Universal Dependencies Version 2 for Japanese

Masayuki Asahara♡†, Hiroshi Kanayama♠, Takaaki Tanaka♣
Yusuke Miyao◇, Sumire Uematsu◇, Shinsuke Mori○
Yuji Matsumoto●, Mai Omura♡, Yugo Murawaki○
♡ National Institute for Japanese Language and Linguistics, ♠ IBM Research AI
♣ NTT Communication Science Labolartories, ◇ National Institute of Informatics
○ Kyoto University, ● Nara Institute of Science and Technology
masayu-a@ninjal.ac.jp†

Abstract

In this paper, we are presenting our work to build Universal Dependencies (UD) resources for Japanese. The UD Japanese resources are built based on automatic conversion from several treebanks. The word delimitation, POS, and syntactic relations of the existing treebanks are ported for the UD annotation scheme. We discuss the issues of the UD scheme found through porting of the Japanese language.

1. Introduction

The Universal Dependencies (UD) project (McDonald et al., 2013) has defined a consistent, cross-linguistic target and syntactic structure representation format. In this presentation, we will show the work of the UD Japanese team. The UD Japanese team was organised by interested people who are developing their own treebanks or parsers. We developed and maintained several UD guidelines (version 2.0) compatible data for Japanese. Most of the data are made through automatic conversion from the existing treebank. The UD annotation guidelines were updated from version 1 to version 2 in early 2017. The automatic conversion enabled us to adapt the existing annotation based on traditional Japanese grammar conventions for the UD annotation guidelines changes. In this paper, we discuss the current issues of UD Japanese resources until today. These issues come from the difficulty to perform cross-linguistically consistent annotation for the different grammatical system from western European languages. The points at the issues related to the conversions are split into the delimitation (word, phrase and clause), undefined policies of UD guidelines, typological systems for UD, and copyright of Japanese language resources.

2. UD Japanese Resources

We have the following UD Japanese resources: UD Japanese-BCCWJ, UD Japanese-KTC (Tanaka et al., 2016), UD Japanese-GSD, UD Japanese-PUD, and UD Japanese-Modern (Omura et al., 2017). Table 1 presents the current status of UD Japanese resources. Below, we describe these resources briefly.

UD Japanese-BCCWJ is UD data based on the 'Balanced Corpus of Contemporary Written Japanese' (hereafter BCCWJ) (Maekawa et al., 2014). The BCCWJ defines 1 million word-scale core data samples in which the morphological information is manually annotated with three layers of word delimitations: Short Unit Word (SUW), Long Unit Word (LUW), and *bunsetsu*. The BCCWJ has several syntactic annotations. The BCCWJ-DepPara (Asahara and Matsumoto, 2016) is a *bunsetsu*-based syntactic dependency and coordinate structure annotation. The BCCWJ-PAS (Ueda et al., 2015) is a predicate-argument

relation annotation with the NAIST Text Corpus annotation schema (Iida et al., 2007). We maintain conversion rules based on these annotations.

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*, namely base phrase, based dependency treebank with its own word delimitation schema and POS tagset. The NTT Japanese Phrase Structure Treebank is a phrase structure-based treebank. The word delimitation and POS are adapted to the UniDic SUW standard. The data is still in version 1.0 schema as of February 2018. We are now modifying UD Japanese KTC from version 1.0 schema to version 2.0.

UD Japanese-GSD (formerly known as UD Japanese) consists of sentences from Wikipedia. The version 2.0 of this annotated corpus was provided for the CoNLL 2017 Shared Task (Zeman et al., 2017). In the release of version 2.0, the sentences have been automatically split into words by IBM's word segmenter. The segmentation errors were removed by adding lexicons specific to the data. In addition, 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 (Tanaka et al., 2016). Complex sentences with parenthesis were removed to avoid parsing errors. In the version 2.1 released in November 2017, manual annotations were merged with the semi-automatic annotations to reduce remaining errors.

UD Japanese-PUD was created in the same manner as UD Japanese-GSD, with the goal of maintaining consistency with UD Japanese-GSD. Since it is a parallel corpus with other languages, no sentences were removed from the corpus, including the ones containing parenthesis.

UD Japanese-Modern (Omura et al., 2017) is a small UD annotation data based on the 'Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines' (CHJ) (Ogiso et al., 2017). The CHJ has morphological information compatible with the BCCWJ. We annotated *bunsetsu*-based syntactic dependency and coordinated structures using BCCWJ-DepPara annotation schema. We also an-

Table 1: Status of UD Japanese resources.

