UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese

Mai Omura and **Masayuki Asahara**
National Institute for Japanese Language and Linguistics
{mai-om, masayu-a}@ninjal.ac.jp

Abstract

In this paper, we describe a corpus UD Japanese-BCCWJ that was created by converting the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a Japanese language corpus, to adhere to the UD annotation schema. The BCCWJ already assigns dependency information at the level of the bunsetsu (a Japanese syntactic unit comparable to the phrase). We developed a program to convert the BCCWJto UD based on this dependency structure, and this corpus is the result of completely automatic conversion using the program. UD Japanese-BCCWJ is the largest-scale UD Japanese corpus and the second-largest of all UD corpora, including 1,980 documents, 57,109 sentences, and 1,273k words across six distinct domains.

1 Introduction

The field of Natural Language Processing has seen growing interest in multilingual and cross-linguistic research. One such cross-linguistic research initiative is the Universal Dependencies (UD) (McDonald et al., 2013) Project, which defines standards and schemas for parts of speech and dependency structures and distributes multilingual corpora. As part of our efforts to import the UD annotation schema into the Japanese language, we defined a part-of-speech (PoS) system and set of dependency structure labels for Japanese, which are documented on GitHub [1], and we are currently preparing reference corpora. This paper describes our Japanese UD corpus **UD Japanese-BCCWJ**, which is based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014), and which we have prepared as part of our efforts to design a Japanese version of UD.

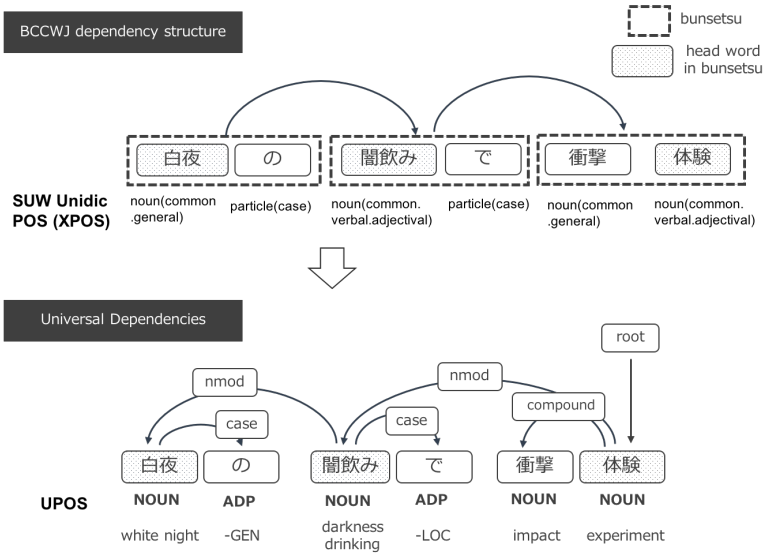
[1] https://github.com/UniversalDependencies/

Previous applications of UD to Japanese corpora can be found in Table 1, which is based on (Asahara et al., 2018). Tanaka et al. (2016) have published a Japanese UD treebank, UD Japanese-KTC, which was converted from the Japanese Phrase Structure Treebank (Tanaka and Nagata, 2013). Other corpora include an unlabelled UD Japanese treebank derived from Wikipedia, UD Japanese-GSD, and a Japanese-PUD corpus, UD Japanese-PUD (Zeman et al., 2017), derived from parallel corpora, but all of these have had to be partially manually corrected. According to Table 1, UD Japanese-BCCWJ is the largest UD Japanese corpus. Furthermore, it is the second largest of all UD corpora and includes many documents across various domains as shown in Table 3.

Existing Japanese-language corpora tagged with dependency structures include the Kyoto University Text Corpus (Kurohashi and Nagao, 2003) and the Japanese Dependency Corpus (Mori et al., 2014). These corpora frequently use **bunsetsu** as the syntactic dependency annotation units for Japanese. Also, the BCCWJ, based on UD Japanese-BCCWJ, is annotated using a bunsetsu-level dependency structure (Asahara and Matsumoto, 2016), which we must thus convert from a bunsetsu-level dependency structure to a Universal Dependencies schema. Figure 1 shows an example of BCCWJ with the UD annotation schema.

