Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022), pages 31–37
Language Resources and Evaluation Conference (LREC 2022), Marseille, 25 June 2022

CHJ-WLSP: Annotation of 'Word List by Semantic Principles' Labels for the Corpus of Historical Japanese

Masayuki Asahara♣♮†, Nao Ikegami♢, Tai Suzuki♠‡, Taro Ichimura◯, Asuko Kondo♠, Sachi Kato♡, Makoto Yamazaki♣
♣National Institute for Japanese Language and Linguistics, ♮Tokyo University of Foreign Studies
♢Saitama University, ♠University of Tokyo, ◯ Kyoto Prefectural University, ♡ Mejiro University,
‡ Professor Emeritus at the University of Tokyo
† masayu-a@ninjal.ac.jp

Abstract

This article presents a word-sense annotation for the Corpus of Historical Japanese: a mashed-up Japanese lexicon based on the 'Word List by Semantic Principles' (WLSP). The WLSP is a large-scale Japanese thesaurus that includes 98,241 entries with syntactic and hierarchical semantic categories. The historical WLSP is also compiled for the words in ancient Japanese. We utilized a morpheme-word sense alignment table to extract all possible word sense candidates for each word appearing in the target corpus. Then, we manually disambiguated the word senses for 647,751 words in the texts from the 10th century to 1910.

Keywords: Historical Japanese, Word Sense Annotation

1. Introduction

The 'Corpus of Historical Japanese' (CHJ) (NINJAL, Japan, 2022) is a large-scale diachronic corpus based on the texts from the late 7th century to the early 20th century, which is a word-segmented and morphological information (POS) annotated corpus. The 'Word List by Semantic Principles' (WLSP) (NINJAL, Japan, 2004) is a large-scale Japanese thesaurus that includes 98,214 entries with syntactic and hiearchical semantic categories. The historical version of Word List by Semantic Principles (*Nihon Koten Taisho Bunrui Goihyo*, hWLSP) (Miyajima et al., 2014) is a thesaurus based on the old word senses of the vocabulary. These two language resources are compiled of the contemporary and historical words in the same word sense hierarchy.

This paper presents annotation of the WLSP/hWLSP sense labels for the Corpus of Historical Japanese. Annotating word senses (syntactic and semantic categories) for the historical corpus enables us to explore the historical changes in words. The distribution of syntactic and semantic categories shows the changes in writing styles and the difference in semantic contents. The word sense labels can also be used as the semantic index to search the texts. Furthermore, the word sense labels help the novice to read classical literature.

We present the annotation procedure and basic statistics. Section 2 presents the used language resources of WLSP and CHJ. Section 3 presents the annotation procedure with the goal. Section 4 presents the basic statistics of the label distributions. Section 5 is conclusions and our current issues.

2. Prerequisites

2.1. Word List by Semantic Principles

Word List by Semantic Principles (WLSP)[1] is one of the major thesauri for contemporary Japanese. The first version of the WLSP was released in 1964 by Kokuritsu Kokugo Kenkyusho, and a newer, expanded version was published in 2004 (NINJAL, Japan, 2004). Its comma-separated values (CSV) file of the expanded version can be used for research purposes.[2]

The data include more than 90,000 words with four syntactic categories (nominal words, verbal words, modifier words, and others) and several hierarchical semantic levels. The categories are indicated with one integer digit to the left of a radix point and four fractional digits to the right of the radix point. Table 1 shows an example of the word '昨年 (Last Year)', which is assigned a value of 1.1642. Here, the first '1. 体' presents the syntactic part **Class**, which is referred to as the 'Nominal Word', while '1642' presents the hierarchical semantic part, as follows: the first digit **Division**, '.1 関係', refers to the top-level semantic category 'Relation'; the two digits **Section** '.16 時間' refer to the second-level semantic category 'Time'; and the four digits **Article** '.1642 過去' refer to the finest-grained semantic category 'Past Time'. These five digits are therefore referred to as the **Article number**. The syntactic categories are 1. 体 Nominal Word, 2. 用 Verbal Word, 3. 相 Modifier Word, and 4. 他 Other (e.g., Conjunction, Interjection, Greeting). The semantic categories are .1 関係 Relation, .2 主体 Subject, .3 活動 Action, .4 生産物 Product, and 5. 自然 Nature. Though the thesaurus defines word senses for content words, the word senses for functional words and symbols are not defined in the WLSP. Furthermore, proper nouns

[1] `https://clrd.ninjal.ac.jp/goihyo.html`
[2] 200,000 yen (+ tax) for commercial use.