	UD Japanese-BCCWJ	UD Japanese-KTC	UD Japanese-GSD	UD Japanese-PUD	UD Japanese-Modern
Tokens (SUW)	1,098K	189K	186K	26K	14K
Copyright	masked surface	masked surface	CC BY-NC-SA	CC BY-SA	CC BY-NC-SA
UD version	v2	v1	v2	v2	v2
Media	Newspaper, Books, Magazines, Whitepaper, Yahoo! Answers, Yahoo! Blogs	Newspaper	Wikipedia	Parallel Corpus	Magazines in 19th century
Method	conversion from *bunsetsu* dependency	conversion from phrase structure tree	integration of bunsetsu-level manual annotations and word-level rule-based parser outputs	integration of bunsetsu-level manual annotations and word-level rule-based parser outputs	conversion from *bunsetsu* dependency
Plan	will be released in March 2018	the modified version (v2) will be released in September 2018	annotating by BCCWJ-DepPara schema from scratch	annotating by BCCWJ-DepPara schema from scratch	will be released in March 2018

notated predicate-argument relations based on the NAIST Text Corpus annotation schema. We utilised the conversion script used for UD Japanese-BCCWJ, because the two corpora share the same annotation schema. There are two characteristic syntactic structures in modern Japanese. One is inversion in Sino-Japanese literary styles. The other is predicative adnominals.

3. Issues

3.1. Overview

The first issue is word delimiters. Japanese has no overt word delimitation. We have to define what is the syntactic word in UD for such languages. The same issue can be found in the discrimination between phrases and clauses. Because obligatory cases in Japanese can be omitted, the definition of adjective and adverbial clauses is vague which requires some clarification for the UD syntactic relations.

The second issue is undefined policies in the UD guidelines. The design of POS can be split into two principles: one is lexicon-based morphology level; the other is the usage-based syntax level. The design of syntactic relations also has several levels. One level is that the relations are defined only by the surface form. Another level is that the relations are defined by the contextual words. The UD guidelines are not entirely clear as to define which levels are appropriate for the designs of POS and syntactic relations.

The third issue is the guidelines related to linguistic typology. The UD guidelines are biased toward subject prominent languages such as English. However, Japanese is one of topic prominent languages. For example, in the topic prominent languages, whereas the expletive label `expl` cannot appeared, a label for the topic marker is needed.

Though Japanese is a strictly head-final language, the UD guidelines define head-initial construction for coordinate structures. We hope to introduce the head directionality parameter for UD guidelines.

The last issue is the copyright of the original texts for Japanese language resources.

3.2. Word Delimitation

Word delimitation is a critical issue for building UD annotation, because Japanese is written without word boundaries. UD guidelines specify that the basic units of annotation are *syntactic words*.

We tend to define morphemic units which are smaller than the word unit in order to maintain unit uniformity. Therefore, when we define the morpheme unit as the Universal Dependency word unit, we must annotate the compound word construction, as defined in the morphological layer of Japanese linguistics. Although the smaller word unit can be produced with high precision, this is not suitable for Japanese syntactic dependency annotation. In Japanese NLP, *bunsetsu* (base phrase) tends to be used as a syntactic dependency annotation unit. The morphology level including multi-word expressions is encapsulated within a bunsetsu. Therefore we can concentrate on the annotation of purely syntactic phenomena.

We must define the *syntactic words* for UD annotation for Japanese. We used the BCCWJ morpheme annotation standard, which is based on UniDic word boundary definition. The definition contains three layers: SUW, LUW, and *bunsetsu*. SUW can be produced by the morphological analyser MeCab.[1] with UniDic[2] LUW and *bunsetsu* can be produced by the pre-trained chunker Comainu.[3]

