BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'

Masayuki ASAHARA♢ Yuji MATSUMOTO♣
♢ National Institute for Japanese Language and Linguistics, National Institutes for the Humanities, Japan
♣ Nara Institute of Science and Technology, Japan.

Abstract

Paratactic syntactic structures are difficult to represent in syntactic dependency tree structures. As such, we propose an annotation schema for syntactic dependency annotation of Japanese, in which coordinate structures are separated from and overlaid on *bunsetsu*(base phrase unit)-based dependency. The schema represents nested coordinate structures, non-constituent conjuncts, and forward sharing as the set of regions. The annotation was performed on the core data of 'Balanced Corpus of Contemporary Written Japanese', which comprised about one million words and 1980 samples from six registers, such as newspapers, books, magazines, and web texts.

1 Introduction

Researchers have focused much attention on syntactic dependency parsing, as evidenced in the development of treebanks of many languages and dependency parsers on these treebanks. Most of the developed dependency treebanks have been word-based. However, treebanking based on *bunsetsu* (base phrase unit) has been adopted by the Japanese NLP community, due to the nature of the Japanese *bunsetsu* dependency structure, such as strictly being head-final and projective on the *bunsetsu* units.

Several annotation schemas for the *bunsetsu*-based treebanks are accessible in selected Japanese corpora. First is the Kyoto Text Corpus Schema(hereafter KC)(Kurohashi and Nagao, 1998), which is used for newspaper articles. Second is the Corpus of Spontaneous Japanese (Maekawa, 2003) Schema (hereafter CSJ)(Uchimoto et al., 2006).

We propose a novel annotation schema for the Japanese *bunsetsu* dependency structure, in which we also annotate coordinate and apposition structure scopes as segments. In this standard, we define the detailed inter-clause attachment guideline based on (Minami, 1974) and also introduce some labels to resolve errors or discrepancies in the upper process of *bunsetsu* and sentence boundary annotation.

We applied the annotation schema for the core data of 'Balanced Corpus of Contemporary Written Japanese' (Maekawa et al., 2014) which comprised data from newspaper(PN), books(PB),magazines(PM), white paper(OW), Yahoo! Answers(OC), and Yahoo! Blogs(OY). The core data includes 1.2 million words. We manually checked the annotation three times in seven years. This annotation schema is, thus, named BCCWJ-dependency parallel structure annotation (hereafter BCCWJ).

Contributions of the paper are summarised in the following:

- We developed a one-million-word *bunsetsu*-based dependency annotations on a balanced corpus that is comprised of newspaper, books, magazines, whitepapers, and web texts.
- We introduced a new annotation schema for coordinate structures and appositions.
- We defined inter-clause attachments by the clause type.
- We resolved the errors of the upper process (word-segmentation and POS tagging layer) in the annotation schema, such as *bunsetsu* and sentence boundaries.

Proceedings of the 12th Workshop on Asian Language Resources,
pages 49–58, Osaka, Japan, December 12 2016.

In this article, we focus on the annotation schema of coordination and apposition structures in the dependency treebank. Section 2 presents an overview of the annotation schema. Section 3 describes the details of the annotation schema on the coordination and apposition structures. Section 4 shows the inter-clause attachment annotation schema. Section 5 illustrates the basic statistics of the annotation data. Section 6 discusses the conclusion of this article.

2 Overview of the Annotation Schema

Table 1: Comparison of *bunsetsu*-based dependency structure annotation schema

Label	BCCWJ	(group)	CSJ	KC
Normal	D	-	no label	D
Parallel	D	(Parallel)	P	P
Parallel (non-constitutent conjunct)	D	(Parallel)	I	I
Apposition	D	(Apposition)	A	A
Apposition (Generic)	D	(Generic)	A2	A
Right to Left	D	-	R	undef
No attachment	F	-	undef	undef
(for *Bunsetsu*)	BCCWJ	-	CSJ	KC
Concatenate *Bunsetsu*	B	-	B+	undef
(Misc)	BCCWJ	(segment)	CSJ	KC
Filler	F	-	F	undef
Smiley	F	-	undef	undef
Sentence conjunction	F or D		C	D
Interjection	F or D	-	E	D
Vocative	Z	-	Y	undef
Disfluency/Self-correction				
(one *bunsetsu*)	D	-	D	undef
(more than one *bunsetsu*)	D	-	S(S:S1, S:E1)	undef
Non speech sound	F	-	no label	undef
Whitespace, URL	F	-	undef	undef
Inversion/non-projective	D	-	X	undef
Foreign word	D	(Foreign)	undef	undef
Archaic word	D	(Foreign)	K(K:S1,K:E1)	undef
Sentence end	Z	-	undef	undef
Grammatical error	undef	-	S	undef

We present the overview of the annotation schema of the BCCWJ by establishing a comparison with two other linguistics annotation schemas using *bunsetsu*-based dependency structure. Table 1 illustrates the comparative differences of the BCCWJ annotation schema from those in the KC and CSJ.