In this paper, we describe the conversion of the BCCWJ to the UD annotation schema. To accomplish the conversion, the following information must be combined: word-morphological information, bunsetsu-level dependency structure, coordination structure annotation, and predicate argument structure information. We also attempt to convert the BCCWJ to a UD schema, which allows us to respond to changes in the tree structures based on ongoing discussions in the UD commu-

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 117–125
Brussels, Belgium, November 1, 2018.



(There is an) impact experiment on the darkness drinking party on white night.

Figure 1: Summary of conversion of BCCWJ to UD. (The sample is from `PB_00001`). The left example is the BCCWJ schema, bunsetsu-level dependency structure, and the right is the Universal Dependencies schema.

Table 1: Comparison of existing UD Japanese resources.

Treebank	Tokens	Version	Copyright	Media
UD Japanese-BCCWJ	**1273k**	v2.2	masked surface	Newspaper, Books, Magazines, Blogs, etc.
UD Japanese-KTC	189k	v1.2	masked surface	Newspaper
UD Japanese-GSD	186k	v2.1	CC-BY-NC-SA	Wikipedia
UD Japanese-PUD	26k	v2.1	CC-BY-SA	Wikipedia Parallel Corpus
UD Japanese-Modern	14k	v2.2	CC-BY-NC-SA	Magazines in 19th century

Table 2: Genres in BCCWJ core data. Please refer to Table 3 about the number of sentences/tokens.

Abbr.	description
OC	Bulletin board (Yahoo! Answers)
OW	Government white papers
OY	Blog (*Yahoo! Blogs*)
PB	Books
PM	Magazines
PN	Newspaper

nity. The next section is a brief description of our current conversion.[2]

2 Balanced Corpus of Contemporary Written Japanese (BCCWJ)

The *Balanced Corpus of Contemporary Written Japanese* (BCCWJ) (Maekawa et al., 2014) is a 104.3-million-word corpus that covers a range of genres including general books and magazines, newspapers, white papers, blogs, Internet bulletin board postings, textbooks, and legal statutes. It is currently the largest balanced corpus of Japanese. The copyright negotiation process has also been completed for BCCWJ DVD purchasers. [3]

All BCCWJ data are automatically tokenized and PoS-tagged by NLP analysers in a three-layered tokenization of Short Unit Word (SUW), Long Unit Word (LUW), and bunsetsu as in Figure 2.[4] There are subcorpora to be checked manually to improve their quality after analysis, as well as a subcorpus of the 1% of the BCCWJ data called 'core data' consisting of 1,980 samples and 57,256 sentences with morphological information (word boundaries and PoS information). Table 2 describes each genre in the BCCWJ core data. The distribution, including the BCCWJ core data, is shown in Figure 3. The UD Japanese-BCCWJ is based on the BCCWJ core data.

The BCCWJ provides bunsetsu-level dependency information as BCCWJ-DepPara (Asahara and Matsumoto, 2016) including bunsetsu dependency structures, coordination structures, and information on

[2]UD Japanese-BCCWJ was released in Universal Dependencies on 2018 March; however, we noticed and addressed some problems after release, and so the development version is as described in this paper.

[3]`http://pj.ninjal.ac.jp/corpus_center/bccwj/en/`

[4]The details of these layers are described in Section 3.1.

Table 3: Genre distribution including BCCWJ core data. A description of each genes is given in the Table 2.

Type \ Genre		OC	OW	OY	PB	PM	PN	total
Documents	train	421	45	214	58	63	286	1,087
	dev	259	9	129	13	12	27	449
	test	258	8	128	12	11	27	444
	total	938	62	471	83	86	340	1,980
Sentences	train	2,838	4,456	3,278	7,196	9,546	13,487	40,801
	dev	1,650	780	1,920	1,131	1,510	1,436	8,427
	test	1,619	589	1,722	1,351	1,486	1,114	7,881
	total	6,107	5,825	6,920	9,678	12,542	16,037	57,109
Tokens(SUWs)	train	50,415	168,909	51,310	174,394	177,947	300,786	923,761
	dev	29,961	31,471	32,164	27,315	30,328	29,528	180,767
	test	29,624	26,421	28,485	29,612	28,183	26,434	168,759
	total	110,000	226,801	111,959	231,321	236,458	356,748	1,273,287

predicate-argument structures through BCCWJ-DepPara-PAS (Ueda et al., 2015). This information is exploited in the conversion of BCCWJ to the UD schemas.