NINJAL[4] defined five sorts of word unit definitions by operationalism. The most fine-grained unit is NINJAL Minimum Unit Word. SUW (Short Unit Word: 短単位) is constructively defined by the NINJAL Minimum Unit Word (最小単位). MUW (Middle Unit Word: 中単位) is a basic unit where a sound may change at the beginning or the ending of a word and/or an accent may change (Uchimoto and Den, 2008). The Middle Unit Word defines voiced compound ("rendaku") (van de Weijer et al., 2005).
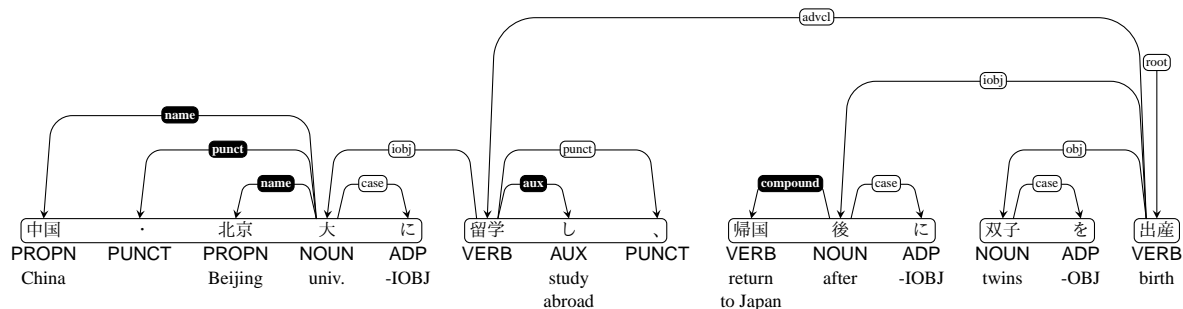

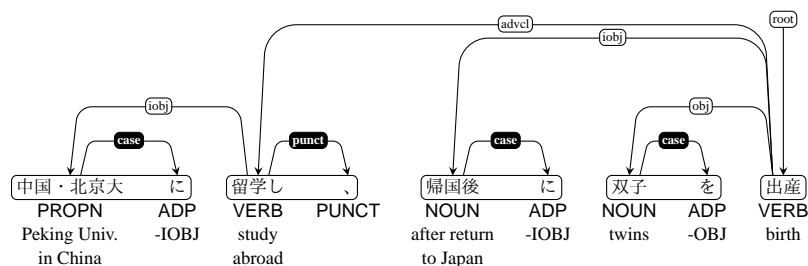[1] `taku910.github.io/mecab/`
[2] `unidic.ninjal.ac.jp/`
[3] `osdn.net/projects/comainu/`
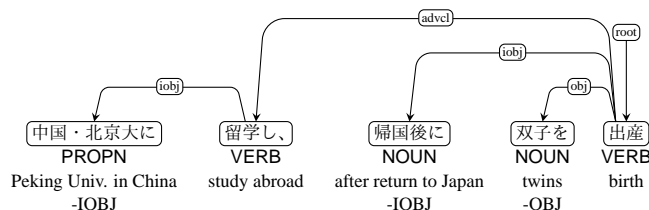[4] National Institute for Japanese Language and Linguistics.

Figure 1: Word delimitation standards in Japanese.

The 'named entity' and 'compound word construction' related labels (in the black ground) which show up SUW disappear in the LUW (Long Unit Word: 長単位) layer. In the most coarse-grained level, *bunsetsu*, case marker and auxiliary verbs that show up in LUW disappear. Currently, all UD Japanese resources (in versions 1 and 2) are based on SUW word delimitations. Figure 1 shows the SUW, LUW and *bunsetsu* examples. The group of words in a squircle defines a *bunsetsu* chunk.

UD schema defines that the basic units of annotation are syntactic words, though a renowned typologist failed to identify words consistently across languages (Haspelmath, 2011). However, SUW is not suitable for syntactic dependency annotations. We need to define syntactic word delimitation for UD Japanese. Structuralists (Hattori, 1960; Miyaoka, 2015) discuss Japanese syntactic word definition. A blog[5] also discusses the plausible unit for UD Japanese resources. These discussions are not based on any language resources.

3.3. POS

In Japanese NLP, we have several POS tagsets used in IPADIC, JUMAN, and UniDic. IPADIC and its POS tagset have not been maintained for 10 years. The JUMAN POS tagset is based on the Masuoka-Takubo POS tagset (Masuoka and Takubo, 1992). UniDic has a two layered POS tagset for SUW and LUW delimitations. The two layered UniDic POS tagset is split in to two categories: lexicon-based (語彙主義) and usage-based (用法主義). The lexicon-based approach involves 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 depending on the context. The POS labels of UniDic SUW are lexicon-based. The POS labels are maintained in a large-scale POS-tagged lexicon and used in semi-Markov model-based morphological analysers. Usage-based labelling is determined by contextual information in a sentence. The POS labels of UniDic LUW are usage-based. The POS labels are produced by chunking from the UniDic SUW sequences with contextual features. The UD guidelines do not present POS design princi-
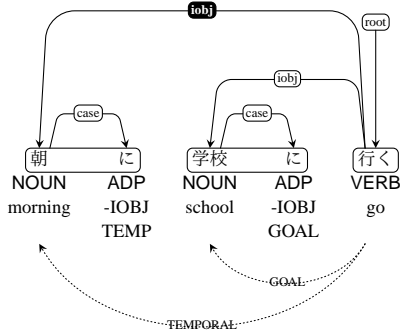
Figure 2: Double indirect object markers.

