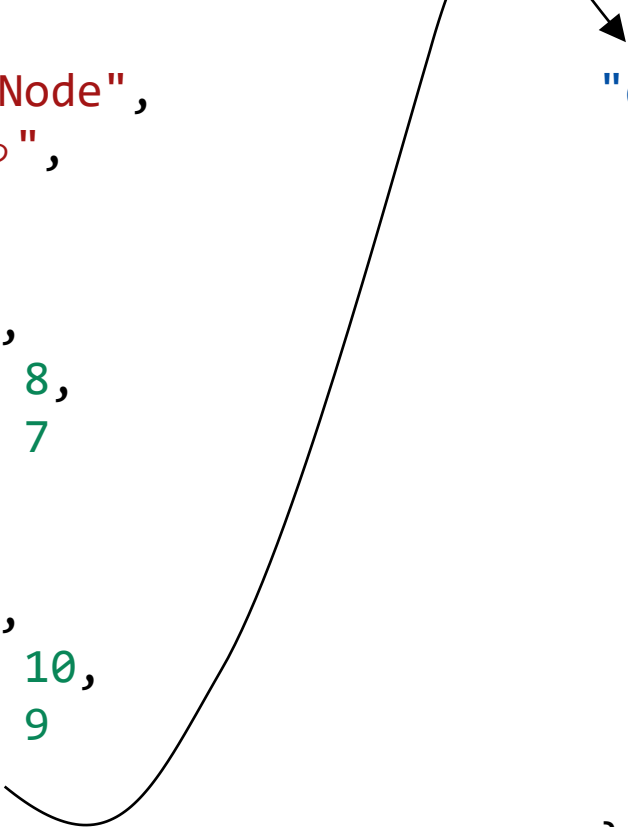


# nlcst-parse-japanese (2)



Parser output for <sup>MO CHI</sup>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**

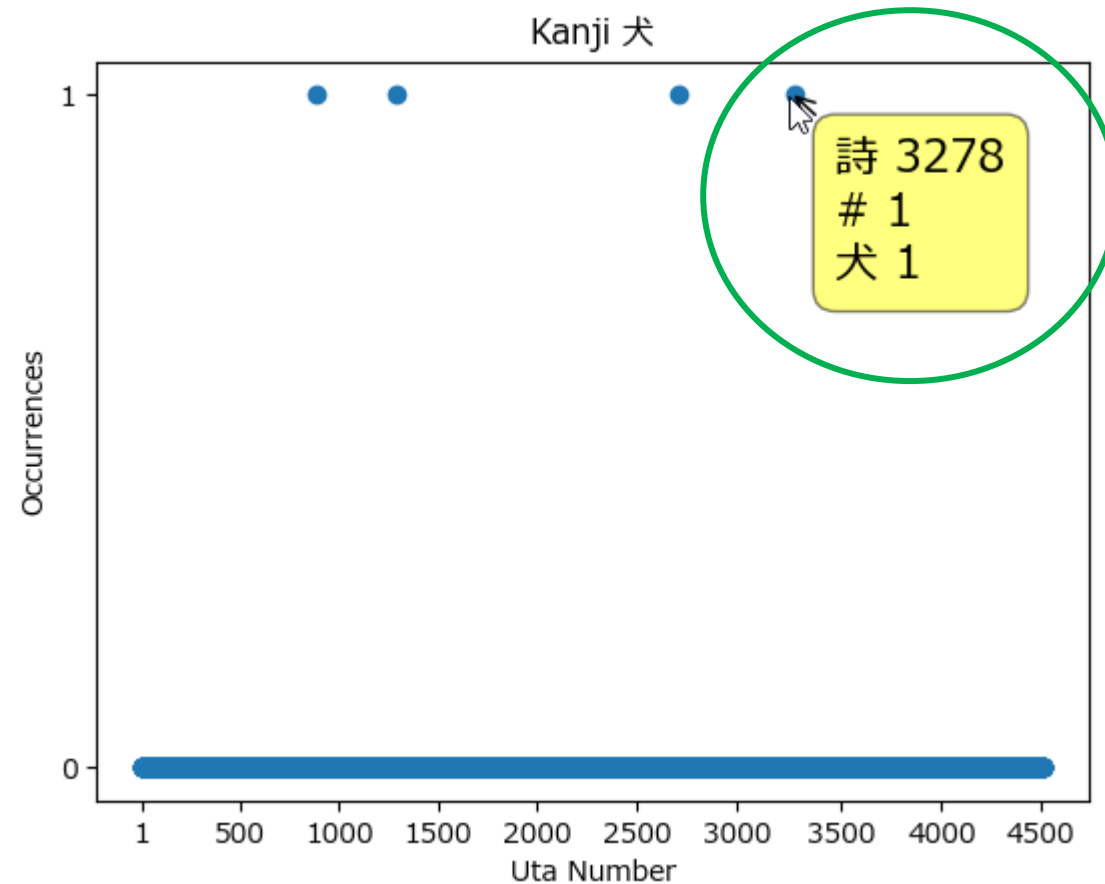
A toolkit to validate Japanese characters



