Annotation of 'Word List by Semantic Principles' Labels for the Balanced Corpus of Contemporary Written Japanese

Sachi Kato **Masayuki Asahara** **Makoto Yamazaki**
National Institute for Japanese Language and Linguistics, Japan
Tokyo, Japan
{yasuda-s, masayu-a, yamazaki} AT ninjal DOT ac DOT jp

Abstract

This article presents a word-sense annotation for the Balanced Corpus of Contemporary Written Japanese: a mashed-up Japanese lexicon based on the 'Word List by Semantic Principles' (WLSP). The WLSP is a large-scale Japanese thesaurus which includes 98,241 entries with syntactic and hierarchical semantic categories. 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 182,166 content words in the texts.

1 Introduction

Semantic information annotation is an important linguistic resource to explore synonyms by semantic category or figurative expressions by discrepancies in the co-occurrence of semantic categories. These annotations can also be used as training and evaluation data for word-sense disambiguation tasks. Among Japanese language resources, the EDR corpus (Yokoi, 1995) and the RWCP corpus (Toyoura et al., 1996) include word-sense information based on the gloss of the dictionaries. The Balanced Corpus of Contemporary Written Japanese (hereafter BCCWJ) (Maekawa et al., 2014) was compiled using a sampling method that preserved representativeness of actual usage. Word-sense information from the Iwanami dictionary is annotated in a subset of the BCCWJ, and the data are utilized in the SemEval-2010 Japanese WSD Task (Okumura et al., 2010). Several all-word sense disambiguation methods are proposed in the benchmark data. A thesaurus-based word-sense tagged corpus has also been proposed. (Bond et al., 2012) developed annotation data for Japanese WordNet, which is a translation of English data.

In this study, we develop a new semantic information annotation of BCCWJ. The semantic information is based on the thesaurus 'Word List by Semantic Principles, Revised and Enlarged Version' (hereafter WLSP) (Kokuritsu Kokugo Kenkyusho, 2004). This article presents the design of the annotation and basic statistics of the data. We use a subset of the core data of BCCWJ as the target corpus for word-sense annotations. The data include books (BCCWJ sample ID: PB), magazines (PM), and newspapers (PN). The BCCWJ has morpheme information annotations such as word boundary and part-of-speech annotations, based on a morphological analyzer dictionary, UniDic. (Kondo et al., 2018) developed an alignment table between UniDic morpheme information and WLSP word-sense labels. We extract all possible word-sense candidates for the text by the alignment table. We manually disambiguate the word senses of all content words by their contextual information. When the word-sense candidates are not appropriate in the context, we newly assign the WLSP word-sense label manually. The data include 347,094 morphemes making up 182,166 content words and 164,928 function words.

The rest of the paper is organized as follows. Section 2 presents how we annotated the WLSP word senses on BCCWJ, employing the language resources of BCCWJ, WLSP, and alignment table. Section 3 presents the basic statistics of the devel-

oped language resource. Section 4 presents a conclusion and future directions.

2 Annotation Procedures

2.1 Thesaurus: WLSP

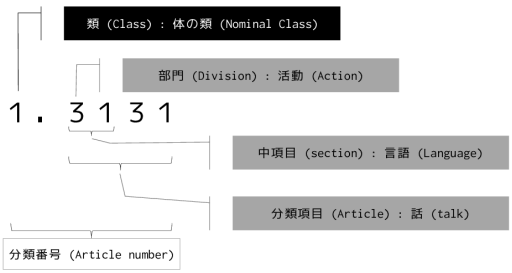

Figure 1: Structure of Word-Sense Labels in WLSP

We present the structure of the WLSP database (Kokuritsu Kokugo Kenkyusho, 2004). The WLSP was a pioneer among Japanese thesauri, and was first published in 1964 (Kokuritsu Kokugo Kenkyusho, 1964). The WLSP assigns a 5-digit article number for each lexical entry, which indicates a syntactic category and a hierarchical semantic category. Figure 1 indicates an example for '話' (talk).

