Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary

Yamamoto Hilofumi
Tokyo Institute of Technology
yamagen@ila.titech.ac.jp

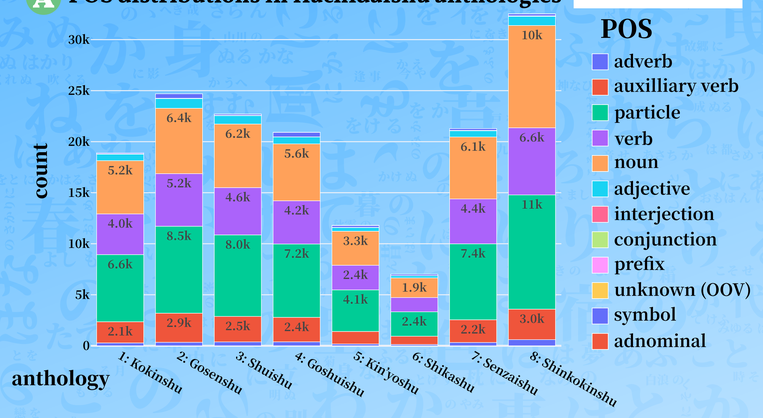
Hodošček Bor
Osaka University
bor@lang.osaka-u.ac.jp

We open sourced an annotated vocabulary of the Hachidaishū, a collection of eight classical Japanese anthologies (ca. 905-1205).

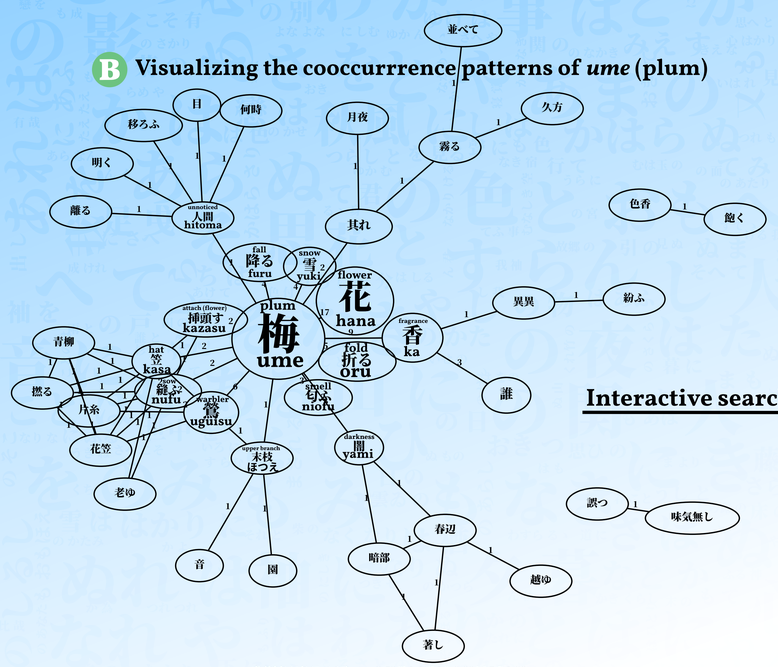
Originally developed in 2009, opening up the dataset was hampered by uncertanties in copyright status of the original documents. A superset of this dataset was released under the CC BY-SA 4 license by the ROIS-DS Center for Open Data in the Humanities (CODH), allowing us to publish it under an open licence. Here we explain the dataset format I 2 and showcase potential uses A B .

Example from notebook →

A **POS distributions in Hachidaishū anthologies**



B **Visualizing the cooccurrrence patterns of *ume* (plum)**



Interactive search site →

GitHub →

I **Poem per line, Part-of-Speech (POS) annotated dataset: `hachidaishu-pos.txt`**

anthology ID+poem ID

EXAMPLE

```
10001 年/名/とし の/格助/の 内/名/うち に/格助/に 春/名/はる は/係助/は ・・・
```

- 年/名/とし: tosi(year)/noun/toshi
- の/格助/の: no(of)/connecting particle/no
- 内/名/うち: uchi(inside)/noun/uchi
- に/格助/に: ni(indicates time)/case particle/ni
- 春/名/はる: haru(spring)/noun/haru
- は/係助/は: wa(topic)/binding case/wa

FORMAT: surface form/POS/reading **OR** surface form/POS:lemma-kanji:lemma-reading/reading (for verbs/adjectives (POS having conjugation))

```
いは/ハ四−未:言ふ:いふ/いは
```

2 **Fully annotated vocabulary in database format: `hachidai.db`** →

EXAMPLE/FORMAT

space-delimited format; database ID; Part-of-Speech (POS); word (morpheme) forms

```
01:000001:0001 A00 BG-01-1630-01-0100 02 年 年 とし 年 とし   year /tosi/
01:000001:0001 A10 BG-01-1911-03-1800 02 年 年 とし 年 とし
01:000001:0002 A00 BG-08-0061-07-0100 61 の の の の の        'of' particle /no/
                                                              a year /hitotose/
01:000001:0010 B00 BG-01-1950-14-0100 02 一とせ 一年 ひととせ 一年 ひととせ
01:000001:0010 C00 BG-01-1950-01-0300 19 一 一 いち 一 いち    one /ichi/
01:000001:0010 C01 BG-01-1630-01-0100 02 年 年 とし 年 とし   year /tosi/
01:000001:0020 A00 BG-02-3120-01-0100 47 いは 言ふ いふ 言は いは
01:000001:0021 A00 BG-03-3012-03-2600 74 ん む む む む
01:000001:0021 A10 BG-09-0010-02-0102 74 ん む む む む
```

possible semantic variants

possible decompositions/ semantic variants

- anthology ID
- poem ID
- token sequence ID
- token type
- WLSP semantic category (from *Bunruigoihyou*, floppy ed.) (Nakano et al., 1994)
- POS number
- surface form
- lemma (kanji)
- lemma (kana)
- conjugated form (kanji)
- conjugated form (kana)

zenodo
DOI 10.5281/zenodo.4744170