# mplcursors



- Interactive data selection tooltips for Matplotlib



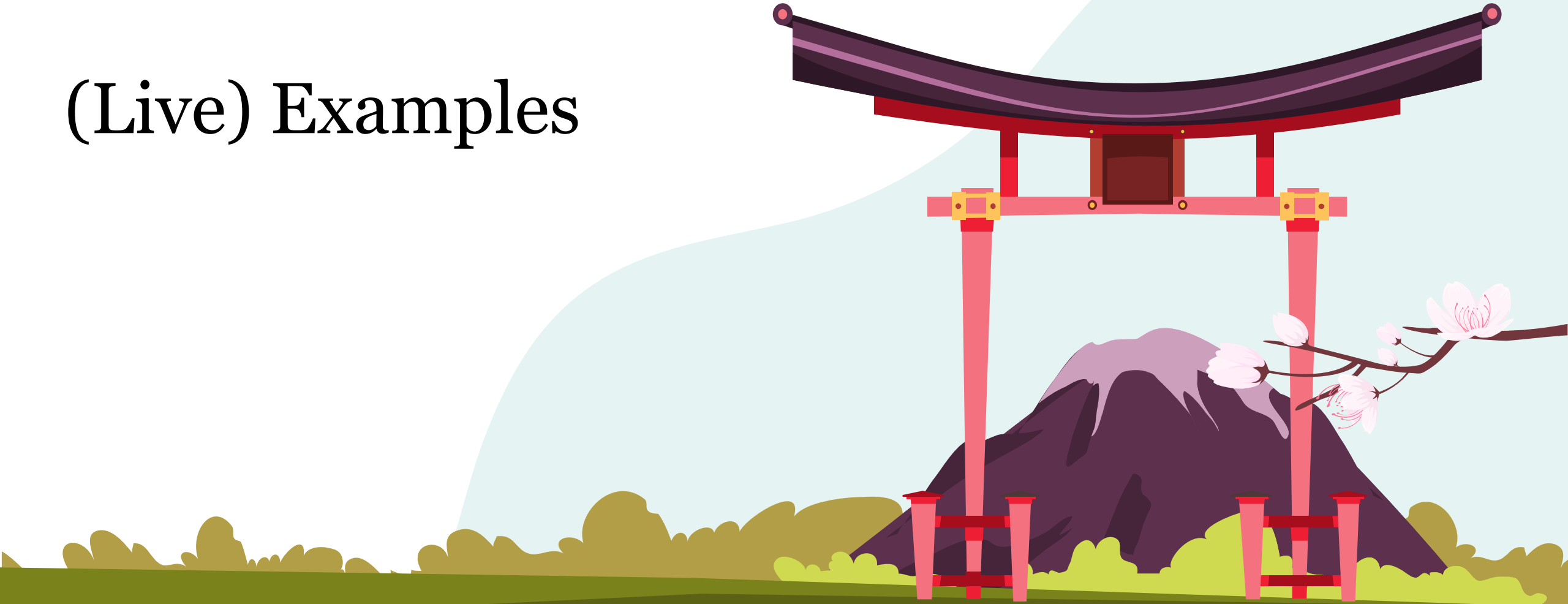


# Look at the Code





# (Live) Examples





# 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 <sup>WARE</sup>我 or <sup>WA GA</sup>我が
- 君 is virtually always used in <sup>KIMI</sup>君 or <sup>OO KIMI</sup>大君
  - 君 used to mean either “*you*” or “*ruler*”  
large, big
- *Kanji* connected with nature are generally used in many different words
  - <sup>YAMA KI</sup>山, <sup>TA</sup>木, <sup>HANA</sup>花, <sup>KUSA</sup>草, <sup>NO TORI</sup>野, <sup>HARU</sup>鳥, <sup>YORU</sup>春, 夜, ...



# Words containing 山



YAMA

山 (497)

YAMA BE

山辺 (31)

(SAN SEN)|(YAMA KAWA)

山川 (28)

YAMA MICH

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

- TODO
- + Zipf's law





# Order of Examples



- TODO: zadefinovat order v kteremu budou veci ukazovany
- Případně mít vždycky slide k exemplu a něco napsáno ohledně toho co má být na slidu vidět – co je výsledek



# Conclusion





# 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>



# References (2)



- <https://www.britannica.com/art/haiku>
- <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>



# References (Pictures) (1)



- [https://www.iconfinder.com/icons/6900234/alligator\\_crocodile\\_crocodile\\_icon\\_predator\\_icon](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/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](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.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\\_1589\\_061380.jpg](https://www.historyofcreativity.com/content/6/s_creator.001_1589_061380.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>