The first digit represents the syntactic category '類' (class), which consists of four sub-categories, as follows:

- 1 体: nominal class
- 2 用: verbal class
- 3 相: modifier class
- 4 他: other (interjection, conjunction)

The numbers after the decimal point represent the hierarchical semantic category. The first decimal place represents '部門' (division), which has five sub-categories, as follows:

- .1 関係: relation
- .2 主体: subject
- .3 活動: action
- .4 生産物: product
- .5 自然: nature

The first and second decimal places represent a '中項目' (section) of 43 labels, and the four digits represent an '分類項目' (article) of 519 labels.

The WLSP also has more fine-grained information, such as '段落番号' (paragraph number), '小段落番号' (small paragraph number), and '語番号' (word number). There is a total of 98,241 entries registered in the WLSP. In the case of polysemous words, each sense is registered as a separate entry in the WLSP.

Table 1 shows examples of WLSP entries. These four pieces of information article number, paragraph number, small paragraph number, and word number can identify each entry in the WLSP.

2.2 Target Data: BCCWJ

The BCCWJ (Maekawa et al., 2014)[1] constitutes the target data for the annotation. The BCCWJ has a million words of core data, which are sourced from books (PB), magazines (PM), newspapers (PN), white papers (OW), Yahoo! Answers (OC), and Yahoo! Blogs (OY). The annotation priority is defined for the data. The core data are analyzed by two kinds of word units, which are 'Short Unit Words' (SUWs) and 'Long Unit Words' (LUWs) with UniDic part-of-speech (PoS) tag sets. In the present study, we annotate PN, PB, and PM samples in that order of annotation priority, from A to E. We have finished PB(A), PB(B), PM(A), PM(B), PN(A) and PN(B) samples based on this annotation priority on the SUW word delimitation.

The understanding of parts-of-speech in Japanese corpora can be split into two philosophies: lexicon-based (語彙主義) and usage-based (用法主義). The lexicon-based approach involves extracting all possible categories for one word as labels. For example, the label '名詞-普通名詞-サ変形状詞可能' means that the word can be a noun, verbal noun, or adjective. The labels are maintained in a large-scale, PoS-tagged lexicon and are used in semi-Markov-model-based morphological analysers. Usage-based labelling, in contrast, is determined by the contextual information in a given sentence. While the PoSs of the SUWs in the BCCWJ are lexicon-based, the PoSs of the LUWs are usage-based. The PoS tagset is called UniDic PoS. UniDic is a morphological an-

[1] pj.ninjal.ac.jp/corpus_center/bccwj/en/

Table 1: Entries of WLSP

class 類	division 部門	section 中項目	article 分類項目	article number 分類番号	paragraph number 段落番号	small paragraph number 小段落番号	word number 語番号	word 見出し	reading 読み
体 Nominal	活動 Action	言語 Language	話・談話 Talk	1.3131	1	1	1	話	はなし
体	活動	言語	話・談話	1.3131	1	1	2	話	わ
体	活動	言語	話・談話	1.3131	1	1	2	トーク	とおく
					...				
体 Nominal	活動 Action	言語 Language	問答 Q/A	1.3132	1	1	1	問答	問答

alyzer dictionary with around 400,000 SUW word entries. UniDic includes PoS, conjugation, pronunciation, and lemma information.

2.3 Alignment Table between UniDic and WLSP

(Kondo et al., 2018) developed an alignment table between UniDic and the WLSP[2]. The UniDic lemma ID is aligned with the WLSP article number, that is, the word-sense label in our annotation. We use the alignment data to extract all possible word senses for both SUWs and LUWs. The table represents many-to-many relationships, in which many occurrences in an entry relate to many occurrences in another entity. Table 2 shows the alignment table. 'BunruiNumber' is the WLSP article number, the WLSP label, and the full WLSP number[3]. 'Lemma ID' is the morpheme identifier in UniDic. BCCWJ is assigned the Lemma ID for all morpheme entries. Therefore, the alignment table enables us to extract all possible word senses by the WLSP article numbers.

2.4 Annotation Procedures (SUWs)

The annotator chooses the most possible (most appropriate) word sense for the target content word based on the contextual information. When no WLSP article number can be assigned by the alignment table, we manually annotate the article number by checking the category hierarchy.

We present the annotation procedures for the SUWs. The target words are all content words from the corpus. The annotator chooses the most possible senses from the spreadsheet, as shown in Figure 2,

Figure 2: Worksheet for Annotator (SUW part)

from the automatically assigned word-sense candidates (highlighted in the figure).