Table 1: Example Entry from the ‘Word List by Semantic Principles’

「昨年」 ‘Last Year’: 1.1642			
Syntactic Category	Semantic Category		
	Top Level	Second Level	Finest Level
Class	Division	Section	Article
体	関係	時間	過去
Nominal Word	Relation	Time	Past Time
1.	.1	.16	.1642

are not defined in the WLSP. Therefore, the functional words and proper nouns tend to be out-of-vocabulary. The historical version of WLSP defines the same word sense hierarchy as the contemporary version of WLSP for the ancient literature in Japan. Whereas the contemporary version of the article number is with a period (e.g., 1.1642), the historical version of the article number is without a period like (e.g., 11642).

2.2. Corpus of Historical Japanese

Corpus of Historical Japanese (CHJ) is a diachronic corpus from the Nara period to the Meiji and Taisho eras. This corpus enables advanced concordance search by annotating morpheme information to the sentences. It can be used online through the search system ‘Chunagon’[3] free of charge.

We annotated WLSP/hWLSP word sense labels (Article numbers) for the subset of CHJ. Table 2 shows the annotation target samples. Samples are the identifier for the literature in CHJ. Descriptions are the title in Japanese and their (literal) English translation. Year is the year of the establishment of the literature. Words are the word count in the annotated samples[4]. At the moment, we annotated 647,751 words from 11 samples.

3. Annotation Procedures

3.1. Goal

We show the goal of the large-scale word sense annotation procedure. Table 3 an annotation example of Taketori Monogatari. The pSample and pStart columns are the offset information in the CHJ. The corpus is word segmented and morphological information annotated. The table shows orthToken (surface form) and the lemma of the original corpus. Though space in the table did not permit us to insert the POSs for the words, the annotator can also see the POS labels and annotate the word sense labels in the Article Num. column. ‘野山’ (*hills*) is annotated the contemporary Article Number 1.5240. Class 1. is 体 (Nominal Word); Division .5 is 自然 (Nature); Section .52 is 天地 (Heaven and Earth); Article .5240 is 山野 (Hilly areas). 交じる (*goes into*) annotates the historical Article Number 21532[5]. Class 2 is 用 (Verbal Word); Division 1 is 関係 (Relation); Section 15 is 作用 (Interaction); Article 1532 is (Enter). Note that the functional words, symbols, and some of the proper nouns are not annotated.

3.2. Annotation Work Flow

Firstly, in order to establish the method of the large-scale word sense annotation on CHJ, we performed the word sense annotation on contemporary Japanese.

The BCCWJ and CHJ are word segmented and POS based on UniDic POS tagset[6]. We compiled the alignment table between WLSP and UniDic Lemma ID: WLSP2UniDic (Kondo and Tanaka, 2020) [7]. The WLSP2UniDic can be used as the word sense assigner with the morphological analyzer ChaMame (stand alone version) [8]. The word sense assigner annotates all possible word sense label candidates in WLSP for the UniDic lemmaID.

We performed the word sense annotation on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) with the word sense assigner. The annotators resolve word sense disambiguation for all possible word sense label candidates. If the word should be assigned other than the word sense label candidates, the annotators assign the most appropriate word sense label for the out-of-vocabulary sense. In this process, the annotator check the WLSP lookup tools of CradleExpress[9] in Section 7. As a result, we annotated 347,094 words and published as BCCWJ-WLSP (Kato et al., 2018)[10].

After finishing BCCWJ-WLSP, we performed the word sense annotation on the CHJ. Previously, Ikegami (Ikegami, 2017) annotated the WLSP labels for adjectives of the early middle Japanese based on hWLSP. We annotated 0900 竹取, 0934 土佐, 1212 方丈, 1336 徒然 samples by the same work flow of BCCWJ-WLSP. The

[3] https://chunagon.ninjal.ac.jp

[4] The annotation of 1642 虎明 is for not whole data.

