Japanese clause classification annotation
on the ‘Balanced Corpus of Contemporary Written Japanese’

Satomi Matsumoto♡, Masayuki Asahara♣, Setsuko Arita♢
♡ Ritsumeikan University
♣ National Institute for Japanese Language and Linguistics
♢ Ritsumeikan University

Abstract

Inter-clause syntactic and semantic structures are important to process semantic reasoning. This paper presents clause boundaries and class annotation on the ‘Balanced Corpus of Contemporary Written Japanese’. The annotation is based on the Tori-Bank labelset, which provides the most fine-grained clause classes. We reformulated the legacy syntactic pattern into a syntactic-dependency-based pattern. Two annotators modified the automatically extracted clause boundary candidates. In this study, we investigate the patterns of disagreement in the annotation.

1. Introduction

Clause boundary detection and classification are important issues in the detection of causal and temporal relations between two events. This paper presents clause boundaries and annotations applied to the ‘Balanced Corpus of Contemporary Written Japanese’ (BCCWJ) (Maekawa et al., 2014); the annotation is based on the surface pattern of the morphemes. Then, the clause boundaries are categorized by their syntactic and semantic classes in Tori-Bank (Ikehara, 2007). Though the clause classes were designed for Japanese-English machine translation, the nevertheless serve as a basis for reasoning on inter-clause relations.

The original Tori-Bank patterns are based on a legacy POS tagset. We reproduced the clause patterns and adapt them to the UniDic POS tagset and word segmentation schema with Bunsetsu-based dependency structure. Then, we annotated the clause boundaries for syntactic and semantic classes on newspaper samples from the ‘BCCWJ’. Next, we evaluated the syntactic or semantic classes of clause boundaries that tended to show annotation discrepancies.

2. Annotation Schema for Clause Classification

The clause boundary classification is based on the Tori-Bank schema (Ikehara, 2007). Tori-Bank is a corpus developed at Tottori University in 2007 in order to compile a Japanese semantic pattern dictionary for compound and complex sentences. The clause boundary patterns are hierarchically defined, in four layers. The top level of the classification consists of Nominal Clauses (補足節:HS), Adnominal Clauses (名詞修飾節:MS), Adverbial Clauses (副詞節:FU), and Coordinate Clauses (並列節:HR). The second level of the classification is made up of 26 classes, and the third level is made up of 52 classes. We use the third-level labels for our annotation. Below, we describe the Tori-Bank clause labels and provide annotated examples from BCCWJ.

2.1. Nominal Clauses

Nominal clauses (補足節:Hosoku-Setsu, HS) are classified into noun clauses (HSa), interrogation clauses (HSb), and quotation clauses (HSc) at the second level.

Noun clauses (HSa) are then classified at level three, into *koto* (コト) form (HSa100), *no* (ノ) form (HSa200), *tokoro* (トコロ) form (HSa300), and clause + case particle form (HSa400). Below, (1) is an example of HSa100. ‘こと’ *koto* appears at the end of the clause. (2) is an example of HSa200. ‘の’ *no* appears at the end of the clause. (3) is an example of HSa300. ‘ところ’ *tokoro* appears at the end of the clause. These words are relative pronouns. (4) is the example of HS400. This is a zero (relative) pronoun before the case particle “に”.

(1) 制度を　育て上げることが ぜひとも 必要。
seido-wo sodateagerukoto-ga zehitomo hitsuyou.
‘It is really needed to develop the system.’
`HSa100 (`*no*`),`
`[BCCWJ Sample ID: PN1c_00001]`

(2) 赤字額が　最も　多い <u>の</u> は
akajigaku-ga mottomo ooi no-ha
東京都大江戸線の　三百十一億円
toukyouooedosen-no sanbyakujuuichiokuen
だった
datta
‘Oedo Line in the Tokyo Metropolitan had the largest deficits at 31.1 billion yen.’
`HSa200 (`*koto*`), [PN2e_00001]`