Figure 3: Topic marker.

Figure 4: Double subject sentence.

Figure 5: Indirect object marker with topic marker.

Table 2: The UD dependency labels between word/phrase and clause.

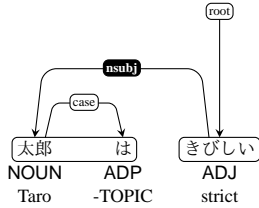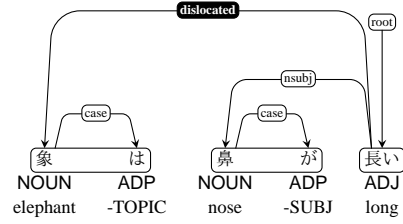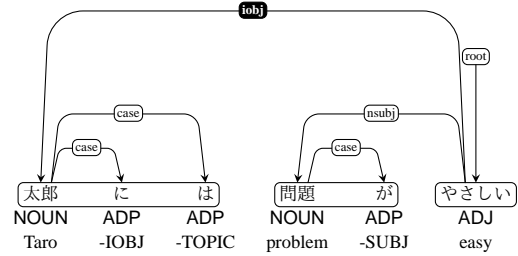| word/phrase	clause
nsubj	csubj
obj, iobj	ccomp, xcomp
amod	acl
advmod	advcl

ples with this respect. We used usage-based POS tags from UniDic based on lexicon/corpora/morphological analysers to align the Universal POS tags.

3.4. Japanese case marker and syntactic relations

A phrase-based treebank includes syntactic relation information for UD. However, a *bunsetsu* dependency-based treebank does not include the syntactic relation information. Our policy is to produce UD syntactic relations from the surface forms of *bunsetsu* dependency-based dependency structures.

We assigned the label nsubj to nominal phrases with the case marker '가' (ga), the label obj to nominal phrases with the case marker 'を' (o), and the label iobj to nominal phrases with the case marker 'に' (ni), respectively.

Though the indirect object marker 'に' (surface case) is given the label iobj, it can appear in more than one phrase with different deep cases for one predicate. The example in Figure 2 shows two indirect objects with the deep cases of TEMPORAL and GOAL presented below the sentences. If the UD guideline define deep case based labels for the syntactic relations, we need to assign the label obl to the TEMPORAL case. It is difficult to determine deep cases. Though the information to discriminate iobj from obl is available in BCCWJ-PAS or compatible annotation, it is not reproducible for other language resouces.

Japanese obligatory cases for predicates do not necessarily appear overtly in Japanese sentences. If '가' does not overtly appear in the clause, the label nsubj is assigned to a nominal phrase with a topic marker 'は' (wa). Figure 3 shows an example of nsubj indicated by the topic marker. The topic marker 'は' and subject marker '가' can appear for the same predicate. Figure 4 shows such a double sub-

ject sentence. The UD guidelines specify that the example needs the label dislocated.

The topic marker 'は' can immediately follow an indirect object marker 'に' as in Figure 5.

The labels for the topic marker are instable because of the design of UD syntactic labels. The design of UD syntactic labels are biased to subject-prominent languages. We argue for a new label for the topic marker for the topic-prominent languages such as Japanese, Korean, Chinese and Indonesian.

3.5. Clause

The UD dependency labels are designed to split between word/phrase and clause (Table 2). The difference between clause and word/phrase is vague in Japanese, because obligatory cases, including subject, do not necessarily overtly appear in the sentences.

We define the dependency label csubj only for the clause with a particle (準体助詞) 'の' (no) with UD POS SCONJ in UD Japanese resources (Figure 6).