[5] As we stated previously, the Article Number of hWLSP is without a period, whereas the one of WLSP is with a period.

[6] https://clrd.ninjal.ac.jp/unidic/en/

[7] https://github.com/masayu-a/WLSP2UniDic

[8] https://ja.osdn.net/projects/chaki/releases/p15635

[9] https://cradle.ninjal.ac.jp/wlsp

[10] https://github.com/masayu-a/BCCWJ-WLSP

Samples	Descriptions	Year	Words
0900 竹取	Taketori Monogatari (*lit. The Tale of the Bamboo Cutter*)	10th century	12,757
0934 土佐	Tosa Nikki (*lit. Tosa Diary*)	10th century	8,208
1100 今昔	Konjaku Monogatari-shu (*lit. Anthology of Tales from the Past*)	Heian period	175,598
1212 方丈	Hojoki (*lit. Square-jo Record*)	1212	5,402
1220 宇治	Uji Shui Monogatari (*lit. Gleanings from Uji Dainagon Monogatari*)	13th Century	120,705
1252 十訓	Jikkin-sho (*A Miscellany of Ten Maxims*)	1252	90,177
1336 徒然	Tsurezuregusa (*Essays in Idleness*)	ca. 1330	40,834
1642 虎明	Toraakira-bon Kyogen [a]	1642	5,448
1895 太陽	Taiyo *The Sun* (Magazine) [b]	1895	46,394
1904 小読	1st Jinjo Shogaku Tokuhon (Textbook) [c]	1904	45,334
1910 小読	2nd Jinjo Shogaku Tokuhon (Textbook)	1910	96,894
Total			647,751

[a] https://iss.ndl.go.jp/books/R100000002-I000008304623-00

[b] https://viaf.org/viaf/184683725/

[c] https://dglb01.ninjal.ac.jp/ninjaldl/bunken.php?title=kokutei1

Table 2: Annotation Targets

pSampleID	pStart	orthToken	lemma	Article Num.	Class	Class Label	Division	Division Label
20-竹取 0900_00001	250	野山	野山	1.5240	1	体	5	自然
20-竹取 0900_00001	270	に	に					
20-竹取 0900_00001	280	まじり	交じる	21532	2	用	1	関係
20-竹取 0900_00001	310	て	て					
20-竹取 0900_00001	320	竹	竹	1.5401	1	体	5	自然
20-竹取 0900_00001	330	を	を					
20-竹取 0900_00001	340	とり	取る	2.3811	2	用	3	活動
20-竹取 0900_00001	360	つつ	つつ					
20-竹取 0900_00001	380	、	、					

Translation: *While (the old man) goes into mountains and collects bamboos,*

Table 3: Annotation Example of Taketori Monogatari

annotator can see the translation from ancient Japanese to contemporary Japanese. Through the work, we compiled the alignment table between hWLSP and UniDic Lemma ID: WLSP2UniDic_historical [11].

Finally, we performed large-scale word sense annotation for the other samples in CHJ. The word sense candidates are extracted from both WLSP2UniDic and WLSP2UniDic_historical. The annotator resolved the polysemous words using the translation.

4. Statistics

4.1. Statistics: Syntactic Categories (Class)

Table 4 shows the basic statistics of syntactic categories (Class). We annotated 647,751 words in total. 353,890 words are 'Unlabelled', which are functional words, symbols, and proper nouns, since article numbers of these words are not defined in WLSP.

In order to explore the statistical biases of syntactic categories for the samples, we performed a chi-square test for the contingency table, excluding unlabelled word[12]. Table 5 shows the standardized residuals, which are measures of the strength of the difference between observed and expected values. The standardized residuals are standard and normal distribution (a mean of zero and a standard deviation of one). Therefore, when the absolute value of the statistics is over 1.96, the data shows the difference of the significance level 0.05[13].