(3) ボールが けれる <u>ところ</u> まで 回復している
booru-ga kereru tokoro-made kaifukushiteiru
‘I have recovered to the point of being able to kick a ball.’
`HSa300 (`*tokoro*`), [PN2f_00002]`

(4) 実務レベルで　協議する <u>に</u>　とどまっている
jitsumureberu-de kyougisuru-ni todomatteiru
‘Stop at the negotiation at the practical level.’
`HSa400 (zero pronoun),`
`[PN3a_00002]`

Interrogation clauses (HSb) are subclassified into alternative/choice questions (HSb100), in (5), and question with interrogative words (HSb200), in (6).

(5) 復調の めどがたっているの か 表情は、
fucchouno medogatatteiruno-ka hyoujou-ha
明るかった。
akarukatta
‘Perhaps due to the increase in the possibility of recovery, his facial expression brightened.’
`HSb100 (alternative/choice question), [PN2f_00002]`

(6) 減税額が 実際に どの程度 違うの か
genzeigaku-ga jissai-ni donoteido chigauno-ka
検証してみよう。
kenshoushitemiyou
‘Let ’ s verify to what extent the amount of tax reduction actually differs. ’
`HSb200 (question with interrogative word), [PN4c_00002]`

Quotation clauses (HSc) are classified into direct quotations (HSc100), as in (7), and indirect quotations (HSc200), as in (8). We classified them by the presence or absence of quotation marks, ‘「」’.

(7) 「... 卒業証書は 出す」と 言った
“... sotugyoushousho-ha dasu”-to itta
‘ “··· I will submit your graduation certificate,” he said.’
`HSc100 (direct quotation), [PN4g_00003]`

(8) 目立つ 生徒だったと いう。
medatsu seitodata toiu
‘It is said that he was a prominent student.’
`HSc200 (indirect quotation), [PN1c_00001]`

2.2. Adnominal Clauses

Adnominal clauses (名詞修飾節: Meishishushoku Setsu, MS) are classified into relative clauses (MSa), apposition clauses (MSb), and clauses with contractive expressions (MSc), clause with functional expressions (MSd), and clause with collocational expressions (MSe).

Relative clauses (MSa) are those in which the modifiee is an argument of the subordinate clause’s end predicate. Relative clauses are classified into restrictive (MSa100), as in (9), and non-restrictive (MSa200), as in (10). These two then each subdiscriminated by whether the modifiee is a normal noun (MSa100) or a proper noun (MSa200).

(9) 呼び出して 注意する 先生も いたが
yobidashite chuuisuru sensei-mo ita-ga
‘Although there were also teachers who summoned and warned students.’
`MSa100 (relative, restrictive), [PN1c_00001]`

(10) この日は 腰の 重い 安芸乃島に
konohi-ha koshi-no homoi akinoshima-ni
快勝。
kaishou.
‘Achieved an easy win over Akinoshima, who was slow to act, today.’
`MSa200 (relative, non-restrictive), [PN1e_00004]`

Apposition clauses (MSb) is that the modifiee has an appositive relation with the clause.

(11) 低迷する 日本経済の 「負の側面」を
teimeisuru nihonkeizai-no “funosokumen”-wo
象徴する 結果に なった
shouchousuru kekka-ni natta
‘Became results that symbolized the “ negative sides ” of the sluggish economy of Japan.’
`MSb (apposition), [PN1e_00001]`

Contractive adnominal clauses (MSc) are adnominal clauses that are neither relative nor apposition clauses.

(12) 試合は 1 点を 争う 展開。
shiai-ha itten-wo arasou tenkai
‘The match developed into a competition for one point.’
`MSc (contractive), [PN1e_00003]`

Functional adnominal expressions (MSd) are pairings of an adnomial clause and a modifee to express a functional meaning. They are subclassified into functional adnominal expressions with relative pronouns (MSd100), as in (13), functional sentence-end expressions (MSd200), as in (14), idiomatic expressions (MSd300), as in (15), and functional adnominal expressions in adverbial usage, as in (16).