We define the dependency label ccomp only for the clause with a case particle 'と' (to) with UD POS ADP in UD Japanese resources (Figure 7).

xcomp is a label generally used for the open clausal complement of a verb or an adjective that is a predicative or clausal complement without its own subject. However, most subjects in Japanese clauses do not overtly appear. Therefore, we do not use xcomp for Japanese.
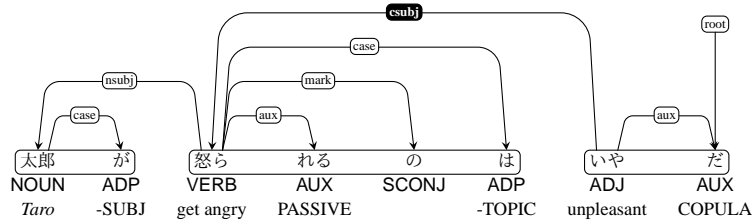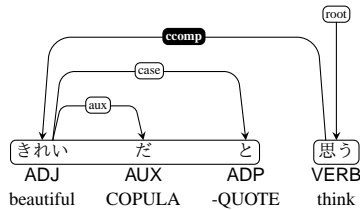
Figure 6: Clausal subject `csubj`.
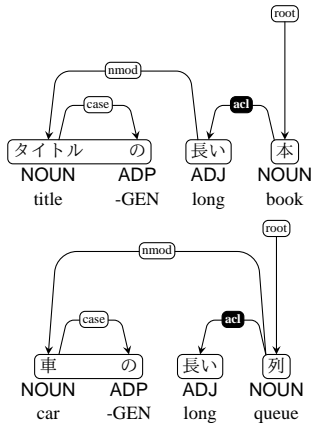

Figure 7: Clausal complement `ccomp`.

Figure 8: Adjective clause `acl`.

In the case of adjectives, attributive and predicative usages are not exclusive. There are two examples of attributive usages given in Figure 8. Whereas the latter example does not have an overt sense subject, the former has an overt sense subject and can also indicate predicative usage. The absence of an overt sense subject does not mean that the adjective is not used in the predicate. Therefore, we regard both adjectives as clauses, and we assign `acl`. The same goes for `advcl` in which there are adverbial forms (連用形) of adjectives.

Furthermore, `acl` has an ambiguity with `appos`. The UD guidelines (v2) define the label (`acl`) for an adjective clause with a predicate-argument relation (関係節内の関係) between the modifier and head (nominal phrase). The Japanese language has a similar clause – an appositional clause (関係節外の関係) without predicate-argument relations between the modifier and head. The difference between the adjective clause and appositional clause is not clearly defined in Japanese, since the case for predicates including adjectives tends to be omitted. Figure 9 shows some ambiguous examples. The arcs below the

sentences indicate predicate-argument relations compatible with BCCWJ-PAS annotation. One interpretation (in the upper example) is that 「中間報告」 (interim report) can be the subject of the 「設け」 (set) as an adjective clause. The other interpretation (in the lower example) is that 「考え」 (idea) cannot be the subject of the 「設け」 (set) as an apposition clause. However, both interpretations are assigned `acl` in the UD annotation. Note that the label `appos` is assigned not for clause but for word/phrase.

3.6. Coordinate Structure

The Japanese language is a strictly head-final language with a *bunsetsu*-based dependency structure. However, UD guidelines define head-initial relations for coordinate structures, where in the attachment of conjuctions or punctuation is to the right-side constituent in the coordinate structures. Figure 10 shows simple coordinate structure examples. The grouping of the words indicates the *bunsetsu*. The word-based UD dependency relations intersect the *bunsetsu* boundaries.

Figure 11 shows a coordinate structure with three constituents. The root of the dependency tree is the leftmost constituent in the coordinate structure.

Figure 12 shows an example of non-constituent coordination. The upper example is the coordination of '花子はりんごを二つ買い' and '太郎はみかんを三つ買っ'. The root node is the rightmost word in the leftmost coordinate chunk. The lower example has an ellipsis of 'りんごを' before the numeral. The dependency relations of the omitted node are reduced and attached to the daughter node '三つ'.

Figure 13 shows another example of non-constituent coordination. The upper example is the coordination of '太郎は山に行き' and '花子は川に行っ'. The root node is the rightmost word in the leftmost coordinate chunk. The lower example is the ellipsis of the '行き' part, which is the root node in the upper example. The dependency relations of the omitted node that include the root are reduced and attached to the daughter node '太郎'. We assigned the label `orphan` between '太郎' (SUBJ) and '山' (IOBJ).

These peculiar tree constructions are caused by the head-initial definition of coordinate structures for a strictly head-final language.