The BCCWJ schema defines four labels on the dependency relations: 'D' for normal dependency relation, 'B' for the concatenation to make a longer *bunsetsu*, 'F' for no dependency relation, and 'Z' marks the end of sentence (EOS).

We introduce 'segment' and 'group' to express coordination and apposition structures: Figure 1 demonstrates examples of these expressions. Segment is a region of the subsequence of words in the sentences. Group is a set of segments. Group is used for equivalence class by equivalance relations such as coordinate structures and coreference relations.

In the first example, the rounded corner squares are the conjuncts of a coordinate structure defined by the group 'Parallel'. The conjuncts are defined by the short unit word sequences in the BCCWJ, which

is the smallest morpheme unit in the corpus. Therefore, the conjunct boundary can be defined within a *bunsetsu*. In that case, the hyphenation is used to indicate NOT *bunsetsu* boundary. As illustrated in the second example in Figure 1, the dotted rounded corner squares represent the conjuncts of an appositional structure in the narrow sense defined by the group 'Apposition'. We also define other segment and group in 'Generic', which stands for an apposition structure in the broad sense.

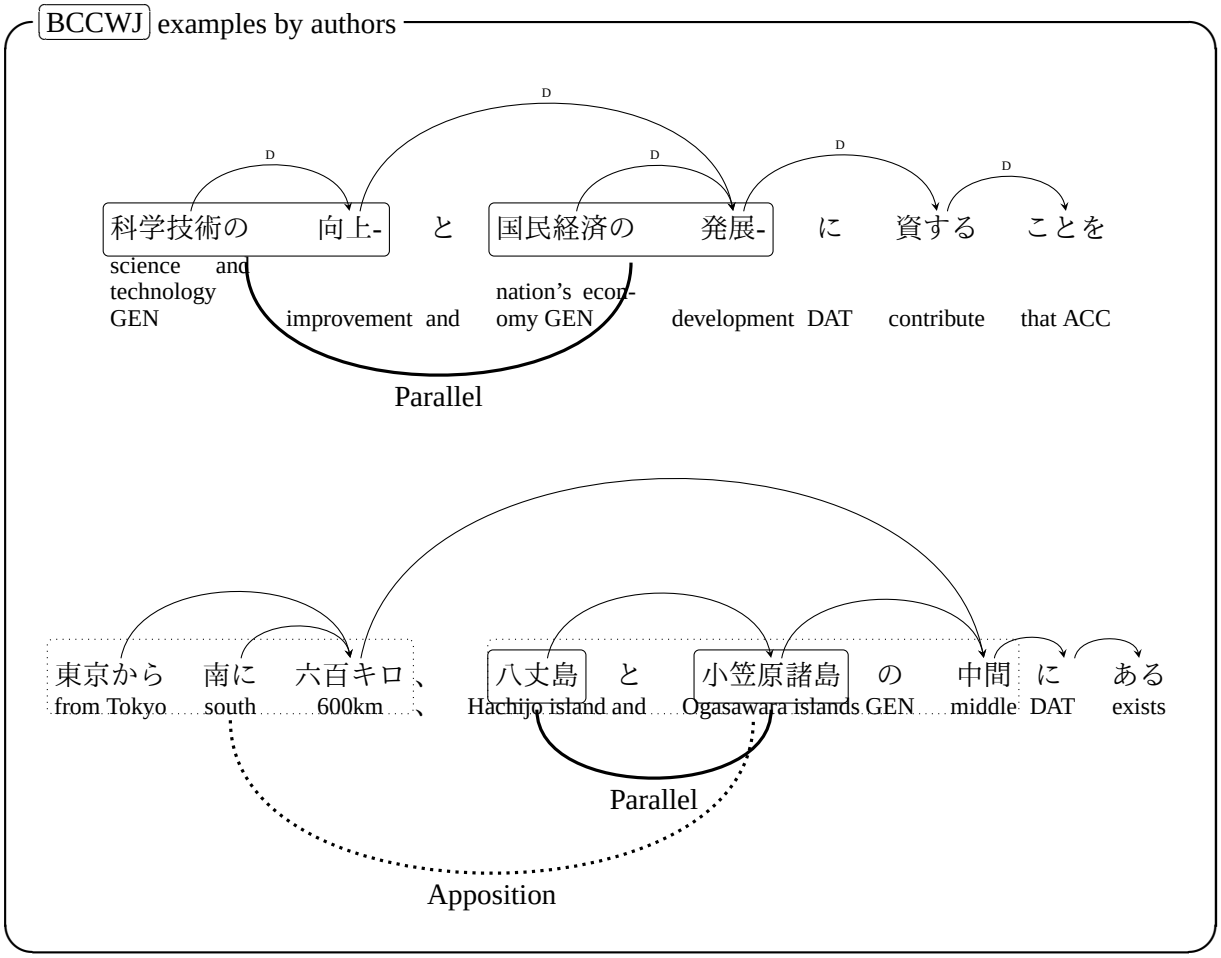


Figure 1: The assignment of 'segment' and 'group' to express coordinate and apposition structure

First, we present the differences of coordination and apposition structures among the annotation standards. In the KC standard, the label 'P' is defined for coordinate structure relation, and the label 'A' is defined for apposition structure relation. For non-constituent conjuncts, the label 'I' is used to avoid non-projective arcs in the dependency structure. The CSJ standard is based on KC, but it further defined apposition structures. The CSJ divide the apposition structure into a narrow sense with label 'A' and a broad sense with the label 'A2': The label 'A2' represents the generic name for the part-of relation or the numerical expression for the attribute-value relation in an apposition structure. In the BCCWJ standard, we avoid expressing coordination and apposition structures by their dependency relation, because these structures in dependency would make the dependency tree structure skewed. As presented above, we assign 'segment' and 'group' to each of the labels, namely, 'Parallel', 'Apposition', and 'Generic'. The subsequent section 3 provides in-depth explanation on this.