3 Conversion of BCCWJ to UD

As shown in Figure 1, there are some differences between the BCCWJ and UD schemas. One concerns PoS: BCCWJ's and UD's PoS Unidic (Den et al., 2007) and Universal PoS (Petrov et al., 2012), respectively (e.g. `noun(common.general)` and NOUN in Figure 1). Second, the structure is different between bunsetsu-level and word-level dependency, for example in the directions and units of dependency (compare BCCWJ with the UD schema in Figure 1). Finally, the bunsetsu-level dependency structures in Japanese have less detailed syntactic dependency roles than the relations in Universal Dependencies like `nmod` and `case`. We need to convert UD Japanese-BCCWJ while taking into consideration the differences between the UD and BCCWJ schemata. In addition, we need to choose or detect apposite word units for the basic word unit based on UD guidelines from SUWs, LUWs, and others because these layers are not always appropriate as given by BCCWJ. Therefore, we convert BCCWJ to UD Japanese-BCCWJ using the following steps:

1. Detect the word unit.
2. Convert Unidic PoS to UD PoS.
3. Convert bunsetsu-level dependency to UD word-level dependency.
4. Attach a UD relation label to each dependency.

We will describe each step in the following sections.

3.1 Word Unit

Japanese, unlike English as well as many other languages, text is not explicitly divided into words using spaces. UD guidelines specify that the basic units of annotation are *syntactic words* [5]. The first task is therefore to decide what counts as a token and what counts as a syntactic word.

All the samples in the BCCWJ are morphologically analysed based on linguistic units called 'Short Unit Words' (SUWs) and 'Long Unit Words' (LUWs), as in Figure 2. SUWs are defined on the basis of their morphological properties in the Japanese language. They are minimal atomic units that can be combined in ways specific to particular classes of Japanese words. LUWs are defined on the basis of their syntactic properties. The bunsetsu are word grouping units defined in terms of the dependency structure (the so-called *bunsetsu-kakariuke*). The bunsetsu-level dependency structure annotations in BCCWJ-DepPara (Asahara and Matsumoto, 2016) rely on LUWs. As shown in Figure 2, the SUWs, LUWs, and bunsetsu exist in a hierarchical relationship: SUW <= LUW <= bunsetsu; SUWs render 魚/フライ/を as three words, LUWs as 魚フライ/を or two words, and bunsetsu as 魚フライを or one word. SUWs and LUWs also entail different PoS systems, as will be described in Section 3.2.

UD Japanese-BCCWJ adopts the SUW word unit, which corresponds to the BCCWJ's basic PoS system, as its fundamental linguistic unit. However, as described in the following sections, usage information associated with LUWs is also required to conform to UD standards and to achieve consistency with annotations for other languages. We will discuss the differences between

[5] http://universaldependencies.org/u/overview/tokenization.html

魚フライを食べたかもしれないペルシャ猫 "It is the Persian cat that may have eaten fried fish."											
SUW	魚	フライ	を	食べ	た	か	も	しれ	ない	ペルシャ	猫
	NOUN	NOUN	ADP	VERB	AUX	PART	ADP	VERB	AUX	PROPN	NOUN
	fish	fry	−ACC	eat	−PAST			know	−NEG	Persia	cat
LUW	魚フライ		を	食べ	た	かもしれない				ペルシャ猫	
	NOUN		ADP	VERB	AUX	AUX				NOUN	
	fried fish		−ACC	eat	−PAST	may				Persia	cat
bunsetsu	魚フライを			食べたかもしれない						ペルシャ猫	

Figure 2: An example of a Japanese word unit: 'It is the Persian cat that may have eaten fried fish' in Japanese.