(13) サラダと 聞いて
sarada-to kiite
思い浮かべた ものは、野菜サラダ
omoiukabetamonoha, yasaisarada
‘When I heard the word “salad,” vegetable salad came into my mind.’
`MSd100 (functional adnomial expression with relative pronoun), [PN1a_00002]`

(14) 廃止の 理由は、 授業時間を
haishi-no riyuu-ha, jugyoujikan-wo
確保する ため
kakuhosurutame
‘The reason for the abolition is to secure class hours.’
`MSd200 (functional sentence end expression), [PN1a_00002]`

(15) 何かに 没頭できる という 点では、
nanika-ni bottoudekiru toiu tendeha,
‘In terms of being able to be absorbed in something,’
`MSd300 (idiomatic expression), [PN3b_00004]`

(16) 焦げ付きを 懐具合に 見合った 範囲に
kogetsuki-wo futokoroguai-ni miatta han’i-ni
抑えたい
osaetai
‘I wish to limit bad debts to a range commensurate with my financial standing.’

MSd400 (functional adnominal expression in adverbial usage), [PN1g_00002]

Example (17) shows a collocational expression with the pattern ‘predicative + conjunctive + の *no*’.

(17) ずっと 入院したまま の 例も 珍しく
zutto nyuuinshitamama-no rei-mo mezurashiku
なかった
nakatta
‘Cases where the patient stayed hospitalized for a long time were not rare.’
MSe (predicative + conjunctive + の *no*), [PN3a_00003]

2.3. Adverbial Clauses

The adverbial clause (副詞節: Fukushi Setsu, FU) is classified by semantic features.
First, (FUa) are temporal clauses, indicating a time-point or the duration of an event, as in (18). The, (FUb) are causal clauses, indicating a cause or result, as in (19).

(18) バックが 暗い中、 ストロボの 光が 鳥に
bakku-ga kurainaka sutorobo-no hikari-ga tori-ni
集光して、
shukou-shite,
‘Against a dark background, concentrate the stroboscope on the bird,’
FUa100 (temporal), [PN1d_00001]

(19) 新聞で 報道され、 逃げ切れないと 思って
shinbun-de houdousare nigekirenai-to omotte
自首した
jishushita
‘He surrendered as he thought he would not be able to escape after being featured in the news.’
FUb100 (causal), [PN1f_00002]

Next, (FUc) are conditional clauses, subclassifiable into nomothetic conditionals (FUc100), as in (20), accidental conditionals (FUc200), as in (21), and imaginary conditionals (FUc300), as in (22).

(20) 世界から 見れば、 一地方大学。
sekai-kara mire-ba ichichihoudaigaku.
‘A regional university from the perspective of the world.’
FUc100 (nomothetic conditional), [PN3b_00001]

(21) 自分が 注意を したら
jibun-ga chuui-wo shitara
逃げ出したことなどを
nigedashitakotonado-o
‘It would not have escaped, if you had paid attention.’
FUc200 (accidental conditional) [PN1b_00003]

(22) 1人に 力点を 置くなら、 断食や
hitori-ni rikiten-wo oku-nara, danjiki-ya
ダイエットには 格好の 状況と
daietto-ni-ha kakkou-no joukyou-to
言えるだろう
ierudarou
‘It could be a situation suitable for fasting and dieting, if the emphasis is put on one person.’
FUc300 (imaginary conditional), [PN3b_00004]

Subsequently, (FUd100) captures attendant circumstances, as in (23), and (FUd200) is aspectual clauses, as in (24). These two are under the ‘attendant circumustances’ label at the second level label (FUd).