The number of word-sense ambiguities for the content words is presented in Table 3 and 4, by tokens and types, respectively. Ambiguous words with more than one sense total 77,344 of the 182,166 tokens (42.45%). Note that the high frequency of ambiguity 8 is because of the verb 'する' (do), which is the most frequently used verb in Japanese.

We have not annotated function words in the data, even if they are defined in the WLSP. However, the list of function words in the WLSP is limited, as shown in Table 5.

The sense selection for SUWs is based on the least contextual information. The etymological sense is chosen for metaphorical or collocational expressions, if the metaphorical or collocational sense is not defined in the WLSP. The metaphorical and collocational senses are resolved in the LUW annotation. The morphological information for BCCWJ SUWs is lexeme based. For example, the PoS 「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可

[2] https://github.com/masayu-a/wlsp2unidic

[3] The article number with paragraph number, small paragraph Number, and word number.

Table 2: Alignment Table between UniDic and WLSP

BunruiNumber	Lemma ID
1.3131, 体-活動-言語-話・談話,1.3131-01-01-01	29980
1.3131, 体-活動-言語-話・談話,1.3131-01-01-02	41106
1.3131, 体-活動-言語-話・談話,1.3131-01-01-02	319033

Table 3: Ambiguities by Tokens

# of Senses	PB	PM	PN	ALL
0	3,996	5,891	5,920	15,807
1	25,972	28,472	34,571	89,015
2	12,900	14,416	15,361	42,677
3	5,199	5,511	4,732	15,442
4	2,212	2,339	2,326	6,877
5	468	520	422	1410
6	794	623	545	1932
7	430	361	273	1064
8	2,254	2,403	2,613	7,270
9	97	103	90	290
10	81	71	32	184
11	31	9	1	41
12	70	67	20	157
≥ 2	24,539	26,423	26,415	77,344
Total	54,504	60,786	66,906	182,166

Table 4: Ambiguities by Types

# of Senses	PB	PM	PN	ALL
0	1,267	2,071	2,487	5,131
1	5,199	5,448	5,600	10,306
2	1,919	2,024	1,898	3,296
3	590	595	583	869
4	195	204	186	261
5	66	55	58	80
6	31	29	26	34
7	20	18	17	21
8	6	5	5	6
9	3	3	3	3
10	1	1	1	1
11	1	1	1	1
12	1	1	1	1
≥ 2	2,833	2,936	2,779	4,573
Total	9,299	10,455	10,866	20,010

能」 can be ‘1 体 Nominal’ and ‘3 相 Modifier’ in the WLSP syntactic category. The ambiguity of the lexeme-based PoS is resolved as ‘usage’ information in the BCCWJ.

When the entry does not appear in the alignment table between the UniDic lexeme ID and the WLSP category, we newly assign the article number for the target entry. In all, 15,807 tokens (8.67%) are not assigned any word-sense candidates. When the entry does not appear in the UniDic lexeme (that is, when it is an unknown word for UniDic), we again newly assign the article number to the target. The examples, which do not appear in the UniDic lexeme, are unknown words, proper nouns, and abbreviations. The unknown words include foreign words such as「ロック」(rock music),「カム」(come), and「トゥゲザー」 (together). In these cases, we newly assign the article number for the word. Note that some words need to define undefined article numbers in the WLSP because of their syntactic category. In such cases, we introduce a new article number for the word.

Let us explain this in further detail. We do not assign the article number for a person’s name. However, we do annotate the article number of the constituents of location or organization names, such as「名古屋/タワー/プラザ/ホール」(Nagoya Tower Plaza Hall). This example is assigned an article number for each SUW of「名古屋」(Nagoya),「タワー」(Tower),「プラザ」(Plaza), and「ホール」(Hall). Abbreviations are extracted from the original forms, and their etymological senses are annotated. In the abbreviated words, coordinations of more than one word appear, such as「厚労」((Ministry of Health,) Labour and Welfare) by「厚生」(Welfare)「労働」(Labour) and「自民」(the Liberal Democratic (Party)) by「自由」(Liberal)「民主」 (Democratic). In such cases, we annotate all senses of each constituent. Paronomasias or puns are also annotated as multiple senses.

