UD_Japanese-CEJC: Dependency Relation Annotation on Corpus of Everyday Japanese Conversation

Mai Omura
NINJAL, Japan

Aya Wakasa
Tohoku University

Hiroshi Matsuda
Megagon Labs, Tokyo,
Recruit Co., Ltd.

Masayuki Asahara
NINJAL, Japan

Abstract

In this study, we have developed Universal Dependencies (UD) resources for spoken Japanese in the Corpus of Everyday Japanese Conversation (CEJC). The CEJC is a large corpus of spoken language that encompasses various everyday conversations in Japanese, and includes word delimitation and part-of-speech annotation. We have newly annotated Long Word Unit delimitation and *Bunsetsu* (Japanese phrase)-based dependencies, including *Bunsetsu* boundaries, for CEJC. The UD of Japanese resources was constructed in accordance with hand-maintained conversion rules from the CEJC with two types of word delimitation, part-of-speech tags and *Bunsetsu*-based syntactic dependency relations. Furthermore, we examined various issues pertaining to the construction of UD in the CEJC by comparing it with the written Japanese corpus and evaluating UD parsing accuracy.

1 Introduction

Universal Dependencies (UD) (Nivre et al., 2016; de Marneffe et al., 2021) is a framework for consistent annotation of grammatical elements including parts of speech, morphological features, and syntactic dependencies in various human languages. UD provides a wide range of corpus types, encompassing written as well as spoken language data (Dobrovoljc, 2022).

The UD Japanese team has also developed and maintained several resources (Asahara et al., 2018), including UD_Japanese-GSD, UD_Japanese-PUD (Asahara et al., 2018) and UD_Japanese-BCCWJ (Omura and Asahara, 2018). Additionally, there are distinct versions of these corpora with long-unit word annotations (Omura et al., 2021). However, all of these resources are currently limited to written Japanese. Therefore, the present study addresses this gap by introducing UD resources for spoken Japanese and leveraging the Corpus of Everyday Japanese Conversation (CEJC). The resulting resource is referred to as **UD_Japanese-CEJC**.

The CEJC (Koiso et al., 2022) was recently released by NINJAL, Japan. This corpus represents a significant advancement in spoken language resources, as it comprises a large-scale collection of Japanese conversations encompassing more than 200 hours. Various types of audio and video data - including chat sessions, consultations, and meetings - were collected for the CEJC corpus. The informants were carefully selected to ensure a balanced representation in terms of gender and age. The resource includes transcriptions and word segmentation information along with Japanese part-of-speech tags. In addition, we have newly annotated *Bunsetsu* (Japanese-phrase unit)-based dependencies for a subset of the CEJC dataset, specifically in a 20-hour segment. Building upon this, Omura and Asahara (2018) have proposed conversion rules to transform the *Bunsetsu*-based dependencies into UD trees. By applying the conversion method proposed by Omura and Asahara (2018), it becomes feasible to transform the CEJC corpus into UD corpus, thereby facilitating the development of a substantial Japanese UD spoken corpus.

We present the outcomes of our endeavor in the development of a spoken UD Japanese corpus using the dialogue-based CEJC. An overview of our work is depicted in Figure 1. The CEJC corpus provides audio and video data along with token mappings for dialogues, enabling the realization of UD mappings. In the following sections, we elaborate on the proposed annotation scheme and present essential statistics of the resulting dataset, drawing upon related research. Furthermore, we evaluate the performance of a parser trained on both the UD Japanese written and spoken corpora. We also highlight the distinctive features of UD_Japanese-CEJC in comparison to written and spoken language, with a specific emphasis on

disfluencies such as reparanda, repairs, and fillers characteristic of dialogue-based UD.

2 Related Work

2.1 Spoken Language Treebanks

Since the seminal work on the Switchboard Corpus (Godfrey et al., 1992; Calhoun et al., 2010), a number of spoken language treebanks have been developed (Marcus et al., 1999; Zen et al., 2019; Hovy et al., 2006). These treebanks have played a crucial role in research pertaining to natural spoken language processing, serving as essential resources for the development of applications such as speech recognition, speech synthesis, speech translation, spoken language understanding, and speech-based dialogue systems. However, the construction of spoken language treebanks poses technical and linguistic challenges in terms of data collection, annotation, and analysis, all of which are more complex compared to their counterparts in text-based treebanks.