(23) 実効性を 考慮して 慎重に
jikkousei-wo kouryoshite shinchou-ni
決めるべきだ
kimerubekida
‘One should decide cautiously, while taking effectiveness into consideration.’
FUd100 (attendant circumstances), [PN3g_00001]

(24) 今と 同じように、 子どもの
ima-to onajiyou-ni, kodomo-no
家庭環境を 把握する
kateikankyou-wo haakusuru
‘Figure out the children’s family environment, just like now.’
FUd200 (aspectual), [PN1a_00002]

The, (FUe) are contrastive clauses, as in (25); (FUf) are objective clauses including necessity and intention, as in (26); and (FUg) conveys degree of action or state, as in (27).

(25) ■■さんは 病院に 運ばれたが、
XX-san-ha byoouinn-ni hakobareta-ga,
間もなく 死亡した
mamonaku shiboushita
‘Mr. X was sent to the hospital, but he died shortly afterwards.’
FUe (contrastive), [PN4f_00001]

(26) 不良債権の 最終処理という
furyousaiken-no saishushoritoiu
外科大手術を するには 大量の
gekadaishujutu-wo suru-ni-ha tairyou-no
輸血が 必要で
yuketu-ga hitsuyou-de
‘Massive blood transfusion is necessary for major surgeries, which are regarded as the final treatment of bad debts.’
FUf (objective), [PN1b_00004]

(27) 日本は アメリカに 言われるまでもなく、
nihon-ha amerika-ni iwarerumademonaku,
国内経済の 安定を 第一に 考えて
kokunaikeizai-no antei-wo daiichi-ni kangae-te

‘Needless to be pointed out by the USA, Japan first considers the stability of domestic economy.’
FUg (degree), [PN1b_00004]

(28) 基本合意の 覚書を 交わした上で、
kihongoui-no oboegaki-wo kawashitaue-de,
合弁会社を 設立
goubengaisha-wo setsuritsu
‘Established a joint corporation upon the exchange of a memorandum of understanding.’
FUh (presuppositional), [PN1g_00002]

In addition, (FUi), as in (29), are means clauses; (FUj), as in (30), are dyadic or binary relation clauses; and (FUk), as in (27), are correlative clauses.

(29) 年末には 上司と 部下が 話し合って
nenmatsu-ni-ha joushi-to buka-ga hanashiatte
次期の 個人目標を つくる
jiki-no kojinmokuhyou-wo tsukuru
‘At the end of the year, the superior and the subordinate have a talk and draft personal goals for the next term.’
FUi (means), [PN5b_00003]

(30) ガスで 作るより スープの 味も まろやか
gasu-de tsukuru-yori suupu-no aji-mo maroyaka
‘The soup has a mellower taste than one made using a stove.’
FUj (dyadic), [PN4a_00001]

(31) でも じっと 眺めているうち、 怖いと
demo jitto nagameteiruuchi, kowai-to
感じ始めた
kanjihajimeta
‘But as he kept staring at it, he started to feel scared.’
FUk (correlative), [PN2b_00002]

(32) 東署で 強盗事件とみて 行方を
higashisho-de goutoujikentomite yukue-wo
追っている
otteiru
‘At the East Police Station, it was regarded as a robbery and tracking was ongoing.’
FUl (conclusive), [PN1f_00002]

Further, (FUm), as in (33), are scenery clauses; (FUn), as in (34), are presuppositional clauses; and (FUo), as in (35), are absolute clauses.