Below, we confirm the statistical biases for syntactic categories. Concerning nominal words 1 . 体, the samples of 0900 竹取 and 1220 宇治 are small rates, and the samples of 1895 太陽 and 1910 小読 are large rates. Concerning verbal words 2. 用, the samples of 1895 太陽 and 1910 小読 are small rates, and the samples of 1100 今昔 and 1220 宇治 are large rates. The rates of nominal and verbal words are in complementary relation. Concerning modifier words 3. 相, the samples of 1252 十訓 and 1100 今昔 are small rates, and the sample of 1336 徒然 is large rates. Concerning other words 4. 他, the samples of 1100 今昔, 1220 方丈, and 1252 十訓 are small rates, and the samples of 1895 太陽 and

[11] https://github.com/masayu-a/WLSP2UniDic_historical

[12] We also performed a chi-square test for the data including labelled data. The tendencies of the statistical biases on the labelled word are nearly the same. However, the result with unlabelled data shows the tendencies of the statistical biases of functional words.

[13] However, we should consider p-value under multiple comparison correction.

	1. 体	2. 用	3. 相	4. 他	Unlabelled	Total
0900 竹取	2,318	2,252	706	72	7,409	12,757
0934 土佐	1,710	1,272	453	45	4,728	8,208
1100 今昔	40,687	29,498	8,518	1,189	95,706	175,598
1212 方丈	1,433	792	342	100	2,735	5,402
1220 宇治	24,214	21,336	6,290	716	68,149	120,705
1252 十訓	19,808	13,039	3,974	460	52,896	90,177
1336 徒然	8,876	6,138	2,688	213	22,919	40,834
1642 虎明	1,255	811	369	88	2,925	5,448
1895 太陽	13,256	6,131	3,116	853	23,038	46,394
1904 小読	10,846	5,312	2,620	794	25,762	45,334
1910 小読	28,915	12,833	6,388	1,135	47,623	96,894
Total	153,318	99,414	35,464	5,665	353,890	647,751

Table 4: Basic Statistics: Syntactic Categories (Class)

	1. 体	2. 用	3. 相	4. 他
0900 竹取	-13.05	12.91	2.57	-3.12
0934 土佐	-3.62	3.40	1.77	-2.74
1100 今昔	-8.26	21.65	-14.30	-10.59
1212 方丈	1.62	-4.53	1.20	6.87
1220 宇治	-30.90	36.19	-0.78	-10.40
1252 十訓	3.96	5.00	-8.94	-10.43
1336 徒然	-7.27	1.26	12.45	-7.42
1642 虎明	-2.45	-1.80	3.96	5.72
1895 太陽	14.61	-25.52	6.22	19.98
1904 小読	9.40	-20.47	5.86	22.42
1910 小読	31.72	-40.03	6.70	6.65

Table 5: Chi-square Test: Syntactic Categories Excluding Unlabelled Words

1904 小読 are large rates.

The appendix 8 includes the distances of syntactic categories evaluation. The figure 2 shows the distances among the samples. The result shows that the neighbouring sample pairs in chronological order are smaller distances than other pairs.

4.2. Statistics: Semantic Categories (Division)

Table 6 shows the basic statistics of semantic categories (Division). The division labels of .2 主体 and .4 生産物 are only defiled in the class 1. 体. So, these two labels are relatively small.

We explore the statistical biases of semantic categories for the samples by chi-square test for the contingency table excluding unlabelled word. Table 7 shows the standardized residuals of chi-squared test. Concerning .3 活動, the samples of 1252 十訓 and 1336 徒然 are large rates, and the samples of 1904 小読 and 1910 小読 are small rates. We regard it as the correlation of the rate 2. 用 in the syntactic category. Concerning .4 生産物, the samples of 1904 小読 and 1910 小読 are large rates, and the sample of 1895 太陽 is a small rate. Concerning .5 自然, the samples of 1904 小読 and 1910 小読 are large rates, and the samples of 1100 今昔 and 1895 太陽 are small rates.

The results of samples around 1900 (1895 太陽, 1904 小読, and 1910 小読) shows that the distributions of semantic category do not show synchronic similarities in the same era. The difference in genres (magazines vs. textbook) are observed in the difference in the distributions of semantic category. The appendix 8 includes the distances of semantic categories evaluation. The figure 3 shows the distances among the samples. The result shows that the neighbouring sample pairs in the chronological order are not smaller distances as the syntactic category distance.