In this context, the UD framework (Nivre et al., 2016; de Marneffe et al., 2021) for spoken language treebanks has emerged as an important development in the field of natural language processing. The UD provides a dependency structure framework (see right side of Figure 1), data format, and guidelines [1] that emphasize commonality across languages. The representation of dependency trees through a common annotation scheme enables language comparisons and improvements in machine translation and other applications. The UD framework also provides a consistent and cross-linguistically applicable set of syntactic annotations are essential for the development of high-quality language processing tools (Straka, 2018; Honnibal et al., 2020).

Dobrovoljc (2022) composed an overview of UD for several spoken languages. UD treebanks for spoken languages vary in size, with relatively large corpora available for Naija (Caron et al., 2019), Norwegian (Øvrelid et al., 2018), and French (Kahane et al., 2021a) in contrast with lower-resource languages such as Beja (Kahane et al., 2021b), Cantonese (Wong et al., 2017), Chukchi (Tyers and Mishchenkova, 2020), and Frisian (Braggaar and van der Goot, 2021). Analyses of spoken language corpora are also being undertaken, for example Kahane et al. (2021a) analyzed examples of spoken dialogue in the Beja, Naija, and French UD treebanks, and examined language phenomena necessary for research on spoken dialogue such as speaker overlap, fillers, and silent pauses.

Yaari et al. (2022) constructed an English UD treebank of 31,264 transcriptions from Hollywood movies. The corpus is multimodal, as it exhibits alignment between audio and video sources. However, it should be noted that the treebank consists of scripted, rather than spontaneous, speech.

2.2 Japanese Spoken Language Resources

Data collection in the Japanese language started with small-scale data, such as reading speech for dialogue systems and speech recognition (Yuichi and Tomoko, 2018). Spontaneous dialogue data continues to be collected as it is recognized to be crucial. Several Japanese spoken language corpora have been constructed in prior studies; e.g., the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2003), Nagoya University Conversation Corpus (NUCC) (Fujimura et al., 2012), SMOC corpus (Yamazaki et al., 2020). (Koiso et al., 2022) in the Table 1 also compiled a list of Japanese spoken language resources that includes spontaneous dialogue corpora.

Each type of data is associated with different research purposes, formats, and annotations. In particular, there has been no unified syntactic annotation in Japanese, and UD format treebanks of spoken Japanese have not been developed to date. Our study aimes to construct the UD version of CEJC as described in Section 3.

2.3 UD Japanese

The UD Japanese team has built several resources with UD Japanese-KTC (Tanaka et al., 2016) as the point of departure, wherein data are based on their constituent trees (Tanaka and Nagata, 2013). As of v2.5, UD Japanese-BCCWJ offers intuitive suitability for Japanese syntax along with an abundance of existing resources. Consequently, more recent UD Japanese resource have been based on a corpus of *Bunsetsu*-based syntactic dependencies. Bunsetsu is Japanese base phrase unit of syntactic dependencies.

Furthermore, NINJAL negotiated with stakeholders to inherit and manage data continuously for the GSD and PUD corpora. The data were manually annotated according to their UniDic-based morphological information (Den et al.,