Second, we present the labels for the case to violate the projective or strictly head final constraints. The KC standard does not define special labels for such violation, because KC analyses texts that are derived from newspaper articles; therefore the dependency structures do not tend to violate these constraints. In the CSJ standard, the label 'X' is defined for the inversion of a non-projective arc, whereas the label 'R' represents the relation from right to left. In the BCCWJ standard, though both

non-projective structure and right-to-left relation are permitted, we use the label ‘D’ to define a normal dependency relation.

Third, we present the labels to resolve errors or discrepancies in the upper process. In the KC standard, all annotations are performed in the same research group. Hence, they do not define any special labels for these errors or discrepancies. However, in the CSJ standard, the discrepancy of *bunsetsu* boundaries is inherent to the original CSJ source, namely, speech. As such, the *bunsetsu* boundaries can be inserted by a speech pause or an interval. In the syntactic layer, we sometimes need to concatenate more than one item into one *bunsetsu*. In that case, the label ‘B+’ is introduced. In the BCCWJ standard, the *bunsetsu* and sentence boundaries are annotated by other research group based on morphology. As a result of some discrepancies between the morphology and syntactic layer research group, we have decided to introduce the labels ‘B’ for the *bunsetsu* and ‘Z’ for sentence boundaries. Note that, we permit nested sentence in the BCCWJ standard.

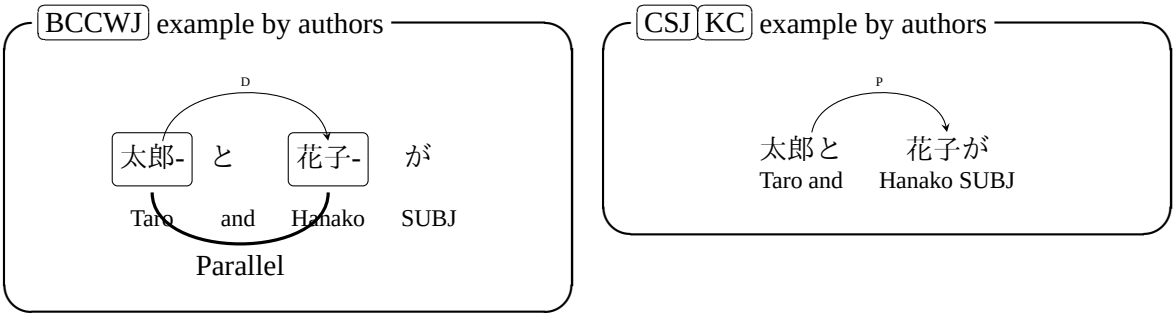
Fourth, we present the labels to avoid annotating the dependency relation. In the KC standard, the target data is from newspaper articles and tends to be normative. Therefore, no special label is assigned to syntactic dependency relation. In contrast, the CSJ standard defines the label ‘D’ for disfluency, ‘F’ for filler, ‘C’ for conjunction, ‘E’ for interjection, ‘Y’ for call, ‘N’ for no dependency attachment, and ‘K’ for archaic words. In the BCCWJ standard, we define the label ‘F’ for filler or no dependency attachment and ‘Z’ for sentence end or call. We also define the segments of ‘Foreign’ for the foreign language region and ‘Disfluency’ for the disfluency region. In the segments, the dependency attachment is to the neighbouring right *bunsetsu*.