3.7. Copyright

UD Japanese BCCWJ is based on BCCWJ. The annotation will be provided under an open license. However, users need to buy the original text in the form of the DVD edition
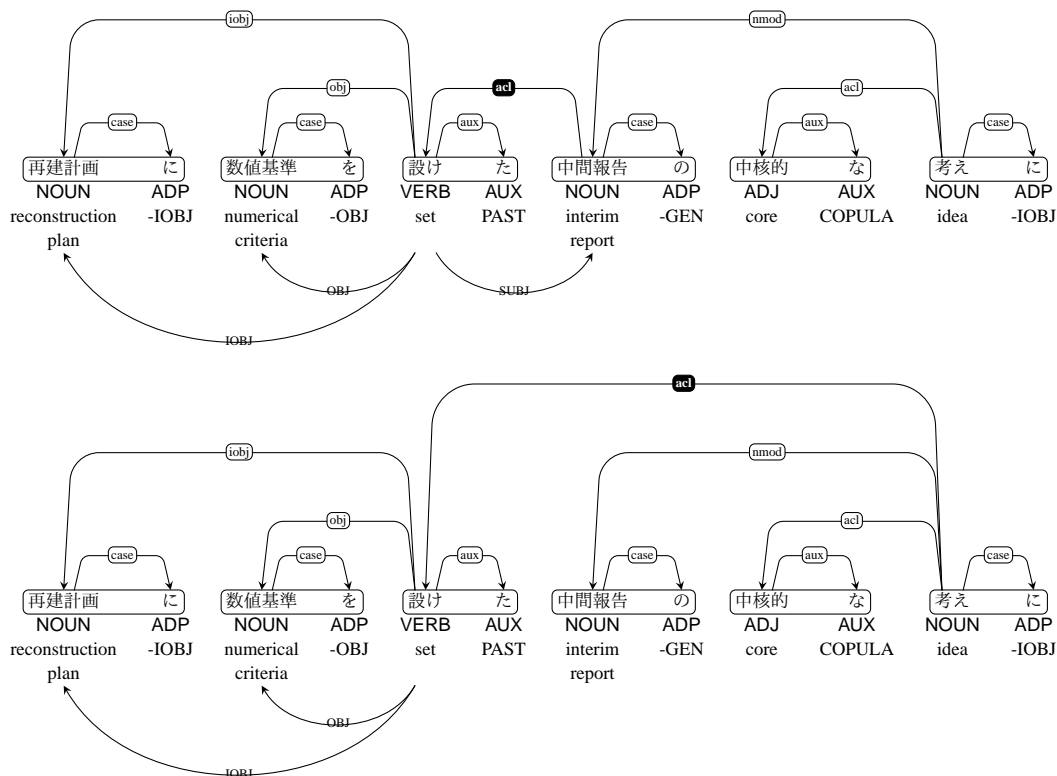
Figure 9: Ambiguous example of adjective clause (`acl`) (LUW delimitation).
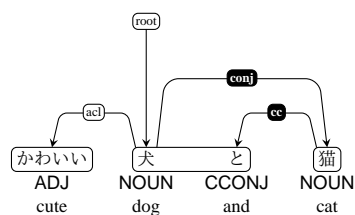

Figure 10: Simple coordinate structures.

of BCCWJ.[6]

UD Japanese-KTC is the first UD annotation schema data for Japanese. The data is based on Mainichi Newspaper in 1995. Though the annotation is available under the Creative Commons License (CC BY-SA), users are required to purchase the original text data CD-ROM.[7]

UD Japanese-GSD and UD Japanese-PUD data are based on open data. The original text is also under Creative Commons License (CC BY-NC-SA for UD Japanese-GSD and CC BY-SA for UD Japanese-PUD).

UD Japanese-Modern is based on CHJ. The copyright of the original text has already expired. We will provide the data under open license.

A newswire company asked us to provide a million scale newspaper texts for the UD Japanese project. We will anno-

tate BCCWJ-like morphological information and *bunsetsu*-based syntactic structures for the data. The data will published on the Linguistic Data Consortium by the newswire company.

4. Conclusion

This paper presented our work to build UD compatible language resources for Japanese. We discussed the issues of using the UD annotation scheme on Japanese language. We are still working on conversion rules around word delimitation, clause definitions and coordinate structure.

Table 1 presents the plans to develop UD Japanese resources.

UD Japanese-KTC is based on the UD schema of version 1.0. The data is derived by conversion rules based on phrase structure trees. We are adapting the conversion rules to the version 2.0 schema. The newer version will be released on September 2018.

UD Japanese-GSD and **UD Japanese-PUD** are not based on UniDic word boundaries (SUW/LUW). We are annotating the word boundaries of SUW and LUW with morphological information in UniDic definition. The next step is to annotate *bunsetsu*-based dependency structures based on BCCWJ-DepPara schema.

