論文 / 著書情報
Article / Book Information

Title	Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary
Authors	Hilofumi Yamamoto, Bor Hodoscek
Citation	Proceedings of JADH conference, Vol. 2021, , pp. 82-87
Pub. date	2021, 9

Open source datasets of the Hachidaishū for the research of classical Japanese poetic vocabulary[1]

Hilofumi Yamamoto[2], Bor Hodošček[3]

1 Introduction

The present paper addresses the curation and publication of an open dataset on Zenodo (https://zenodo.org/) for classifying the vocabulary of the Hachidaishū (ca.905–1205). While the dataset was mainly developed in 2009 (Yamamoto 2009), it could not be published until now due to uncertainties in its copyright status. Even if the copyright of the classical text itself has expired, it is still under a reprint copyright which prevents publishing without a clear precedent. In 2016, the National Institute of Informatics (NII) Center of Open Data for the Humanities released the Nijūichidaishū, the Japanese classics dataset created by the National Institute of Japanese Literature (NIJL) under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license (Kitamoto 2017), which allows us to release our dataset.

Zenodo is a data repository for researchers to store their datasets, founded in 2013 (https://www.openaire.eu/zenodo-is-launched) by OpenAIRE (Open Access Infrastructure for Research in Europe) and CERN (European Organization for Nuclear Research). Researchers can upload up to 50GB per dataset, regardless of their research field, and can also cite code, datasets, or things relating to their research that are available on Github.

We will report on the publishing of the Hachidaishū vocabulary dataset, explain the ways in which the dataset will allow researchers to conduct semantic classification of words in the Hachidaishū, and succintly document the dataset containing part of speech tags and kanji readings.

2 Material

The Hachidaishū is a collection of eight anthologies of classical Japanese poetry compiled by the order of Emperors during the 300 years from the Kokinshū (ca.905), the first anthology written in Japanese kana characters, to the Shinkokinshū (1205).(Table 1) The main text is based on a collection of the Nijūichidaishū created by NIJL, and the text is now distributed by both NIJI and NII.(National Institute of Japanese Literature 2016) NIJI provides a poem string search service. NII provides a batch download service for all text

[1] This work was supported by KAKENHI, Grant-in-Aid for Scientific Research (C: 18K00528).
[2] Tokyo Institute of Technology
[3] Osaka University

data along with the original images using IIIF (International Image Interoperability Framework). The license for these texts is Creative Commons by SA 4.0 International. The conditions for redistribution are the same.

In addition to the main texts mentioned above, we use, as reference materials, the equivalent data to the Hachidaishu¯ texts contained in the CDROM of the Shimpen Kokka Taikan,(Shin-pen Kokkataikan Henshū Committee 1996) and those in the Nijūichidaishū database.(Nakamura et al. 1999). We use them as references to attach the readings of Kanji characters to each word.

As mentioned above, since the Hachidaishū consists of Japanese poems published across 300 years, and each collection is published approximately 20 to 80 years apart (42 year intervals on average), it is suitable for research into longitudinal changes in poetic vocabulary.

Table 1: The list of anthologies in the Hachidaishū: the number of poets in each anthlogy is based on Shimpen Kokka Taikan (Kokka Taikan Editorial Committee 1996).

	name	order	established	editors	poems
1	Kokin	Daigo tennō	ca. 905	Kino Tomonori, Kino Tsurayuki, Ōshikochino Mitsune, Mibuno Tadamine	1,100
2	Gosen	Murakami tennō	ca. 951	Kiyoharano Motosuke, Kino Tokifumi, Ōnakatomino Yoshinobu, Sakanoueno Mochiki Minamotono Shitagō	1,425
3	Shūi	Kazan'in	ca. 1007	Kazan'in	1,351
4	Goshūi	Shirakawa tennō	1086	Fujiwarano Michitoshi	1,218
5	Kin'yō	Shirakawain	ca. 1125	Minamotono Toshiyori	665
6	Shika	Sutokuin	ca. 1151	Fujiwarano Akisuke	415
7	Senzai	Goshirakawain	1188	Fujiwarano Toshinari	1,288
8	Shinkokin	Gotobain	1205	Minamotono Michitomo, Fujiwarano Ariie, Fujiwarano Ietaka, Fujiwarano Masatsune, Jakuren Fujiwarano Sadaie	1,978