3 Examples of Coordination and Apposition Structures

In this section, we exemplify the dependency annotation standards of coordination and apposition.

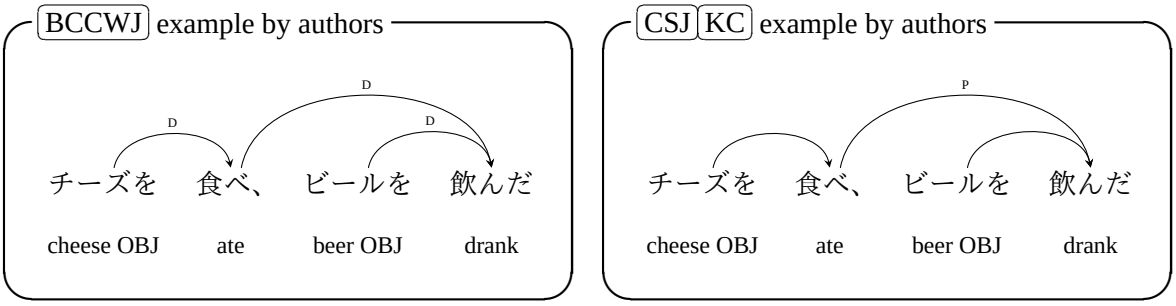
3.1 Coordination of nominal phrases

In the BCCWJ standard, coordinate structures of nominal phrases are represented by segments with the label ‘Parallel’ with grouping. The dependency arc is labelled ‘D’. However, in the case of CSJ and KC, the coordination of nominal phrases is expressed by the dependency arc labelled ‘P’.