UD Japanese-BCCWJ and **UD Japanese-Modern** are developped based on version 2.0 schema. These two data will be released in March 2018.

5. Acknowledgements

---

[6]pj.ninjal.ac.jp/corpus_center/bccwj/
en/dvd-index.html
[7]www.nichigai.co.jp/sales/mainichi/
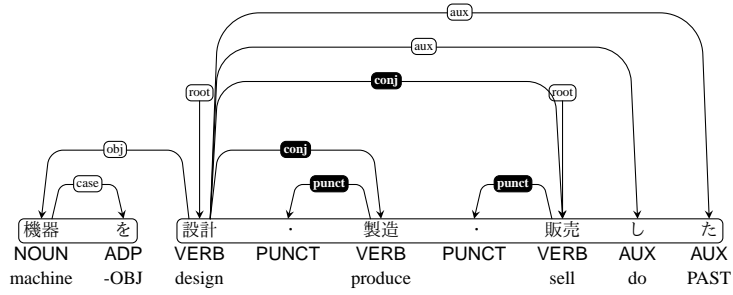mainichi-series.html

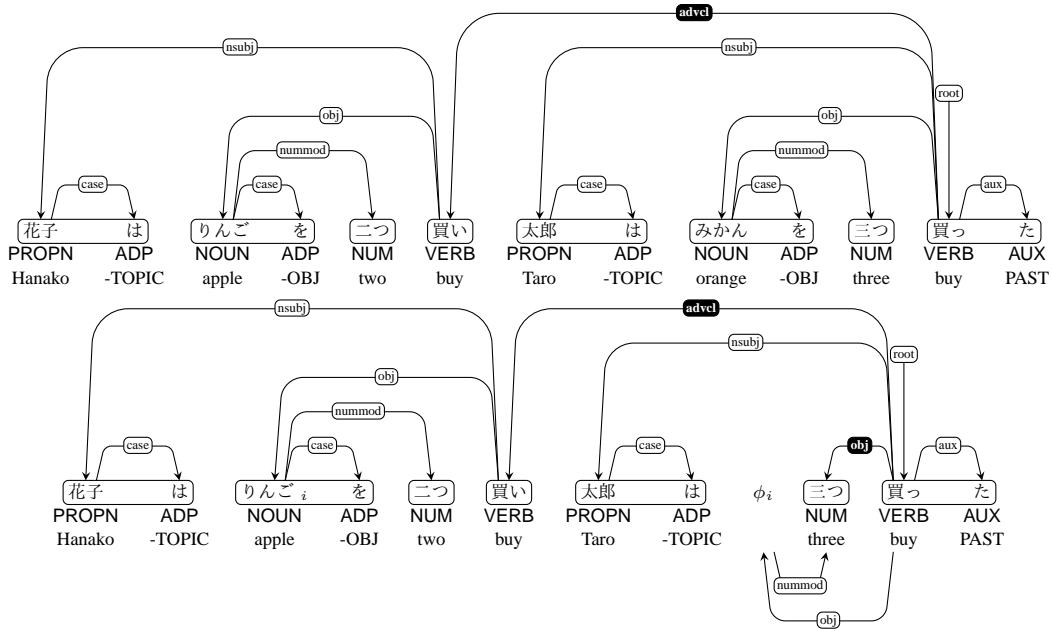Figure 11: Coordinate structure with three constituents.

Figure 12: Nominal phrase ellipsis in the non-constituent conjunct coordination.

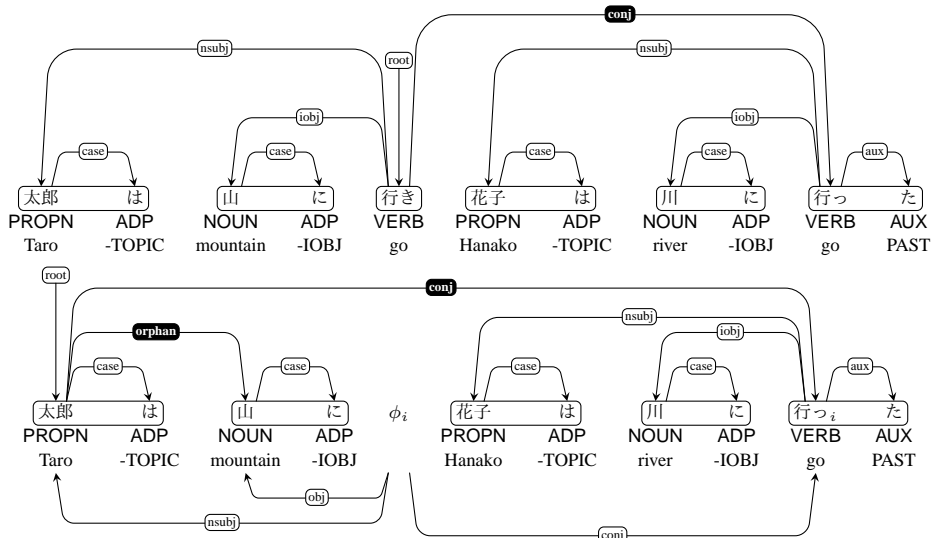Figure 13: Predicate ellipsis in the non-constituent conjunct coordination.