3 Methods

First, we developed our own dictionary and system to divide the lines of poems of the Kokinshu into units (Yamamoto 2007). We did not accept conjunctions or compound verbs as valid part of speech categories, and adopted the shortest possible unit. In this way, we completed a dictionary that describes the words and concatenation rules of the Kokinshū. After that, the same dictionary was used to divide into units the subsequent collection, the Gosenshū. We checked the words and concatenation patterns that were included in the collection but were not included in the dictionary one by one, and added the missing words and patterns to the dictionary. The same process was repeated, and the dictionary was expanded from the Kokinshū to the Shinkokinshū according to the order of their establishment. In order to prevent the contents added to the dictionary from affecting the unit-divisions processed so far, and to maintain the consistency of the divided units, we

checked each poem processed. If any differences were found with the previously processed results, the differences were output, the content added to the dictionary was reviewed, and the consistency of the dictionary was adjusted. The Hachidaishū Part-of-Speech Dataset was created through the above process.

Because of the variety of notations in Japanese, a metacode was added to indicate the word meaning according to the form in which the word appears. The metacode indicating the lexical system is assigned by referring to the old version of the Word List by Semantic Principles (WLSP) by the National Institute for Japanese Language and Linguistics (NINJAL) (Nakano et al. 1994), and does not correspond to the new version (Asahara 2016, Kato et al. 2018). The classification metacodes are newly added since there are no metacodes for non-independent words such as particles, auxiliary verbs, conjunctions, and proper nouns such as place names (utamakura) and personal names as well as missing words in the WLSP.

4 Construction of two datasets

We will explain the construction of two datasets, the Hachidaishū vocabulary dataset (vocabulary dataset) and the Hachidaishū part-of-speech dataset (POS dataset). Table 2 indicates the construction of the vocabulary dataset.

We will explain the data offset with the first line in Table 2: [01:000001:0007 A00 BG-02-1527-01-0102 47 き 来 く 来 き]. A line consists of 7 columns separated by spaces. The first column ‘‘01:000001:0007’’ consists of 3 fields: 1) anthology ID as indicated in Table2, 2) number of poem, and 3) sequential ID of the token. ID 01 indicates the Kokinshū in this case. The second column indicates the type of token: type A is a single token; type B is a compound token; type C is a breakdown of type B. A00 indicates a single token; A01 indicates a single token and it has another meaning/metacode; B00 indicates a compound token; B01 indicates a compound token which has another meaning/metacode; C00 indicates the first element of the B00/B01.. breakdown; C01 indicates the second element of the B00/B01.. breakdown. The third column ‘‘BG-02-1527-01-0102’’: classification ID based on semantic categories according to WLSP.(Nakano et al. 1994) The fourth column indicates a POS number used in the morphological analysis system, Chasen.(Matsumoto et al. 2002) The fifth column indicates surface form: a form appears in literary works. The sixth column indicates lemma in kanji writing. The seventh column indicates lemma in kana writing. The eighth column indicates conjugated form in kanji. The nineth column indicates conjugated form in kana.

Table 2: Data structure of Hachidaishu vocabulary dataset; an example from the Kokinshu¯ Poem #1; left aligned tiny typefaces are for reference and not included in the dataset.