(33) 欧米のような 基盤が ない中、 市民の
oubeinoyouna kiban-ga nainaka shimin-no
実質的な 参加が 得られるように
jisshitsuteki-na sanka-ga erareruyou-ni
‘In order to gain real participation of citizens, while having no foundation like the Western countries,’
FUm (scenery), [PN3g_00001]

(34) 酒を 飲ませない 以外は、 同様の
sake-wo nomasenai igai-ha douyou-no
扱い。
atsukai.
‘The same treatment, except the prohibition to drink alcohol.’
FUn (restrictive), [PN4g_00003]

(35) 無理して 頑張る 必要は ないが、
murishite ganbaru hitsuyou-ha nai-ga,
私は 布団や カーテンなどの大物を
watashi-ha futon-ya kaatennado-no
洗濯するのが 好き
oomono-wo sentakusurunoga suki
‘Although there is no need to push myself too hard, I like washing large objects such as mattresses and curtains.’
FUo (absolute), [PN3b_00004]

Finally, (FUp) covers other adverbial clauses: (FUp100), as in (36), are functional (auxiliary verbal) expressions; (FUp200), as in (37), are idiomatic expressions; and (FUp300), as in (38), are adverbial phrases (not clauses).

(36) 医療の 倫理を 逸脱した 行為と
iryou-no rinri-wo itsudatsu-shita koui-to
いわざるを えないだろう
iwazaru-wo enaidarou
‘It is definitely an act that deviates from medical ethics.’
FUp100 (functional), [PN2b_00002]

(37) なりふり 構わず 資金を
narifurikamawazu shikin-wo
調達しようとした
choutatsushiyoutoshita
‘Tried to raise funds by fair means or foul.’
FUp200 (other, idiomatic), [PN3b_00001]

(38) 50 歳を 過ぎて なぜか エステサロンに 来た
50sai-wo sugi-te naze-ka esutesaron-ni kita
‘Somehow, I came to the esthetic salon after I had passed the age of 50.’
FUp300 (other, adverbial), [PN1b_00003]

2.4. Coordinate Clauses

Coordinate clauses (並列節; Heiretsu Setsu, HS) are classified into resultative (HRa) and contrastive (HRb) at the second level. The resultative clauses are (sub)classified again into exhaustive list (HRa100), as in (39); exemplification (HRa200), as in (40); accumulation (HRa300), as in (41); parallels (HRa400), as in (42), and negation coordination (HRa500), as in (43). An example of a contrastive clause (HRb) is (44).

(39) 年利 3% 借入れ、 三十年の
nenri 3% kariire, 30nen-no
元利均等返済方式で 返済する場合
ganrikintouhensaihoushiki-de hensaisurubaai
‘The circumstance of repaying the debt with an annual interest of 3% through a 30-year level-payment plan.’
HRa100 (exhaustive listing), [PN4c_00002]

(40) カラスよけの 糸を 張り、 ひなの ための
karasuyoke-no ito-wo hari, hina-no tame-no
筒形シェルターを 置くなど、
tsutsugatasherutaa-wo okunado,
恒久的な 営巣地に するため
koukyuuteki-na eisouchi-ni suru-tame

‘In order to make it a permanent nesting place, tie crow-repelling string and set up a cylindrical shelter for chicks.’
HRa200 (exemplification),
[PN5b_00002]

(41) 病気の パターンごとの
byouki-no pataangoto-no
入院日数だけでなく、 医療費の
nyuuinnissuudakedenaku, iryouhi-no
全国平均値も 示された
zenkokuheikinchi-mo shimesareta

‘Shown were not only the number of hospitalization days by the pattern of diseases, but also the national average medical fees.’
HRa300 (accumulation),
[PN3a_00003]

(42) 条約締結国に 国内の
jouyakuteiketukoku-ni kokunai-no
無形文化遺産の 保護や 目録の
mukeibunkaisan-no hogo-ya mokuroku-no
作成を 求めるとともに、
sakuse-wo motomerutotomo-ni,
国際協力のための 基金設置などを
kokusaikyourokunotameno kikinsecchinadowo
盛り込んでいる。
morikondeiru.

‘Incorporated the establishment of funds for international cooperation, along with the requests for State Parties to protect domestic intangible cultural heritage and produce catalogues.’
HRa400 (parallel), [PN4g_00001]

(43) 義務ではなく、 各学校の 判断で
gimudehanaku, kakugakkou-no handan-de
行われる。
okonawareru.