SUWs and LUWs in Section 5.1.

3.2 Conversion to Universal PoS tags

UD has adopted Universal PoS tags, version 2.0 (Petrov et al., 2012), as a system for aggregating the parts of speech of all languages; in this system 17 distinct parts of speech are defined. For the Japanese-language version of UD, we defined the UD parts of speech by constructing a table of correspondences using UniDic (Den et al., 2007) and the Universal PoS tags. For SUWs, BCCWJ adopts a PoS system based on a word's possible lexical categories. For example, the PoS tag `noun(common.adverbial)` (名詞-普通名詞-副詞可能) means that the word can be a common noun (普通名詞) or an adverb (副詞). In contrast, LUWs are used to specify PoS tags based on *usage principles*, which resolve usage ambiguities based on context. The `noun(common.adverbial)` tag in the SUW PoS system resolves to a common noun or an adverb depending on context. We selected the SUW PoS system because SUWs are the base annotation of word units of the BCCWJ; broadly speaking, there is no significant difference between the SUW and LUW PoS systems for our purposes.

However, for certain words we need to use a LUW PoS system based on usage principles in order to conform to the UD standards and to achieve consistency with other languages. For example, in the case of a nominal verb (`noun(common.verbal_suru)`, which can add -する) or nominal adjective (`noun(common.adjectival)`, which can add -な), the SUW PoS system, based on lexical principles, is not appropriate because if a word is a verb or adjective depending on the context, the SUW PoS system cannot detect this. Instead, here we use LUW PoS tags based on usage principles that resolve ambiguities based on context. The LUW PoS tags based on usage principles have the advantage of being easier to map onto other languages, and the reduced ambiguity associated with word endings makes it easier to specify the conditions for a VERB or ADJ tag.

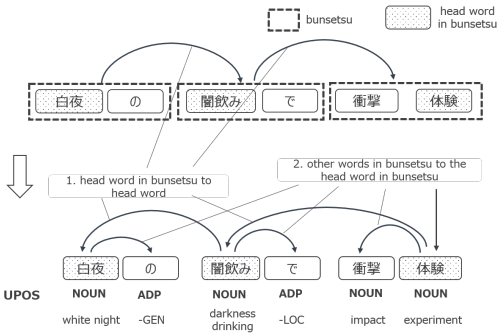


Figure 3: Illustration of the conversion of bunsetsu-level dependency to UD word-level dependency.

Table 4 shows the mapping between Universal PoS tags and UniDic based on these principles. Note that the mapping is for Unidic SUW PoS; using Unidic LUW PoS would be simpler, as described in the following Section 5.1. The fact is, however, that there are several problems involved in using LUW PoS, as will be described presently.

3.3 Conversion of dependency structure

For syntactic information for Japanese, we use BCCWJ-DepPara (Asahara and Matsumoto, 2016), which includes bunsetsu dependency and coordination information for the BCCWJ. In order to convert bunsetsu-level into word-level dependencies, we identify the head word in the bunsetsu and then attach all other elements in the bunsetsu to the head word, as in Figure 3. Note that the UD dependency arrow is from the head to the dependent word, whereas the BCCWJ dependency arrow is from the dependent to the head word; this is merely a notational issue and the substantive description is the same. Moreover, the head-word

[6]Japanese uses various suffixes to make an adjective phrase using a noun, -的; to express an honorific meaning, such as -さん; and so on. However, we use the NOUN for the time being for various reasons.

Table 4: Some of example of labeling rule UPOS, which of the number is about forty.