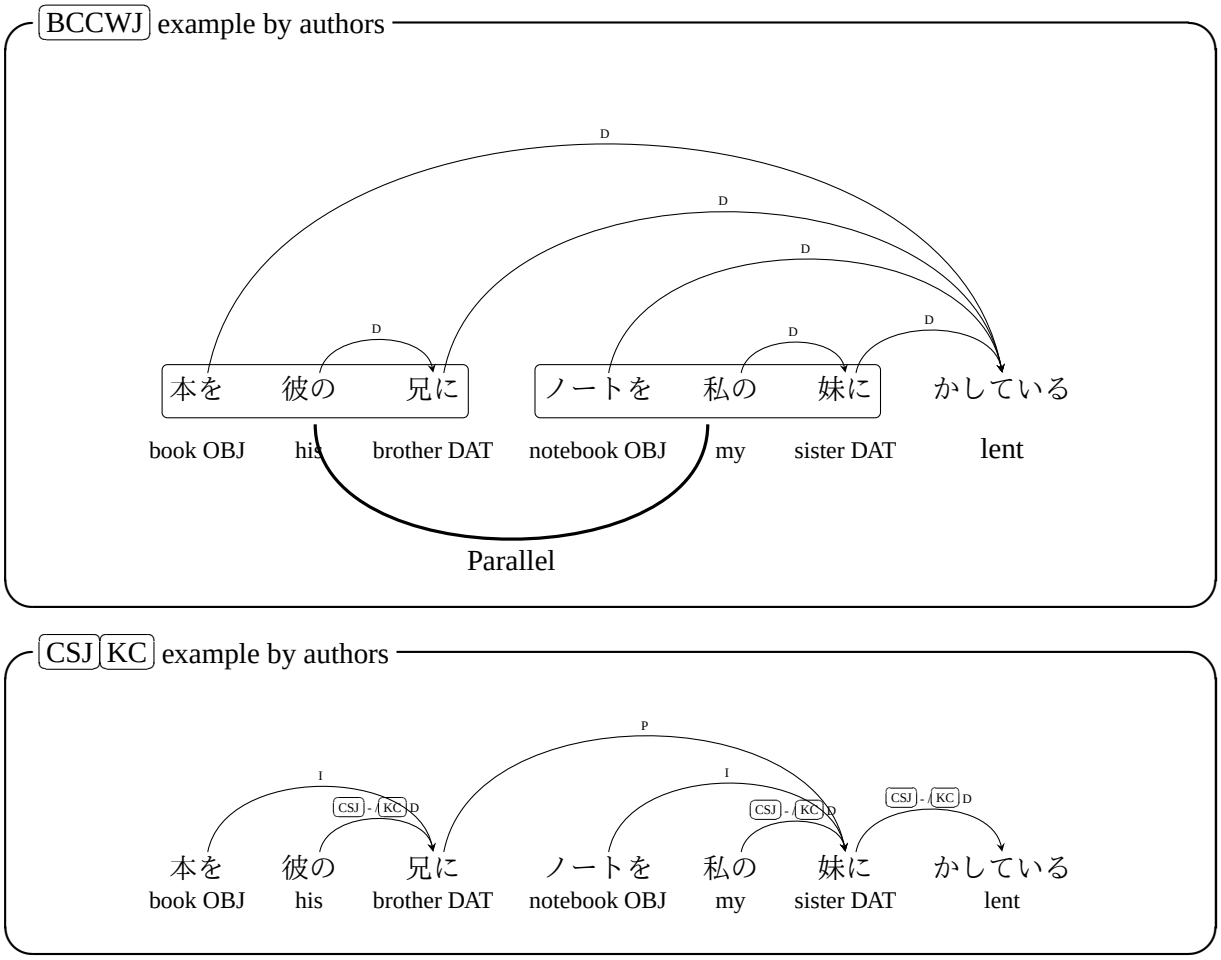
3.2 Predicate coordination

Since the identification of a predicate coordination is difficult, the BCCWJ standard does not focus on using labels or segments to define these structures. We regard a predicate coordination as a normal dependency attachment (labelled ‘D’). As a comparison, the CSJ KC standards label ‘P’ for predicate coordination.



3.3 Non-constituent coordination

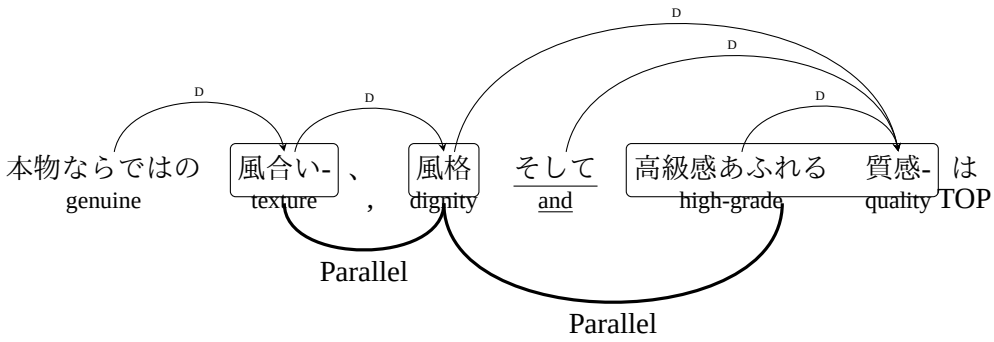
The non-constituent coordinate structure may violate projective or double ‘を (wo: object marker)’ constraints. The CSJ KC standards define the label ‘I’ to show the scope of such coordination and to maintain projective constraints. However, in the BCCWJ standard, we only define the segments on non-constituent coordination and normal dependency attachment with the label ‘D’.



3.4 Coordination with more than two constituents

In the BCCWJ standard, coordination with more than two constituents is expressed by segments which are attached to the rightmost *bunsetsu* within the right adjacent coordinate constituent with the label ‘D’. In the example, ‘風合い (texture)’, ‘風格 (dignity)’, and ‘高級感あふれる質感 (high-grade quality)’ are expressed by grouping the segments. The conjunction ‘そして (and)’ (underlined in the below figure) attaches the rightmost *bunsetsu* within the rightmost coordinate constituent with the label ‘D’.