Table 5: Functional Words in WLSP

Lexeme	PoS	description
れる	Aux V	passive
など	Adv PostP	etc.
まで	Adv PostP	goal
られる	Aux V	passive
せる	Aux V	causative
たい	Aux V	desire
だけ	Adv PostP	only
たり	Adv PostP	perfect
くらい	Adv PostP	degree
ほど	Adv PostP	degree
ばかり	Adv PostP	degree
ずつ	Adv PostP	each
のみ	Adv PostP	only
つ	Adv PostP	perfect
させる	Aux V	causative

2.5 Annotation Procedures (LUWs)

We also annotate article numbers for LUWs. The annotation procedure for LUWs is nearly the same as that of SUWs. When the entry is registered in the alignment table between the UniDic lexeme ID and the WLSP, we just choose one sense among the possible senses. However, in the case of LUWs, most of the tokens are not registered in the alignment table. In such cases, we newly introduce the article number for the entry.

Multiword expressions of function words appear in the LUWs, such as「ていく」(*te-iku*),「てくる」(*te-kuru*), and「にとって」(*ni-totte*). These words have different senses from the SUW constituents. We annotate the article numbers for these function words in the LUWs. When collocational expressions appear, we annotate their article numbers in the longer unit.

3 Basic Statistics

This section presents the basic statistics for the annotations. All the information is based on SUW annotation.

Table 6 shows the rates of syntactic categories (classes) in the three registers. In the book (PB) register, whereas 1. nominal class rate (PB: 55.01%) is smaller than other registers (PM: 62.81%, PN: 73.53%), 2. verbal class rate (26.53%) is larger than others (PM: 21.51%, PN 16.40%). In the newspaper (PN) register, 3. modifier class rate (PN: 6.65%) is smaller than others (PB: 13.18%, PM 11.43%).

Table 7 shows the rate of the top semantic category (class) in each of the three registers. The variance of semantic category rates is smaller than that of syntactic category rates. Still, in the PN (Newspaper) samples, the Subject (.2) is larger rate, and the Nature (.5) is smaller rate than other registers.

Table 8 shows the labelled ‘out of vocabulary’ (OOV) words in the alignment table. OOV lexemes are those where although the lexeme is not registered in the alignment table, the WLSP label is assigned in the corpus by the annotators. Most OOV lexemes are nominals: 9,040 tokens and 3,651 types. In addition, some OOV words are not assigned (5,304 words), since they are substrings of longer named entities, symbols, and collocations. ‘OOV senses’ are those where although the lexeme is registered and the sense is not registered in the alignment table, the WLSP label is assigned in the corpus by the annotators. Most OOV senses are also nominals: 2,133 tokens and 647 types.

Table 9 shows the top 5 frequent article numbers in each of the three registers. Books include large portion of action verbs and existential relations. Magazines and newspapers include large portions of numeral expressions and numeral suffixes.

4 Conclusions

In this study, we present a word-sense-annotated corpus based on the WLSP thesaurus. The annotation speed depends on the annotator and samples, but is very roughly 100-300 words per hour. It has taken around 2 years’ annotation work to get to the state of the work presented here, since 2016.

Our future work will proceed as follows. First, we will explore writing styles among the registers based on the annotations. Whereas the distribution of semantic categories shows small variance, the distribution of syntactic categories shows large variance among registers. Second, we will annotate the function words in the corpus with semantic labels. In the WLSP, the word senses of function words are not entirely defined. Table 5 shows the word-sense-defined function words in WLSP; it contains only 15 entries. We have to define new word-sense la-

Table 6: Statistics of Syntactic Categories (Content Word Only)

Register	1. Nominal	2. Verbal	3. Modifier	4. Other	not assigned	TOTAL
PB Books	29,966 (55.01%)	14,396 (26.53%)	7,179 (13.18%)	1,206 (2.21%)	1,727 (3.17%)	54,474 (100.00%)
PM Magazines	38,182 (62.81%)	13,076 (21.51%)	6,946 (11.43%)	883 (1.45%)	1,699 (2.80%)	60,786 (100.00%)
PN Newspapers	49,196 (73.53%)	10,973 (16.40%)	4,452 (6.65%)	407 (0.61%)	1,878 (2.81%)	66,906 (100.00%)
TOTAL	117,344 (64.42%)	38,445 (21.10%)	18,577 (10.20%)	2,496 (1.37%)	5,304 (2.91%)	182,166 (100.00%)