SUW POS	Basic form	LUW POS	UD rel
adjective_i(bound)		auxiliary	AUX
adjective_i(bound)		adjective_i(general)	ADJ
adnominal	^[こそあど此其彼]の (ko/so/a/do/ko/ka-no)		DET
adnominal	^[こそあど此其彼](ko/so/a/do/ko/ka)		PRON
verb(bound)	為る (suru)		AUX
verb			VERB
noun(proper.*.*)			PROPN
noun(common.adverbial)		adverb	ADV
noun(common.adverbial)			NOUN
prefix		adverb	NOUN
suffix			NOUN[6]

in the bunsetsuis selected as the rightmost content word after separating content and function words; for example, the head-word is 体験‘experiments’ in 衝撃体験‘impact experiments’ in Figure 3. [7]

While BCCWJ-DepPara includes dependency information, it does not include syntactic dependency roles corresponding to the Universal Dependencies relations (de Marneffe et al., 2014) (such as the labels nsubj, obj, and iobj). We therefore determined and assigned the UD relation labels based on the case-marking (particle(case|binding|adverbial)) or predicate-argument structure information in BCCWJ-PAS (Ueda et al., 2015). This predicate-argument structure information is semantic-level information, so *basically* we use the case-marking, and the predicate-argument information is just for reference. Since Japanese, unlike languages such as English, can omit core arguments and case-marking and the case-marking *always* corresponds with grammatical arguments in UD relations, predicate-argument structure is necessarily expressed by the case marker. For example, the case marker は *ha* usually indicates a nominal subject nsubj, but also frequently appears as a topic marker. [8]

Table 5 shows the rules for assigning UD relations. These conversions combine various rules like bunsetsu information, case information, and coordination relations between the head word and the dependent word.

Our current rules, which are unable to identify clauses, thus cannot effectively handle clause-related labels such as csubj, advcl, and acl; this is because clauses in Japanese are vaguer than in English, as described in Section 5.2. In the future, we will solve this problem by establishing criteria for identifying clauses.

Table 5: Some of example of rules for assigning UD relations, which of the number is about sixty. It is more detailed in the actual implementation.

Rule	Label
root of sentence and head word in bunsetsu.	root
have UD POS NUM	nummod
have UD POS ADV	advmod
include case 'ga' (nominative case) in bunsetsu	nsubj
include case 'o' (accusative case) in bunsetsu	obj
have UD POS VERB and the dependency have UD POS VERB if the relation is above bunsetsu.	aux
have UD POS VERB and the dependency have UD POS VERB if the relation is not above bunsetsu	compound

Table 6: MISC field on UD Japanese-BCCWJ. It is a development version, so may be changed.

label	description
BunsetuBILabel	BI-tags on bunsetsu (B=top of bunsetsu, I=others.)
BunsetuPositionType	Type of bunsetsu
LUWBILabel	BI-tags on LUW. (B=top word of LUW, I=others.)
LUWPOS	LUW Unidic POS tag.

BCCWJ-DepPara also contains coordinate structure information, but our current conversion rules do not yet have defined rules related to coordinate structures such as cc and conj. The issue will be presented in (Kanayama et al., 2018).

3.4 Format

Through this process we can convert the BCCWJ to a UD schema. UD Japanese-BCCWJ is formatted by CoNLL-U. UD Japanese-BCCWJ provides the word form, lemma of the word form, universal part-of-speech tag, language-specific part-of-speech tag (Unidic POS), and Universal Depen-

[7] As described in (Kanayama et al., 2018), this property affects coordinate structures.

[8] Please refer to Section 3.4 in (Asahara et al., 2018) for a discussion of case markers in Japanese.

dencies relation. Note that the provided POS is the **SUW** POS serves as the language-specific PoS tag in UD Japanese-BCCWJ.

UD allows us to insert any annotation using the `MISC` field, so we can give syntactic information using this field for LUW word units and bunsetsu. This information may be useful for Japanese parsing. Table 6 summarizes the MISC fields in UD Japanese-BCCWJ.

4 Parsing by genre

UD Japanese-BCCWJ is attractive in that it includes documents in various genres. We present the parsing results that indicate differences by genre. In this paper we do not show part-of-speech tagging results, because there are some Japanese POS tagging tools (for example, Kudo et al. (2004)'s implementation, MeCab), which make it easier to convert Unidic to UD POS, as mentioned.