BCCWJ 00033_B_PB35_00013 in BCCWJ



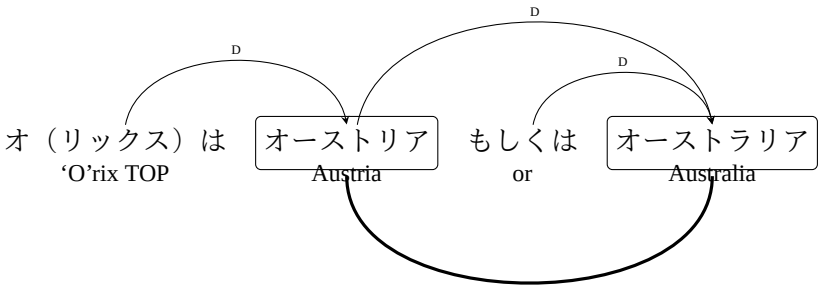
In contrast, the CSJ standard labels ‘C’ for the conjunction. However, the illustration is omitted due to space limitation.

3.5 Forward sharing

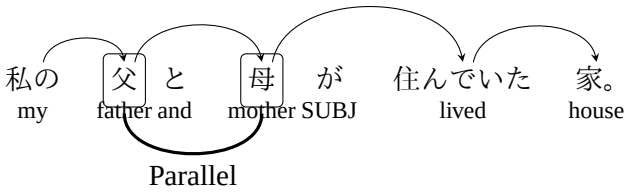
Forward sharing is a unique trait of a coordinate structure, in which one *bunsetsu* attaches all constituents in the coordination.

In the example below, ‘オ（リックス）は (Orix TOP)’ attaches both ‘オーストリア (Austria)’ and ‘オーストラリア (Australia)’. Attaching the leftmost constituent of the coordination means forward sharing. Note that since Japanese language is essentially a strictly final language, we are not concerned about backward sharing.

BCCWJ 00620_B_OC06_02188 in BCCWJ



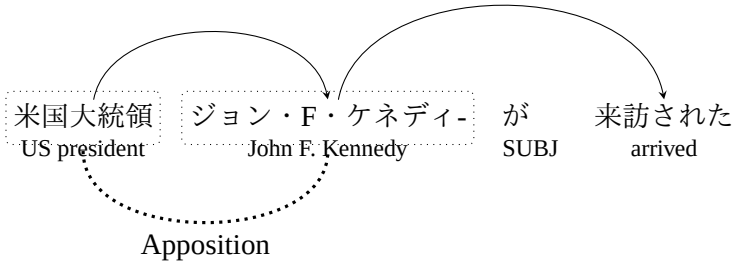
BCCWJ example by authors



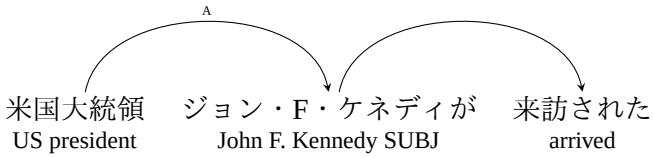
3.6 Apposition in the narrow sense

In the BCCWJ standard, apposition structures are also expressed by segments and groups. The example below illustrates that the appositive noun phrases, namely, ‘米国大統領 (US president)’ and ‘ジョン・F・ケネディ(John F. Kennedy)’ are grouped and labelled ‘Apposition’. However, in the KC CSJ standards, these appositive noun phrases are expressed by the dependency arc with the label ‘A’.

BCCWJ example by authors



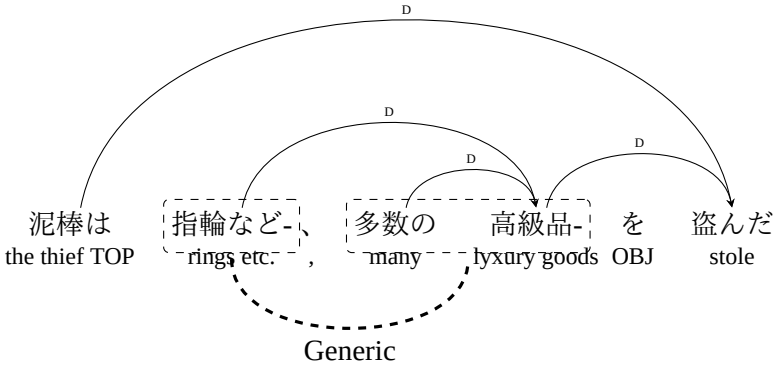
KC CSJ example by authors



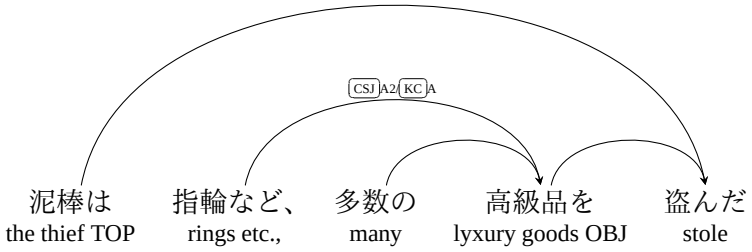
3.7 Generic – Apposition in a broad sense