Table 7: Statistics of the Top Semantic Categories (Content Word Only)

Register	.1 Relation	.2 Subject	.3 Action	.4 Product	.5 Nature	not assigned	TOTAL
PB Books	25,193 (46.25%)	6,575 (12.07%)	15,783 (28.97%)	2,352 (4.32%)	2,844 (5.22%)	1,727 (3.17%)	54,474 (100.00%)
PM Magazines	28,982 (47.68%)	6,683 (10.99%)	17,270 (28.41%)	3,003 (4.94%)	3,149 (5.18%)	1,699 (2.80%)	60,786 (100.00%)
PN Newspapers	30,518 (45.61%)	11,006 (16.45%)	19,551 (29.22%)	2,063 (3.08%)	1,890 (2.82%)	1,878 (2.81%)	66,906 (100.00%)
TOTAL	84,693 (46.49%)	24,264 (13.32%)	52,604 (28.88%)	7,418 (4.07%)	7,883 (4.33%)	5,304 (2.91%)	182,166 (100.00%)

Table 8: Statistics of Out of Vocabulary Words in the Alignment Table

		1. Nominal	2. Verbal	3. Modifier	4. Other	TOTAL
OOV lexemes	tokens	9,040	300	1,279	81	10,699
	types	3,651	187	244	34	4,116
OOV senses (IV lexemes)	tokens	2,133	277	488	18	2,917
	types	647	158	194	7	1,007

Table 9: The Top 5 Frequent Article Numbers

Rank	PB article number	count	rate	PM article number	count	rate	PN article number	count	rate
1	2.3430 Verbal Action-Act-Act	2,468	4.53%	1.1960 Nominal Relation-Quantity-Numeral	4,800	7.90%	1.1960 Nominal Relation-Quantity-Numeral	6,908	10.32%
2	2.1200 Verbal Relation-Existence-Existence	2,406	4.42%	2.3430 Verbal Action-Act-Act	2,606	4.29%	1.1962 Nominal Relation-Quantity-Numeral Suffix	2,893	4.32%
3	1.1960 Nominal Relation-Quantity-Numeral	1,669	3.06%	2.1200 Verbal Relation-Existence-Existence	1,844	3.03%	2.3430 Verbal Action-Act-Act	2,729	4.08%
4	3.1010 Modifier Relation-Thing-Demonstrative	1,362	2.50%	1.1962 Nominal Relation-Quantity-Numeral_Suffix	1,298	2.14%	1.2590 Nominal Subject-Public-Location_Name	2,452	3.66%
5	1.2000 Nominal Relation-Existence-Existence	1,187	2.18%	3.1010 Modifier Relation-Thing-Demonstrative	1,008	1.66%	2.1200 Verbal Relation-Existence-Existence	1,330	1.99%

bels for presently sense-undefined function words. Third, we will develop a supervised all-word WSD model based on this corpus.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Number 17H00917, and 18H05521 and a project of the Center for Corpus Development, NINJAL.

References

Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A Sense-tagged Corpus of Japanese. In *The 6th International Conference of the Global WordNet Association (GWC-2012)*, pages 56–63.

Kokuritsu Kokugo Kenkyusho. 1964. Word List by Semantic Principles.

Kokuritsu Kokugo Kenkyusho. 2004. Word List by Semantic Principles (revised and enlarged version).

Asuko Kondo, Makiro Tanaka, and Masayuki Asahara. 2018. Alignment Table between UniDic and 'Word List by Semantic Principles'. In *Proceedings of JADH2018*, pages 125–128.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48:345–371.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2010. Semeval-2010 task: Japanese WSD. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 69–74, Uppsala, Sweden, July. Association for Computational Linguistics.

Jun Toyoura, Takenobu Tokunaga, Hitoshi Isahara, and Ryuichi Oka. 1996. Development of a RWC Text Database Tagged with Classification Code (in Japanese). In *IEICE, NLC96-13*, pages 89–96.

Toshio Yokoi. 1995. The EDR Electronic Dictionary. *Commun. ACM*, 38(11):42–44, November.