We use UDPipe (Straka and Straková, 2017) as a tool to train the parsing model and evaluate the parsing accuracy. UDPipe is a trainable pipeline for tokenization, tagging, lemmatization, and dependency parsing from CoNLL-U format files. The parsing uses Parsito (Straka et al., 2015), which is a transition-based parser using a neural-network classifier. We use default parameters in UDPipe. [9] We use the labelled attachment score (LAS) and unlabelled attachment score (UAS) as evaluation metrics.

The results are shown in Table 7 and Table 8. The columns in the Tables represent the parsing model by genre, the rows the genre tests, and 'all' is the full core data, so a given cell represents the result of evaluating the genre parsing model by the genre test set.

Whereas the genres of OW, PB, PM, and PN contain more than 200K tokens, the genres of OC and OY contain only around 100K, tokens as shown in Table 3.

It is in principle one of the advantages of UD Japanese-BCCWJ that it can utilize a relatively large scale sub-corpus. In fact, however, the UAS results show that if a genre has more than 200K tokens, the result from using only the in-domain data is better than that with the data for all 1.2 million tokens, including the out-domain data.

[9]The version using UDPipe is 1.2.1-devel, and executes with no options.

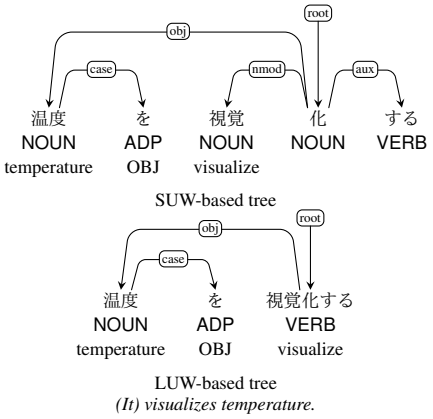


Figure 4: PoS variation between SUW and LUW

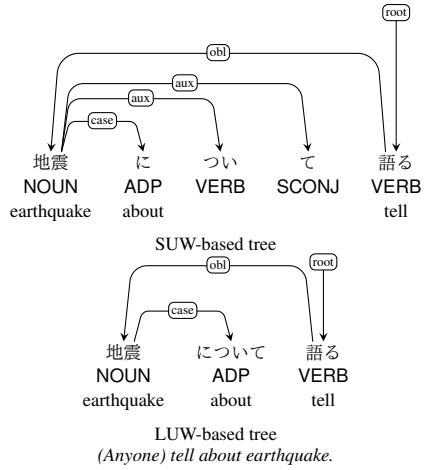


Figure 5: Multi-Word Expression

5 Discussion

In this section, we will take up a problem related to UD Japanese that centres on UD Japanese-BCCWJ. The overall discussion of UD Japanese is summarized by (Asahara et al., 2018).

We must also still discuss the issue of coordinate structures in Japanese. The issue will be presented in (Kanayama et al., 2018).

5.1 Word units

The choice of word unit is one of the important issues in UD Japanese. BCCWJ includes three sorts of word unit standards, as noted: SUWs, LUWs, and bunsetsu. We used SUWs for UD Japanese-BCCWJ.

However, the UD project stipulates that word delimitation in the UD standard should be for 'syntactic words'. LUWs in BCCWJ are thus a more preferable word delimitation standard than SUWs.

Figure 4 shows the difference between SUW PoS and LUW PoS. The top of Figure 4 shows the

Table 7: Results of unlabeled attachment score (UAS).

train \ test	OC	OW	OY	PB	PM	PN	all.
OC	89.70	81.99	88.46	87.93	88.45	87.21	**90.49**
OW	80.21	**88.62**	78.08	83.66	84.74	84.95	88.55
OY	86.35	79.54	86.15	84.62	85.67	84.66	**88.21**
PB	89.23	86.23	88.34	**91.56**	90.91	90.63	91.48
PM	87.28	85.57	86.64	89.65	**89.74**	89.32	89.67
PN	86.40	87.66	85.88	88.65	89.31	**91.20**	90.83
all.	86.64	84.84	85.71	87.74	88.18	88.00	**89.89**

Table 8: Results of LAS (Labeled attachment score). LAS consider the UD relation label unlike UAS.