[1] https://universaldependencies.org/

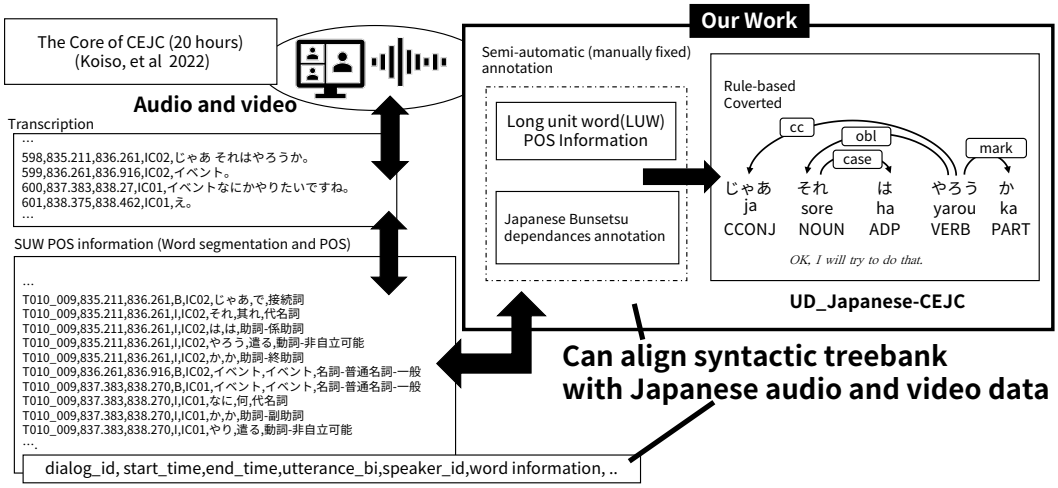

Figure 1: The overview of out building UD_ Japanese-CEJC. (The sample is dialog T010_009 from CEJC)

2008), NINJAL Short Unit Word (SUW) delimitation, NINJAL Long Unit Word (LUW) delimitation, and Bunsetsu (base phrase)-based syntactic dependencies on the original text. The UD Japanese team developed conversion rules from the two-word delimitation and Bunsetsu-based syntactic dependencies to SUW-based UD (Asahara et al., 2018). The Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) is one of large written Japanese corpora. The corpus serves as a model to annotate UD Japanese GSD and PUD with SUW, LUW, and Bunsetsu-based syntactic dependencies (Asahara and Matsumoto, 2016). Likewise, Omura and Asahara (2018) constructed UD Japanese-BCCWJ via the conversion rules of UD Japanese-GSD and PUD.

Thus, the design of UD Japanese is based on SUW, LUW, and Bunsetsu-based dependencies. The CEJC includes the SUW and its morpheme information. If we know the LUW and Bunsetsu-based dependencies of the CEJC, we can develop UD resources for spoken Japanese via the methods described in Omura et al. (2021). In Section 3.2, we describe the construction of the LUW and Bunsetsu-based dependencies of CEJC.

3 Design of UD_Japanese-CEJC

The following section provides a concise overview of CEJC and outlines the construction of the UD version of CEJC.

Sound file ID	yes
Text-sound alignment	yes
Speaker ID	yes
Language variety	no
Standard orthography	yes
Capitalization	not applicable
Pronunciation	yes
Speaker overlap	yes
Final punctuation	not applicable
Other punctuation	not applicable
Incomplete words	yes
Fillers	yes
Silent pauses	yes
Incidents	yes
Text-video alignment	yes
Dialog act	yes (ISO-24617-2)
Intonation label	partially yes

Table 1: Transcription characteristics in CEJC. (cf. Dobrovoljc (2022), Table 2)

3.1 Corpus of Everyday Japanese Conversation

The Corpus of Everyday Japanese Conversation (CEJC) (Koiso et al., 2022) is a large-scale spoken Japanese corpus. It encompasses 200 hours of speech, comprising 577 conversations approximately 2.4 million words and involving a total of 1675 participants. Data are segmented into utterance units based on perceptible pauses and clause boundaries. Transcriptions of the speech audio and video data are provided, and the text is further segmented into word units using SUW and UniDic-based morphological information.

The Core dataset is a subset of CEJC that consists of 20 hours of speech, encompassing 52 di-

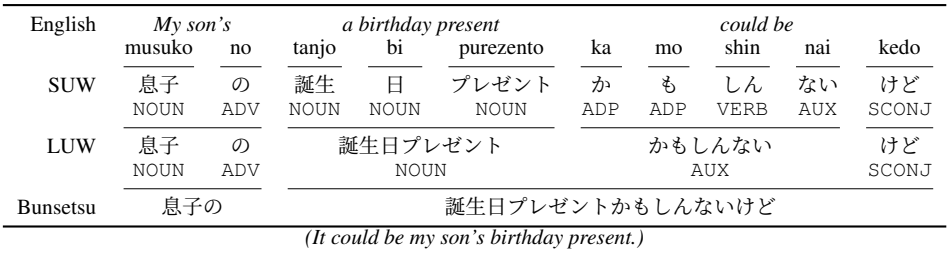

Figure 2: Example of two-way POS annotation (Short and Long unit word) and Bunsetsu of CEJC (refer to T011_005.) The lines above indicate the word boundaries. The parts of speech are represented using universal POS tags for simplicity, but UD_Japanese CEJC can refer to the UniDic part-of-speech tags.

alogues. This subset includes manually annotated and corrected annotations. For this dataset, we annotated LUW and established Bunsetsu-based dependencies. Details pertaining to this annotation process are discussed in the following section.