01:000001:0001 A00 BG-01-1630-01-0100 02 年 年 とし 年 とし	年=tosi (year) とし=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	の=no (particle)
01:000001:0003 A00 BG-01-1770-01-0300 02 内 内 うち 内 うち	内=uchi (inside), うち=uchi
01:000001:0004 A00 BG-08-0061-05-0100 61 に に に に に	に=ni (particle)
01:000001:0005 A00 BG-01-1624-02-0100 02 春 春 はる 春 はる	春=haru (spring), はる=haru
01:000001:0006 A00 BG-08-0065-07-0100 65 は は は は は	は=ha (particle)
01:000001:0007 A00 BG-02-1527-01-0102 47 き 来 く 来 き	き=ki (verb: come), 来 (kanji writing of き)
01:000001:0008 A00 BG-03-1200-02-0900 74 に ぬ ぬ に に	に=ni (auxiliary verb: perfect), ぬ=lemma of に
01:000001:0008 A10 BG-09-0010-01-0101 74 に ぬ ぬ に に	
01:000001:0008 A20 BG-09-0010-03-0200 74 に ぬ ぬ に に	
01:000001:0009 A00 BG-09-0010-04-0300 74 けり けり けり けり けり	けり=keri (auxiliary verb: past)
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	一年=hitotose (a year), 一とせ=hitotose
01:000001:0010 C00 BG-01-1950-01-0300 19 一 一 いち 一 いち	一=ichi (one), いち=ichi
01:000001:0010 C01 BG-01-1630-01-0100 02 年 年 とし 年 とし	年=tosi (year), とし=tosi
01:000001:0011 A00 BG-08-0061-10-0100 61 を を を を を	を=wo (particle)
01:000001:0012 A00 BG-01-1642-02-0100 02 こそ 去年 こぞ 去年 こぞ	去年=kozo (last year), こぞ=kozo
01:000001:0013 A00 BG-08-0061-04-0100 61 と と と と と	と=to (particle)
01:000001:0014 A00 BG-08-0065-14-0100 65 や や や や や	や=ya (particle)
01:000001:0015 A00 BG-02-3120-01-0100 47 いは 言ふ いふ 言は いは	言ふ=ifu (verb: say), いは=iha (predicative form)
01:000001:0016 A00 BG-03-3012-03-2600 74 ん む む む む	ん=n (colloquial form of mu), む=mu (auxiliary verb: inference)
01:000001:0016 A10 BG-09-0010-02-0102 74 ん む む む む	
01:000001:0017 B00 BG-01-1641-02-0100 02 ことし 今年 ことし 今年 ことし	今年=kotosi (this year), ことし=kotosi
01:000001:0017 C00 BG-03-1000-01-0100 57 この この この この この	
01:000001:0017 C01 BG-01-1630-01-0100 02 年 年 とし 年 とし	
01:000001:0018 A00 BG-08-0061-04-0100 61 と と と と と	
01:000001:0019 A00 BG-08-0065-14-0100 65 や や や や や	
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	

Table 3 indicates the construction of the POS dataset. We take the Kokinshū, Poem #1 as an example. It is a line, a poem. Tokens are separated by spaces. Each token consists of part-of-speech elements separated by slashes. The first column ”10001” contains two elements: the first digit indicates an anthology ID and the rest is a poem ID. The second column and the followings are the information of each token. In the case of nouns and particles, i.e., words that are not conjugated, they are shown in the following format: text/POS/reading. In the case of verbs and adjectives, i.e., words that are conjugated, they are shown in the following format: text/POS:lemma-kanji:lemma-reading/reading.

Table 3: Data structure of the Hachidaishu part-of-speech dataset; the first 5 digits are the anthlogy and poem ID; upper original; lower English translation; *its POS cannot be deterimined.

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 き/カ変-用:来:く/き に/完-用:ぬ:ぬ/に けり/過-終:けり:けり/けり ーとせ/名/ひととせ を/＊助/を こそ/名/こぞ と/格助/と や/係助/や いは/ハ四-未:言ふ:いふ/いは ん/推-終体:む:む/む ことし/名/ことし と/格助/と や/係助/や いは/ハ四-未:言ふ:いふ/いは ん/推-終体:む:む/む

10001 toshi(year)/noun/toshi no(of)/connecting_particle/no uchi(inside)/noun/uchi ni(indicates_time)/\
case_particle/ni haru(spring)/noun/haru wa(topic)/binding_case/wa ki(come)/kahen_conjugation-conjunctive:\
ku(lemma_kanji):ku(lemma_reading)/ki ni/perfect-conjunctive:nu:nu/ni keri(auxiliary_verb)/\
past-final:keri:keri/keri hitotose(a year)/noun/hitotose wo/case_particle/wo kozo(last year)/noun/kozo \
to/case_particle/to ya/binding_particle/ya iha(say)/yodan_verb-predicative:ifu:ifu/iha n(auxiliary_verb)/\
inference-final:mu:mu/mu kotosi(this year)/noun/kotosi to/case_particle/to ya/binding_particle/ya iha(say)/\
yodan_verb-predicative:ifu:ifu/iha n(auxiliary_verb)/inference-final:mu:mu/mu

5 How to use datasets and access the repository

The POS dataset allows researchers to count the number of tokens for each part of speech; count the number of poems in which the word or part of speech appears; obtain the sequence of words that appear; collect the patterns of the sequence of words; in the case of the vocabulary dataset, to count the number of words in each semantic category, extracting co-occurrence patterns; and so forth. A Jupyter notebook showing examples of Python code needed to conduct some of this is provided in the Github repository alongside the dataset.