Center for Corpus Development, NINJAL.

6. Bibliographical References

Asahara, M. and Matsumoto, Y. (2016). BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Cor-

pus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pages 49–58.

Haspelmath, M. (2011). The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Hattori, S., (1960). *Gengogaku no hoho*, chapter Fuzokugo to Fuzokukeishiki, pages 461–491. Iwanami Shoten. (in Japanese).

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese Text Corpus with Predicate-argument and Coreference Relations. In *LAW '07: Proceedings of the Linguistic Annotation Workshop*.

Kanayama, H., Torisawa, K., Mitsuishi, Y., and Tsujii, J. (2000). A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 411–417.

Kurohashi, S. and Nagao, M., (2003). *Building a Japanese Parsed Corpus – while Improving the Parsing System*, chapter 14, pages 249–260. Kluwer Academic Publishers.

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.

Masuoka, T. and Takubo, Y. (1992). *Kiso Nihongo Bunpou – Kaitei Ban*. Kuroshio Publisher.

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

Miyaoka, O. (2015). *Go to wa Nani ka Saiko (Reconsidering What is the "Word"?)*. Sanseido. (in Japanese).

Ogiso, T., Kondo, A., Mabuchi, Y., and Hattori, N. (2017). Construction of the "Corpus of Historical Japanese: Meiji-Taisho Series I - Magazines". In *Proceedings of Digital Humanities 2017*.

Omura, M., Takahashi, Y., and Asahara, M. (2017). Universal Dependency for Modern Japanese. In *Proceedings of the 7th Conference of Japanese Association for Digital Humanities (JADH2017)*.

Tanaka, T. and Nagata, M. (2013). Constructing a Practical Constituent Parser from a Japanese Treebank with Function Labels. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Language*, pages 108–118.

Tanaka, T., Miyao, Y., Asahara, M., Uematsu, S., Kanayama, H., Mori, S., and Matsumoto, Y. (2016). Universal Dependencies for Japanese. In *Proceedings of LREC 2016*.

Uchimoto, K. and Den, Y. (2008). Word-level dependency-structure annotation to corpus of spontaneous japanese and its application. In *Proceedings of LREC 2008*.

Ueda, Y., Iida, R., Asahara, M., Matsumoto, Y., and Tokunaga, T. (2015). A Predicate-Argument and Coreference Relation Annotation on the 'Balanced Corpus of Contemporary Written Japanese', *gendai nihongo kakikotoba kinkou koopasu ni taisuru jutugo koukouzou kyosanshou kankei annoteeshon* (in Japanese). In *8th Corpus Japanese Linguistics Workshop*, pages 205–214.

van de Weijer, J., Nanjo, K., and Nishihara, T. (2005). *Voicing in Japanese*. Studies in Generative Grammar. Walter De Gruyter Inc.

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gökırmak, M., Nedoluzhko, A., Cinková, S., Hajič jr., J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Kanerva, J., Ojala, S., Missilä, A., Manning, C., Schuster, S., Reddy, S., Taji, D., Habash, N., Leung, H., de Marneffe, M.-C., Sanguinetti, M., Simi, M., Kanayama, H., de Paiva, V., Droganova, K., Martínez Alonso, H., Uszkoreit, H., Macketanz, V., Burchardt, A., Harris, K., Marheinecke, K., Rehm, G., Kayadelen, T., Attia, M., Elkahky, A., Yu, Z., Pitler, E., Lertpradit, S., Mandl, M., Kirchner, J., Fernandez Alcalde, H., Strnadova, J., Banerjee, E., Manurung, R., Stella, A., Shimada, A., Kwak, S., Mendonça, G., Lando, T., Nitisaroj, R., and Li, J. (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*. Association for Computational Linguistics.