Table 2 in (Dobrovoljc, 2022) provides an overview of the transcription characteristics in the CEJC. We present a summary of these characteristics in Table 1. The language variety represented in the CEJC is predominantly limited to speakers of common Japanese residing in Tokyo and surrounding prefectures. It is important to note that Japanese does not follow a capitalization convention. Additionally, the transcription rule employed in the CEJC does not account for punctuation marks. One characteristic of the CEJC is the alignment of video data to speech. All videos were collected by normal and omnidirectional 360-degree cameras[2]. The dataset contains dialog act annotations following the ISO 24617-2 scheme (Iseki et al., 2019). Moreover, the audio files are partially annotated with intonation labels using X-JToBI (eXtended-Japanese ToBI), a framework specifically designed for the analysis of spontaneous Japanese speech, as employed in the CSJ corpus (Maekawa, 2003).

3.2 *Bunsetsu*-based Dependency Annotation

The written Japanese data are segmented into sentences based on sentence end symbols specified by authors. However, because sentence-ending punctuation is absent in spoken dialogue, sentence bounds are significantly less straightforward. To address this, the CEJC developers introduced the concept of utterance units, specifically focusing on long utterance units (Den et al., 2010) characterized by silent pauses and clause boundaries. These long utterance units are identified by syntactic and pragmatic disjuncture within the dialogues. Throughout our annotation process, we treated each utterance unit as a separate sentence, forming a tree structure.

We newly annotated the LUW morphological information and *Bunsetsu* boundaries for the CEJC trees. An example of word delimitation using SUW, LUW, and Bunsetsu is illustrated in Figure 2. The SUW is a minimal language unit that has a morphological function and the LUW definition can be regarded as syntactic words in Japanese based Bunsetsu. For further details, please refer to (Omura et al., 2021) and NINJAL website [3]. The LUW information was initially analyzed using Comainu (Kozawa et al., 2014) and subsequently manually corrected by annotators.

In addition, we annotated Bunsetsu-based dependencies for the CEJC utterance units following the BCCWJ-DepPara annotation scheme (Asahara and Matsumoto, 2016). The Bunsetsu-based dependencies was also analyzed by Cabocha (Kudo and Matsumoto, 2002), manually corrected by annotators. It is important to note that the Japanese language exhibits a strict head-final order within the Bunsetsu units. However, the Bunsetsu dependencies in CEJC encompass linguistic phenomena such as fillers, anastrophes, and predicate ellipses, which are rarely observed in written texts. In cases where a dependent does not have its corresponding head within the utterance units, we position a dummy node as the dependency head at the end of the utterance, as depicted in Figure 3.

3.3 Conversion into UD schema

The UD_Japanese-CEJC corpus was derived from the Bunsetsu dependencies in the core data sub-

[2]Video files include the faces of the main conversation participants who agreed to have their faces published. All other participant's faces are obscured.

[3]https://clrd.ninjal.ac.jp/bccwj/en/morphology.html

Conversion rule	UPOS
...	...
POS of SUW is *punctiction*	PUNCT
...	...
POS of SUW is *adjective*	ADJ
POS of SUW is *noun*	NOUN
...	...
POS of SUW is *verb* & The Bunsetsu is the end of the phrase	VERB
...	...

Conversion rule	DEPREL
Bunsetsu is the end of the phrase & Subject word	*root*
UPOS is PUNCT	*punct*
...	...
Subject word in the Bunsetsu & UPOS is NOUN & Attaching particle 'ga'	*nsubj*
...	...
Bunsetsu is not functional phrase & UPOS is ADJ	*amod*
UPOS is ADP	*case*
...	...

Table 2: The short sample of UD conversion rules is outlined in (Omura and Asahara, 2018). As of July 2023, there are 85 rules for UPOS conversion and 120 rules for DEPREL.