‘It is not an obligation and is carried out at the discretion of each school.’
HRa500 (with negation),
[PN1a_00002]

(44) 昨日は あなたに ほほ笑んだけれど、 今日は
kinou-ha anata-ni hohoenda-keredo, kyou-ha
さようならを 言わなければいけない。
sayounara-wo iwanakerebaikenai.

‘I smiled at you yesterday, but today I have to say goodbye.’
HRb (contrastive), [PN2b_00003]

3. Annotation Procedures

3.1. Overview of Procedures and Target Data

We annotated third-level clause boundary labels for 52 classes on 54 BCCWJ newspaper core data A samples. The sentence boundaries, word segmentation, morphological information, *bunsetsu* (Base phrase), and *bunsetsu*-based syntactic dependency were annotated precedingly. The data consisted of 2,543 sentences and 56,922 morphemes.

The annotation procedure was based on the modification of the automatically extracted clause boundaries. First, clause boundary candidates were extracted using clause patterns with a fourth level of labels. Patterns were defined based on the morphological information and syntactic dependency relations.

3.2. Patterns of Clauses

The original Tori-Bank pattern files were provided through a contract with a data distribution organization; the specification document is available as a PDF file on the website[1]. The original patterns were based on the surface forms of the morphological analyser outputs. We reimplemented the patterns and adapted them for the UniDic POS set, lemma information, and syntactic dependency[2]. The patterns were based on a syntactic dependency structure[3]. Note, the morphological information was manually annotated on the original BCCWJ. The syntactic dependency structure was also annotated on the BCCWJ (Asahara and Matsumoto, 2016).

3.3. Annotation

First, two annotators checked the labeled clause boundary candidates based on the patterns. The first 14 files of the total of 54 files were used in a training phase.

Second, one annotator resolved the inconsistency between the two annotations.

4. Data Statistics

Table 1: Disagreement of Clause Position

	Nom	Adnom	Adv	Coord
Disagreement	102 (20%)	275 (33%)	207 (30%)	47 (24%)
Total (the final)	486	836	701	199

Table 2: Discrepancies of Clause Labels Between Two Annotators

	Nom	Adnom	Adv	Coord
Nominal	8	4	7	8
Adnominal	11	179	4	0
Adverbial	12	6	85	125
Coordinate	1	0	26	2

4.1. Statistics of the Completed Data

The table 5 presents the basic statistics of the completed data in the first and second levels. The 2,543 sentences include 673 Nominal Clauses, 1,103 Adnominal Clauses, 969 Adverbial Clauses, and 293 Coordinate Clauses.

[1] http://unicorn.ike.tottori-u.ac.jp/toribank/data_list.html
[2] https://github.com/X/clause_pattern
[3] https://taku910.github.io/cabocha/

(45) 雪舟作と 伝えられる 花鳥図屏風は、 10 点余りが 知られている。
Sesshusakuto tsutaerareru kachouzubyoubu-ha, 10tenamari-ga shirareteiru.

‘Around 10 folding screens of flower and bird by *Sesshu* were identified.’
MSa100 (relative, restrictive) vs MSa200 (non-relative, restrictive), [PN2b_00002]

(46) 警察当局が 危険人物と 認定した 九百三十二人に対し、
keisatsutoukyoku-ga kikenjinbutsu-to ninteishita 932nin-nitaishi,

‘For 932 people who are regarded as dangerous by the police’
MSa100 (relative, restrictive) vs MSa200 (non-relative, restrictive), [PN2c_00002]

(47) 再建計画に 数値基準を 設けた 中間報告の 中核的な 考えに 反映されている。
saikenkeikaku-ni suuchikijun-wo mouketa chuukanhoukoku-no chuukakuteki-na kangae-ni haneisareteiru.