train \ test	OC	OW	OY	PB	PM	PN	all.
OC	87.35	78.19	85.76	85.06	85.67	84.32	**88.17**
OW	78.36	**87.16**	76.16	82.06	83.03	83.23	87.00
OY	83.31	75.87	83.24	81.43	82.62	81.43	**85.33**
PB	86.60	83.47	85.73	89.21	88.58	88.07	**89.30**
PM	84.32	82.59	83.81	86.63	**87.16**	86.79	87.14
PN	83.65	85.03	83.34	85.93	87.06	**89.28**	88.90
all.	84.04	81.94	83.12	85.10	85.72	85.51	**87.65**

SUW-based PoS. The verb する‘do’and the verbal noun make a compound verb, as in the bottom of Figure 4 in the LUW-based segmentation.

Figure 5 presents a functional multi-word expression について, which includes three words in SUW units and one word in LUW units. We can mask the morphological construction of the syntactic word within a LUW.

However, currently we nevertheless continue to use SUWs as the UD Japanese word delimitation standard. This is because (1) LUWs are difficult to produce with word segmenters, and (2) some functional multi-word expressions in Japanese do not conform to the LUW standards.

5.2 Clause

The UD dependency labels are designed to be split between the word/phrase and the clause. The difference between clauses and words/phrases is vague in Japanese, because cases, including the subject, do not necessarily overtly appear in sentences.

Figure 6 shows an adjective clause and an adjective phrase in Japanese. At the top of Figure 6 is an overt adjective clause with a nominal subject. In contrast, however, in the example at the bottom of Figure 6 it cannot be determined whether the adjective is attributive or predicative, since the nominal subject of adjective predicate can be omitted in Japanese (in this case, しっぽ 'tail' may be omitted). Thus, we define `acl` for all adjectives which attach to noun phrases as the current state.

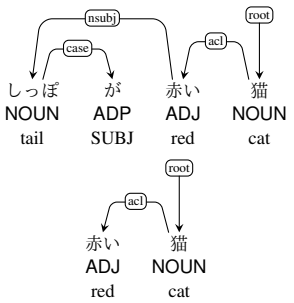


Figure 6: Clause or Phrase.

6 Other UD Japanese resources

In this section, we describe other UD Japanese resources at the time of writing. Table 2 shows a summary of these. As noted, there are five UD Japanese corpora as of March 2018, which in scale constitute the second largest of all UD corpora with the addition of the UD Japanese-BCCWJ.

UD Japanese-KTC (Tanaka et al., 2016) is based on the NTT Japanese Phrase Structure Treebank (Tanaka and Nagata, 2013), which contains the same original text as the Kyoto Text Corpus (KTC) (Kurohashi and Nagao, 2003). KTC is a bunsetsu-level dependency structure like BCCWJ, but with its own word delimitation schema and POS tag set. We are now modifying the UD Japanese KTC from the version 1.0 schema to version 2.0.

UD Japanese-GSD consists of sentences from Japanese Wikipedia that have been automatically split into words by IBM's word seg-

menter. The dependencies are automatically resolved using the bunsetsu-level dependency parser (Kanayama et al., 2000) with the attachment rules for functional words defined in UD Japanese.

UD Japanese-PUD (Zeman et al., 2017) was created in the same manner as UD Japanese-GSD, with the goal of maintaining consistency with UD Japanese-GSD. It is a parallel corpus with multiple other languages.

UD Japanese-Modern (Omura et al., 2017) is a small UD annotation corpus based on the *Corpus of Historical Japanese: Meiji-Taisho Series I Magazines* (CHJ) (Ogiso et al., 2017). The CHJ is large-scale corpus with morphological information of Old Japanese and has morphological information compatible with the BCCWJ. We annotated bunsetsu-level syntactic dependency and coordinated structures using the BCCWJ-DepPara annotation schema and predicate-argument relations, and utilized the conversion script used for UD Japanese-BCCWJ because the two corpora share the same annotation schema. There are two characteristic syntactic structures in Old Japanese. One is inversion, found in Sino-Japanese literary styles. The other is predicative adnominals.