5. Conclusions

This study presents large-scale word sense label annotation on the Corpus of Historical Japanese. We presented the annotation work flow and the basic statistics of the results. The data will publish via `https://github.com/masayu-a/CHJ-WLSP` as the stand-off annotation format [14], and also are shared for the applicant of NINJAL Joint Usage Projects (NINJAL language resources).

Below, we present the current issues of the CHJ-WLSP.

Word sense label granularity: Though the word sense label design is based on WLSP/hWLSP, the granularity of the word sense is limited. For example, the word 'い

[14] Excluding the surface form and lemma from the Table 3.

	.1 関係	.2 主体	.3 活動	.4 生産物	.5 自然	Unlabelled	Total
0900 竹取	2,335	577	1,722	229	485	7,409	12,757
0934 土佐	1,540	360	1,002	167	411	4,728	8,208
1100 今昔	37,538	12,221	21,942	3,615	4,576	95,706	175,598
1212 方丈	1,335	255	637	147	293	2,735	5,402
1220 宇治	24,924	6,726	14,857	2,661	3,388	68,149	120,705
1252 十訓	16,323	5,461	11,789	1,540	2,168	52,896	90,177
1336 徒然	8,279	2,011	5,710	714	1,201	22,919	40,834
1642 虎明	1,122	389	667	113	232	2,925	5,448
1895 太陽	11,403	3,324	6,888	609	1,132	23,038	46,394
1904 小読	8,915	2,960	4,107	1,363	2,227	25,762	45,334
1910 小読	23,521	6,426	10,881	3,025	5,418	47,623	96,894
Total	137,235	40,710	80,202	14,183	21,531	353,890	647,751

Table 6: Basic Statistics: Semantic Categories (Division)

	.1 関係	.2 主体	.3 活動	.4 生産物	.5 自然
0900 竹取	-4.50	-6.55	8.13	-1.87	4.93
0934 土佐	-2.91	-6.03	2.00	-0.08	10.21
1100 今昔	1.89	13.84	1.28	-4.66	-20.33
1212 方丈	3.49	-6.45	-3.97	1.66	7.29
1220 宇治	3.67	-7.73	5.55	2.79	-8.55
1252 十訓	-12.08	4.75	20.08	-6.71	-11.99
1336 徒然	-1.35	-10.51	14.20	-5.42	-3.30
1642 虎明	-2.25	2.28	-0.97	-0.82	3.62
1895 太陽	6.77	1.74	7.86	-16.49	-15.16
1904 小読	-3.34	5.32	-20.51	14.44	22.51
1910 小読	5.06	-5.71	-28.45	14.91	34.26

Table 7: Chi-square Test: Semantic Categories Excluding Unlabelled Words

みじ (lemma: いみじい)' (*extreme*) is assigned the article number 31920 (相-関係-量-程度-程度: Modifier-Relation-Degree-Degree). The word can be used in both positive and negative contexts. The polarity of the word sense is not encoded in the WLSP article number. In order to explore more deep linguistic research, we need to introduce more fine-grained word sense labels. Contextual word embedding techniques might introduce the more fine-grained word sense definition. We need to perform the comparison between the vector spaces of word embeddings and human judgement.

Word unit in Japanese: This work is based on the word delimitation of Short Unit Word (SUD) by NINJAL. The other word delimitation is Long Unit Word (LUW) by NINJAL, which defines the base phrase (文節 Bunsetsu) in Japanese. In some cases, the compound word of LUW cannot be composed by their constituents of SUW word senses. We annotate the LUW word sense labels for 1100 今昔 and 1220 宇治 samples. However, we have not organised language resources.

Balanced Sampling: The word sense label annotation for the ancient languages is quite difficult task. The work should be done by an expert in the literature of that era. We select the target samples when we can hire an expert in the literature. Therefore, the sampling of CHJ-WLSP is not balanced.