‘.. are reflected in the core idea of the interim report (in which/that) is set as the numerical criterion for the restructuring plan’
MSa100 (relative) vs MSb (apposition), [PN1g_00002]

(48) 他派閥からも 引き抜いて 三十人から 五十人の 新派閥を つくることが できるんだ
tahabatsu-kara-mo hikinuite 30nin-kara 50nin-no shinhabatsu-wo tsukurukoto-ga dekirunda

‘We can create a new faction with 30–50 people by hiring from the other factions’
FUi (means) vs FUb (causal), [PN2e_00002]

(49) 各政権の 積み残しを 一手に引き受けて、そのすべてを 処理するという . . .
kakuseiken-no tsuminokoshi-wo ittenihikiukete, sonosubete-wo shorisurutoiu . . .

‘(the new government) took charge of the goods left by the previous governments and processed all of them ... ’
FUd (attendant circumstances) vs FUb (causal), [PN1b_00004]

Table 3: Frequent Discrepancies of Clause Labels in Adnominal Clauses

	Annotator A	Annotator B
46	Relative Clause (Restrictive)	Relative Clause (Non-Restrictive)
27	Relative Clause (Non-Restrictive)	Relative Clause (Restrictive)
23	Apposition Clause	Relative Clause (Restrictive)
14	Adnominal Clause (w/ Contractive)	Apposition Clause
14	Adnominal Clause (w/ Contractive)	Relative Clause (Restrictive)
10	Relative Clause (Restrictive)	Apposition Clause

Table 4: Frequent Discrepancies among Clause Labels in Adverbial Clauses

	Annotator A	Annotator B
33	Means	Causal
12	Attendant Circumstances	Causal

The most frequent type of clause in the second level of nominal clauses was the quotation clause (342). The quotation clauses were marked with a ‘と’ (*to*) marker in reported speech. The frequency of noun clauses was 300. The Japanese noun clauses were marked ‘の’(*no*), ‘こと’ (*koto*), and ‘ところ’(*tokoro*).

The adnominal clauses are classified into relative clauses, apposition clauses, and others, including functional or collocational clauses. The major difference between the relative clauses and apposition clauses is whether the predicate in the clause modifier and the modified noun have predicate-argument relations.

The adverbial clauses were semantically classified into 16 classes in the second level. The most frequent type was causal relations (243). The second most frequent type was attendant circumstances (118).

Finally, the coordinate clauses were classified into the following: resultative (282) and contrastive (11) clauses.

4.2. Disagreement in the Annotation Phases

We investigated disagreements between the two annotators in the first annotation phases. We present only disagreements after the training phase, that is, in the files 15-54.

Table 1 shows disagreement on boundary detection. These disagreements were seen most frequently on adnominal clauses. Because of that, Japanese subject nominal phrases tend to be omitted, and any attributive adjective can become an adnominal clause. We introduced clauses composed of more than one *bunsetsu*. However, the judgments of the two annotators tended to disagree. We refined the definition of the clause based on the existence of a complement for the attributive adjective predicate.

Table 2 shows discrepancies in third-level labels on agreed segments.

The most frequent discrepancies were in the second-level labels in the adnominal clauses. Table 3 shows discrepancies in the third-level labels within adnominal clauses.

It is important in English clause classification to distinguish

Table 5: Second Level Labels

Label	Description	Count
HS: Nominal Clause		671
HSa	Noun	300
HSb	Interrogation	29
HSc	Quotation	342
MS: Adnominal Clause		1103
MSa	Relative	677
MSb	Apposition	213
MSc	Other	122
MSd	Functional	66
MSe	Collocational	25
FU: Adverbial Clause		969
FUa	Temporal	76
FUb	Causal	243
FUc	Conditional, Concessive	96
FUd	Attendant Circumstances	118
FUe	Contrastive	98
FUf	Objective	43
FUg	Degree	3
FUh	Presuppositional	8
FUi	Means	94
FUj	Dyadic	18
FUk	Correlative	6
FUl	Conclusive	18
FUm	Scenery	2
FUn	Restrictive	3
FUo	Absolute	68
FUp	Other	75
HR: Coordinate Clause		293
HRa	Resultative	282
HRb	Contrastive	11