As mentioned, each UD Japanese corpus has been developed in a different manner since the resources are derived from annotation with other standards. For example, UD Japanese-KTC is converted from a phrase structure treebank, while UD Japanese-Modern is based on compatible annotation with UD Japanese-BCCWJ. However, the syntactic structures of Old Japanese are very different from contemporary Japanese, as described above.

Presently we are trying to standardize UD Japanese resources under the UD Japanese-BCCWJ schema by annotating BCCWJ-DepPara with standard syntactic dependency notation for other resources. Then, we will use the conversion rules of this article for the other UD Japanese resources.

7 Summary and Outlook

In this paper, we described a corpus created by converting the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a Japanese language corpus, into the UD annotation schema. There are differences between BCCWJ and UD schemas, and so we have tried to develop and implement rules to convert BCCWJ to UD.

The UD Japanese-BCCWJ was released in March 2018. Note that though the corpus does not include the surface form due to the original text copyright, the BCCWJ DVD Edition purchaser can add the surface form using the scripts in the UD package. However, this is a matter of debate, as described in this paper, so we are going to continue to update it based on ongoing discussion, for instance regarding the apposite word unit for Japanese.

At the time of writing, we have completed the process of UD conversion based on SUWs. We also need to implement a corpus based on LUWs, and will publicly release our Japanese UD data based on both SUW and LUW analyses.

Acknowledgements

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

References

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. 2018. Universal Dependencies Version 2 for Japanese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1824–1831, Miyazaki, Japan.

Masayuki Asahara and Yuji Matsumoto. 2016. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pages 49–58, Osaka, Japan.

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Minematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. *The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics (in Japanese)*.

Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D. Hwang, Yusuke Miyao, Jinho Choi, and Yuji Matsumoto. 2018. Coordinate structures in universal dependencies for head-final languages. In *Proceedings of Universal Dependencies Workshop 2018 (UDW 2018)*, Brussels, Belgium. (to appear).

Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. 2000. A hybrid Japanese parser with hand-crafted grammar and statistics. In

Proceedings of the 18th International Conference on Computational Linguistics (COLING '00), pages 411–417, Saarbrücken, Germany.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Barcelona, Spain.

Sadao Kurohashi and Makoto Nagao. 2003. *Building a Japanese Parsed Corpus – while Improving the Parsing System*, Treebanks: Building and Using Parsed Corpora, chapter 14. Springer, Dordrecht.

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(2):345–371.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4585–4592, Reykjavik, Iceland.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K.Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. B. Castelló, and J. Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 92–97, Sofia, Bulgaria.

Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada. 2014. A Japanese word dependency corpus. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 753–758, Reykjavik, Iceland.

Toshinobu Ogiso, Asuko Kondo, Yoko Mabuchi, and Noriko Hattori. 2017. Construction of the "Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines". In *Proceedings of the 2017 Conference of Digital Humanities (DH2017)*, Montrèal, Canada.

Mai Omura, Yuta Takahashi, and Masayuki Asahara. 2017. Universal Dependency for Modern Japanese. In *Proceedings of the 7th Conference of Japanese Association for Digital Humanities (JADH2017)*, pages 34–36.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2089–2096.

Milan Straka, Jan Hajič, Jana Straková, and Jan Hajič jr. 2015. Parsing universal dependency treebanks using neural networks and search-based oracle. In *Proceedings of Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 2015)*.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto. 2016. Universal Dependencies for Japanese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1651–1658.

Takaaki Tanaka and Masaaki Nagata. 2013. Constructing a practical constituent parser from a Japanese treebank with function labels. In *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL'2013)*, pages 108–118, Seattle, Washington, USA.

Yoshiko Ueda, Ryu Iida, Masayuki Asahara, Yuji Matsumoto, and Takenobu Tokunaga. 2015. Predicate-argument structure and coreference relation annotation on 'Balanced Corpus of Contemporary Written Japanese' (in Japanese). In *Proceedings of the 8th Workshop on Japanese Language Corpus*, pages 205–214, Tokyo, Japanese.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gökırmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Hěctor Martínez Alonso, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadova, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.