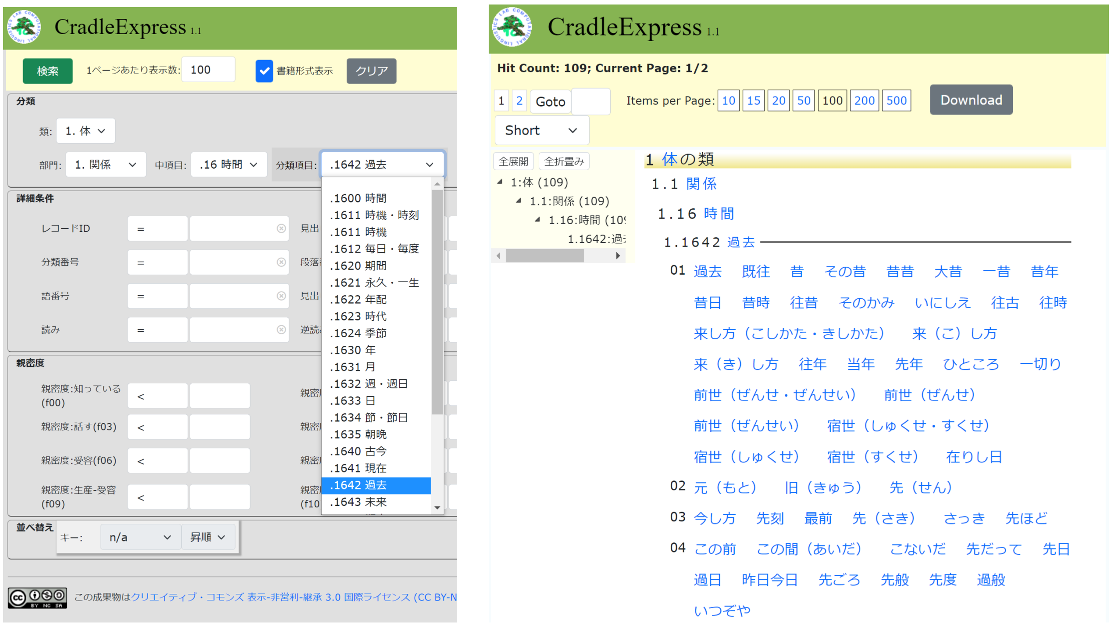
All word WSD: We still have several issues with the manual annotation procedures. The work is very time-consuming. Nevertheless, the constructed historical language resource size is 647,751 words. Moreover, we constructed 347,094 words of word sense labelled data on the contemporary language resources. We can use around one million word sense labelled data. It might be enough for training all word WSD (word sense disambiguation) tools. The tools enable us to reduce the manual annotation cost.

6. Acknowledgments

This work is partially presented in poster session of Japanese domestic conference of The Society for Japanese Linguistics on May 2022. We thank Mr. Hara, Ms. Sogo, and Ms. Nagumo for their assistance with the annotation, and NINJAL colleagues for comments that greatly improved the manuscript. This research was supported by JSPS KAKENHI Grant Numbers JP17H00917(2017-2021), NINJAL Project of Center for Corpus Development (2016-2021), and NINJAL Project 'Evidence-based Computational Psycholinguistics Using Annotation Data' (2022-2027).

7. Appendix: CredleExpress

CredleExpress is a lexicon viewer for WLSP. Figure 1 shows the form. The left figure shows the query.


CradleExpress 1.1

検索 1ページあたり表示数: 100 ☑ 書籍形式表示 クリア

分類

類: 1. 体

部門: 1. 関係 中項目: .16 時間 分類項目: .1642 過去

.1600 時間
.1611 時機・時刻
.1611 時機
.1612 毎日・毎度
.1620 期間
.1621 永久・一生
.1622 年配
.1623 時代
.1624 季節
.1630 年
.1631 月
.1632 週・週日
.1633 日
.1634 節・節日
.1635 朝晩
.1640 古今
.1641 現在
.1642 過去
.1643 未来