In the KC standard, the apposition label ‘A’ is defined in the broad sense, which includes the apposition between examples and generic expressions, and between examples and numeral expressions (attribute-value relation). In comparison, the CSJ standard restricts the label ‘A’ to the narrow sense of apposition, whereas the label ‘A2’ represents apposition in the broad sense.

BCCWJ example by authors



CSJ KC example by authors



4 Inter-clause attachment

Inter-clause attachment is one of issues of annotation consistency among the annotators. We use subordinate clause classes (Minami, 1974) to determine the inter clause attachments. Table 2 shows the annotation schema. The subordinate clause is classified into three classes (i.e. A, B, C). The classes define the scope of the constituents.

The most frequent inconsistency is the attachment of case markers. Whereas the subjective *"-ga"* can attach to class B and C, the topicalization *"-ha"* can attach only to class C. Other case markers such as the objectives *"-wo"* and *"-ni"* can attach to all classes.

In the definition, the annotators need to judge the usages of *"-te"* and the conjunctive form. However, we did not record the judgment. In our future work, we will annotate the class of clauses.

5 Basic Statistics of the BCCWJ-DepPara

In this section, we present the basic statistics of the BCCWJ-DepPara data. Table 3 shows the number of sample files, short unit words (SUW), long unit words (LUW), *bunsetsus*, the dependency arc labels of 'D', 'B', 'F', 'Z', and end of sentences ('EOS'). The label 'F' in both OW and OY registers tends to be larger than of those in the other registers. The OW register includes many item markers, whereas the OY register includes many smiley strings, all labelled 'F'. Since we permit nested sentences, the number of the label 'Z' is more than the number of 'EOS'[1].

Table 4 shows the basic statistics of the coordination and apposition structures. The register 'OW' tends to include many 'Parallel' annotations. Because coordinate structures permit more than two constituents, the average number of constituents (seg/grp) of coordinate structures ranges from 2.19-2.35. However, since the 'Apposition' and 'General' labels are paired constituent structures, the average number of the constituents of these labels is nearly 2.00. Some exceptions of apposition expressions are caused by paraphrasing more than one time in several forms.

6 Conclusion

This article presents the annotation standard of dependency and coordination structures in the BCCWJ-DepPara. In the standard, the coordinate structure was taken out of the dependency structure, and it was, then, expressed by segments and groups.

Due to space limitation, we have omitted the annotation standard related to the inter-clause attachment, in which the scopes of phrases or clauses are defined by Minami's clause classes(Minami, 1974). Though the annotator used the clause classes for judgement, we did not annotate the clause classes on the corpus. Our current work is to annotate the clause classes based on the standard of 'Japanese Semantic Pattern Dictionary – Compound and Complex Sentence Eds.' (Ikehara, 2007).

The data of the BCCWJ-DepPara are accessible at `http://bccwj-data.ninjal.ac.jp/mdl/` for any purchaser of the BCCWJ DVD edition.

Parsing models should be adopted for the BCCWJ standard. (Iwatate, 2012) proposed a model that involves the BCCWJ standard, in which the dependency attachments and coordinate structures are estimated by a dual decomposition method.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18061005, JP23240020, JP25284083, and JP15K12888. The work was also supported by the NINJAL collaborative research project 'Basic Research on Corpus Annotation'. We also appreciate the annotators and contributors of the BCCWJ-DepPara.

[1]Whereas the label 'Z' defines both inner and outer sentence ends, the label 'EOS' defines only the outer sentence ends.