Both the vocabulary and POS datasets can be obtained from the Zenodo repository (Yamamoto and Hodošček 2021a,b), and from Github (URL: url https://github.com/yamagen) as well.

There are advantages to using official repositories like Zenodo and Github: i.e., a DOI specific to the dataset is provided immediately; various bibliographic formats such as Mendeley, BiBTeX, etc. are available; a DOI clarifies the source of the data, ensuring that anyone can use the same dataset and verify the results.

The whole of the dataset is downloadable without user authentication and can be used in the user's preferred environment. Zenodo operates under a license where the data is intended to be downloaded and used. Since it is a portal site type repository, it may be disseminated faster than by publishing it on a personal site. As it is linked with Github, data can be updated and modified and the resulting changes inspected. Also, a DOI is given for each updated version.

6 Conclusion

We published two datasets for studying the Hachidaishū vocabulary on Zenodo and Github.We explained how they were created, their structure, how to use them, and introduced the URLs. The copyright issue has been cleared up, and the full text is now available and can be downloaded to promote research on classical Japanese poetic vocabulary at various user levels. For publications using these data, see Chen et al. (submitted to JADH2021). The datasets presented in the current paper are also licensed under the Creative Commons by SA 4.0 International.

References

- Asahara, Masayuki (2016) "Word List by Semantic Principles (WLSP): a collection of words classified and arranged by their meanings.", https://github.com/masayu-a/WLSP.
- Chen, Xudong, Hilofumi Yamamoto, and Bor Hodošček (submitted to JADH2021) "Token-based seman- tic vector space model for classic poetic Japanese", in *JADH2021 Proceedings of the 11th Conference of Japanese Association for Digital Humanities*.
- Kato, Sachi, Masayuki Asahara, and Makoto Yamazaki (2018) "Annotation of 'Word List by Semantic Principles' Labels for the Balanced Corpus of Contemporary Written

Japanese", in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong: Association for Computational Linguistics.

- Kitamoto, Asanobu (2017) "Center for Open Data in the Humanities (CODH): Activities and Future Plans".
- Kokka Taikan Editorial Committee ed. (1996) *Shimpen Kokka-taikan: CDROM Version*: Kadokawa Shoten.
- Matsumoto, Yuji, Akira Kitauchi, Tatsuo Yamashita, Osamu Imaichi, and Tomoaki Imamura (2002) *Morphological Analysis System ChaSen Version 2.2.9 Manual*, Nara Institute of Science and Technology.
- Nakamura, Yasuo, Yoshihiko Tachikawa, and Mayuko Sugita (1999) *Kokubungaku kenkyu¯ shiry¯okan d¯etab¯esu koten korekushon "Niju¯ichidaishu¯" Sh¯ohobanbon CD-ROM (Database Collection by National Institute of Japanese Literature "Niju¯ichidaishu¯" the Sh¯oho edition CD-ROM)*: Iwanami Shoten.
- Nakano, Hiroshi, Ooki Hayashi, Hisao Isii, Makoto Yamazaki, Masahiko Ishii, Yasuhiko Kato, Tatuo Miyazima, and Akio Tsuruoka (1994) *Bunrui goi hy¯o furopp¯i ban (Word List by Semantic Principles, floppy disk version)*, Vol. 5 of Kokuritsu Kokugo Kenkyu¯jo gengo shori data shu¯ (National Language Research Institute language data), Tokyo: Dainippon Tosho.
- National Institute of Japanese Literature (2016) "The Niju¯ichidaishu¯ Japanese Classics Dataset", http://codh.rois.ac.jp/pmjt/book/200007092/, http://kotenseki.nijl.ac.jp/biblio/200007092.
- Shin-pen Kokkataikan Henshu¯ Committee ed. (1996) *Shimpen Kokka-taikan: CDROM Version* : Kadokawa Shoten.
- Yamamoto, Hilofumi and Bor Hodošček (2021a) "Hachidaishu part of speech dataset", https://doi. org/10.5281/zenodo.4835806.
- Yamamoto, Hilofumi and Bor Hodošček (2021b) "Hachidaishu vocabulary dataset", https://doi.org/10.5281/zenodo.4744170.
- Yamamoto, Hilofumi (2007) "Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems", *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.
- Yamamoto, Hilofumi (2009) "Thesaurus for the Hachidaishu (ca.905–1205) with the classification codes based on semantic principles", *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 5, No. 1, pp. 46–52.