between relative clauses that are restrictive and those that are non-restrictive. Whereas restrictive relative clauses of normal nouns, non-restrictive ones modify proper nouns. In contrast, in Japanese grammar the distinction between these two is vague and not overtly marked. Examples (45) and (46) show disagreeing judgments on which relative clauses were restrictive or non-restrictive. For example, the 花鳥図屏風 'folding screens of flower and bird' in (45) and 九百三十二人 '932 people' in (46) are difficult to specify based on world knowledge.

Moreover, the difference between relative clauses and apposition clauses is vague in the Japanese language, because the subject and object of the predicate can be omitted. Example (47) shows disagreeing judgments between relative and apposition clauses: whereas annotator A regards the example as a restrictive relative clause with the subject 中間報告 'the interim report', annotator B regards it as an apposition relative clause with subject ellipsis. The sentence is too vague to resolve the attachment ambiguity.

The second-most frequent discrepancy is between coordinate and adverbial clauses. This is because the coordinate structure is a syntactic meta-structure, in which coordinate clauses are subcategorized into adverbial clauses in a clause boundary definition (Maruyama et al., 2016).

The third-most frequent discrepancies are within adverbial clauses. Table 4 shows frequent discrepancies within adverbial clauses.

The conjunctive postposition て (*te*) form in Japanese has ambiguities for semantic classification. (48) shows the discrepancy between means and causal relations. 引き抜いて 'hiring' can serve as both the means and the cause for 新派閥をつくる 'create a new faction'. Then, (49) shows the discrepancy between attendant circumstances and causal relations.

4.3. Data release

These discrepancies were resolved in the second phase of checking. One annotator resolved annotation ambiguity through introspection. It was found that most disagreements were caused by oversight.

Label disagreement was caused by homographical patterns, such as in suspended form (連用中止 in Japanese) and て (*te*) form. The annotator of the second check defined a standard for differentiating them. For example, it was determined whether the mutual substitution of these types of clauses could preserve the meaning of the original sentence in a language test and thus resolve the ambiguity between coordinate and adverbial clauses. However, there are also truly ambiguous examples, which cannot be resolved even using contextual information. We put some special notes on examples that may have interpretations other than the classes with which they are annotated.

The final annotation data are available for users of the BCCWJ DVD Edition, published by the NINJAL official[4].

5. Conclusions

We present annotation data on Japanese clause boundaries with syntactic and semantic labels. We reimplemented the Tori-Bank clause patterns in the UniDic POS tagset and syntactic dependency structures. Two annotators modified the clause candidates yielded by the pattern-based analysers, and we explored the segments and labels on which they disagreed and resolved the disagreements.

The clause classes in Tori-Bank were originally designed for machine translation from English to Japanese. Some clause classes relate to for English-specific structures or issues. In our future work, we will refine the Tori-Bank clause class standard for the Japanese language. For example, adverbial clauses can be subcategorized into statement clauses and logical clauses.

Acknowledgments

This work was supported by JSPS KAKENHI Grants Numbers JP15K12888 and JP17H00917 and is a project of the Center for Corpus Development, NINJAL.

6. Bibliographical References

Asahara, M. and Matsumoto, Y. (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.

[4] NINJAL provides services to distribute BCCWJ derived data.

Ikehara, S. (2007). Japanese semantic pattern dictionary – compound and complex sentence eds. – . `http://unicorn.ike.tottori-u.ac.jp/toribank/`.

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.

Maruyama, T., Sato, S., and Natsume, K. (2016). Gendai Nihongo ni-okeru Setsu no Bunruitaikei ni-tsuite, On the clause classification in Contempprary Japanese, (in Japanese). In *22nd Annual Meeting of Gengoshori-gakkai*, pages 1113–1116.