詳細条件

レコードID =
分類番号 =
語番号 =
読み =

親密度

親密度:知っている(f00) <
親密度:話す(f03) <
親密度:受容(f06) <
親密度:生産-受容(f09) <

並べ替え キー: n/a 昇順

この成果物はクリエイティブ・コモンズ 表示-非営利-継承 3.0 国際ライセンス (CC BY-N

CradleExpress 1.1

Hit Count: 109; Current Page: 1/2

1 2 Goto Items per Page: 10 15 20 50 100 200 500 Download

Short

全展開 全折畳み

1:体 (109)
1.1:関係 (109)
1.16:時間 (10
1.1642:過

1 体の類

1.1 関係

1.16 時間

1.1642 過去

01 過去 既往 昔 その昔 昔昔 大昔 一昔 昔年 昔日 昔時 往昔 そのかみ いにしえ 往古 往時 来し方（こしかた・きしかた） 来（こ）し方 来（き）し方 往年 当年 先年 ひところ 一切り 前世（ぜんせ・ぜんせい） 前世（ぜんせ） 前世（ぜんせい） 宿世（しゅくせ・すくせ） 宿世（しゅくせ） 宿世（すくせ） 在りし日

02 元（もと） 旧（きゅう） 先（せん）

03 今し方 先刻 最前 先（さき） さっき 先ほど

04 この前 この間（あいだ） こないだ 先だって 先日 過日 昨日今日 先ごろ 先般 先度 過般 いつぞや

https://cradle.ninjal.ac.jp/wlsp/

Figure 1: CradleExpress

We can choose the syntactic and semantic categories. The right figure shows the query results. By clicking a word in the results, the viewer shows further information about the word.

8. Appendix: Similarities of Syntactic/Semantic Categories among samples

Figure 2 and 3 shows the distances of syntactic and semantic category distributions. The distances are evaluated by the R dist method with euclidean (2-norm) of the frequency vectors of samples. The figures are plotted by corrplot. The larger (blue) circles are longer distanced pairs.

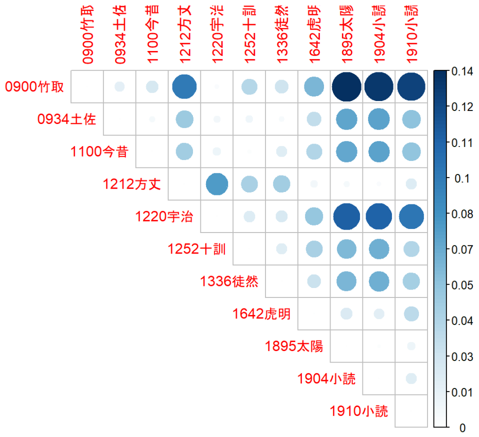


Figure 2: The Distances of Syntactic Category Distributions

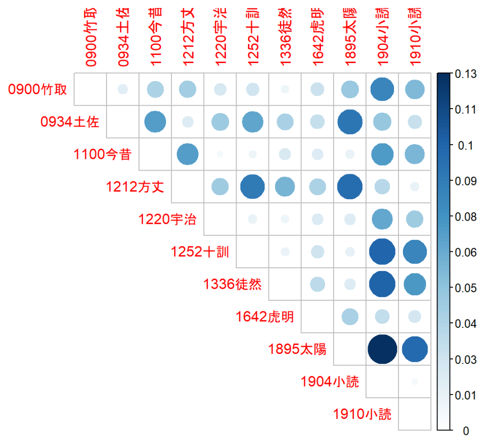


Figure 3: The Distances of Semantic Category Distributions

9. Bibliographical References

Ikegami, N. (2017). Nihongo Rekishi Kopasu Heian-jidai-hen Shutsugen Keiyoushi ni-taisuru Koten Bunruigoihyou Bango Anoteshon (in Japanese) (*lit. hWLSP article number annotations on adjectives in the Corpus of Historical Japanese Heian Period Edition*). In *The 23rd Annual Meeting of the Association for Natural Language Processing, Japan*, pages 310–313.

Kato, S., Asahara, M., and Yamazaki, M. (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*.

Kondo, A. and Tanaka, M. (2020). Construction of an Alignment Table between 'Word List by Semantic Principles' and UniDic. *NINJAL Research Paper*, (18):77–91.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

Miyajima, T., Ishii, H., Abe, S., and Suzuki, T. (2014). *Nihon Koten Taisho Bunrui Goihyo*. Kasama-shoin.

NINJAL, Japan, editor. (2004). *Word List by Semantic Principles, - Revised and Enlarged Edition*. Dainippon-tosho.

NINJAL, Japan. (2022). Corpus of Historical Japanese.