Table 2: Minami's clause classes and their attachments

Classes of subordinate clause		class A					class B												class C					
	Constituent	‡conjunctive forms of adjective）	‡repetition of conjunctive forms	†-*te* (collateral circumstance）	*-tsutsu* (collateral circumstance）	*-nagara* (collateral circumstance）	†-*te* (other than collateral, sequence or citation)	*-to* (assumption)	*-nagara* (contradictory)	*-node* (reason or focus)	*-noni* (reason or focus)	*-ba* (assumption)	*-tara* (assumption)	*-nara* (assumption)	*-temo* (assumption)	†-*te* (sequence)	‡conjunctive forms (sequence)	*-zu* or *zuni*	*-naide*	*-ga*	*-kara* (reason)	*-keredo* (coordination)	*-shi* (coordination)	†-*te* (citation)	‡conjunctive form (coordination)
non-predicative constituents	Noun + Case Particle other than subjective (-ga)	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Stative adverb	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Degree adverb	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Class A clauses	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Subjective (*-ga*)	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Temporal modifier	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Locative modifier	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	jitsuni, tonikaku, yahari, etc.	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Evaluative modifier	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Class B clause	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Topicalization *-ha*	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	osoraku, tabun, masaka, etc.	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Class C clause	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
predicative constituents	Declinable word without dummy noun	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Causative form	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Passive form	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Receiving form	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Respect form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Polite form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Negation form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Past form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Declinable word with dummy noun	+	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
	Volition form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
	Conjecture form	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+

'+' means that the constituent can exist in the clause class
'-' means that the constituent cannot exist in the clause class
†-*te* form is classified into 4 classes based on usages
‡conjunctive form is classified into 4 classes by the usages

References

Satoshi Ikehara. 2007. Semantic typology pattern dictionary. http://asia.shachi.org/resources/862.

Table 3: Basic statistics of the BCCWJ-DepPara (word and dependency labels)

register	samples	SUW	LUW	*Bunsetsu*	'D'	'B'	'F'	'Z'	'EOS'
PN	340	308,504	224,140	116,955	96,892	1,652	2,017	16,394	16,042
					(82.8%)	(1.4%)	(1.7%)	(14.0%)	
PB	83	204,050	169,730	84,733	72,340	1,091	1,425	9,877	9,678
					(85.3%)	(1.2%)	(1.7%)	(11.7%)	
PM	86	202,268	159,883	83,077	67,618	1,187	1,629	12,643	12,542
					(81.4%)	(1.4%)	(2.0%)	(15.2%)	
OW	62	197,011	129,646	68,449	59,320	359	2,927	5,843	5,825
					(86.6%)	(0.5%)	(4.3%)	(8.5%)	
OC	938	93,932	78,770	36,740	29,753	323	428	6,236	6,110
					(81.0%)	(0.9%)	(1.2%)	(17.0%)	
OY	471	92,746	75,242	38,576	29,650	337	1,501	7,088	7,059
					(78.9%)	(0.9%)	(3.9%)	(18.4%)	

The percentages are the number of labels { 'D', 'B', 'F', and 'Z'} / the number of *bunsetsus*.

Table 4: Basic statistics of the BCCWJ-DepPara (coordination and apposition structures)

register	Parallel			Apposition			General		
	seg	grp	seg/grp	seg	grp	seg/grp	seg	grp	seg/grp
PN	8,446	3,844	2.19	3,440	1,713	2.01	1,026	513	2.00
PB	4,640	2,060	2.25	704	352	2.00	304	152	2.00
PM	5,513	2,454	2.24	1,313	651	2.02	280	140	2.00
OW	10,709	4,613	2.32	1,326	662	2.00	656	328	2.00
OC	1,586	715	2.21	292	146	2.00	62	31	2.00
OY	1,603	682	2.35	262	131	2.00	58	29	2.00
Total	32,497	14,368	2.26	7,337	3,655	2.01	2,386	1,193	2.00

Masakazu Iwatate. 2012. *Development of Pairwise Comparison-based Japanese Dependency Parsers and Application to Corpus Annotation, Chapter 7: Joint Inference of Dependency Parsing and Coordination Analysis Using a Dual Decomposition Algorithm*. Ph.D. thesis, Graduate School of information Science, Nara Institute of Science and Technology, Japan.

Sadao Kurohashi and Makoto Nagao. 1998. Building a japanese parsed corpus while improving the parsing system. In *Proceedings of The 1st International Conference on Language Resources & Evaluation*, pages 719–724.

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.

Kikuo Maekawa. 2003. Corpus of spontaneous japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*.

Fujio Minami, 1974. 現代日本語の構造 *(Structure of Contemporary Japanese)*, pages 128–129. 大修館書店 (Taishukan publishing).

Kiyotaka Uchimoto, Ryoji Hamabe, Takehiko Maruyama, Katsuya Takanashi, Tatsuya Kawahara, and Hitoshi Isahara. 2006. Dependency-structure annotation to Corpus of Spontaneous Japanese. In *Proceedings of 5th edition of the International Conference on Language Resources and Evaluation*, pages 635–638.