	UPOS	DEPREL
If *the word* is filter	INTJ	*discourse(:filter)*
If *the word* is disfluency	X	*reparandum*

Table 3: Labeling rules to convert for UD_Japanese-CEJC. The current approach for determining whether the word is filler or disfluencyis to reference the POS information.

set, which consists of 20 hours of transcribed speech. To compile the UD Japanese resource, we applied the conversion rules outlined in (Omura and Asahara, 2018), which are shared across all UD Japanese treebanks, including GSD, PUD, BCCWJ, GSDLUW, PUDLUW, and BCCWJLUW (Omura and Asahara, 2018; Omura et al., 2021) [4]. Table 2 shows a partial set of conversion rules. These rules determine the UPOS (Universal Part-of-Speech) and DEPREL (Dependency Relation Label) in the UD framework. However, it is important to note that the conversion rules primarily consider written Japanese corpora and might not fully capture the specific characteristics of spoken Japanese. As a result, additional rules were introduced to handle fillers and stutters, which are infrequent in written corpora, as shown in Table 3. While these conversion rules provide a valuable starting point, further refinements may be necessary to fully account for the nuances of spoken Japanese.

[4]There are several spoken UD corpora that offer automatic conversion of existing resources; e.g., UD French ParisStories (Kahane et al., 2021a) and Naija NSC (Caron et al., 2019)

In the UD version of the CEJC, the aforementioned utterance units serve as boundaries for dependency trees. According to the UD guideline, other treebanks have their own language-specific guidelines for handling fillers and disfluencies (e.g. Slovenian SST (Dobrovoljc and Nivre, 2016)). Nevertheless, we decided that any fillers and disfluencies dependent on the dummy node are to be converted to the sentence end root to adhere to the single root restriction, as their attachment is inherently ambiguous. Because argument ellipses are common in Japanese and the annotation units in this dataset are based on utterances, we can only define these ellipses as fillers or disfluencies within the scope of the utterance unit. To determine the appropriate attachment of fillers across languages, including those where ellipses are grammatically allowed, a thorough investigation is necessary.

Figure 3 shows an example of Bunsetsu dependencies constructed to the UD framework. The Bunsetsu-dependency structure is converted to UD structures according to rules specified in (Omura and Asahara, 2018). In the case of the figure, the words "tsu" and "n" are a disfluency and filter, respectively, making them dependent upon the root node " deki ta shi".

3.4 Statistics of UD Japanese CEJC

Table 4 presents a statistical analyses of the generated UD_Japanese-CEJC (spoken) corpus in comparison to UD_Japanese-GSD and BCCWJ (written). These statistical values are from version 2.11. The 'Trees' column indicates the numbers of utterance units in CEJC (spoken) and sentences in GSD (written). The 'Tokens' column represents the total count of word tokens in each treebank. The 'Avg.' column displays the average number of word tokens per tree, whereas the 'Bunsetsu' column indicates the total number of Bunsetsu. The automatic conversion of the 20-hour speech transcription has yielded a substantial amount of data that aligns with the corresponding audio and video. However, it is worth noting that the number of words in a dependency tree within a spoken utterance unit tends to be smaller than that in a written sentence. It provides a clear comparison between the statistics in Table 1 of Dobrovoljc (2022) and the specific characteristics of CEJC as a conversational corpus including many phatic expressions like *Aizuchi* in Japanese such as "hai"

Corpora	Unit	Trees	Tokens	Avg. per Tree	Bunsetsu
CEJC	SUW	59,319	256,885	4.3	136,071
	LUW	59,319	231,774	3.9	136,071
GSD	SUW	8,100	193,654	23.9	65,966
	LUW	8,100	150,243	18.5	65,966
BCCWJ	SUW	57,109	1,253,903	21.9	425,751
	LUW	425,751	99,5632	17.4	425,751
CEJC-	SUW	54,599	24,4296	4.7	124,456
	LUW	54,599	219,415	4.0	124,456

Table 4: Statistics of UD Japanese CEJC (spoken), GSD, and BCCWJ (written) (v2.11). CEJC- is a CEJC corpus that omits any words containing solely inapplicable morphological information (non-lexical tokens), filters, or reparandums.

and "ee" ("uhhuh" and "yeah" in English).

Table 5 shows the distribution of UPOS labels of UD_Japanese-CEJC, GSD, and BCCWJ [5]. The spoken data does not include any PUNCT and SYM, as punctuations and symbols were not accounted for. CCONJ and INTJ are larger than the written corpora. Whereas the written data tend to omit PRON, the spoken data tends to include PRON when referencing speakers. X is a token associated with no morphological annotations, such as incidents (laugh, cry, singing, etc.) in the CEJC.

Table 6 shows the distribution of DEPREL labels of UD Japanese CEJC, GSD and BCCWJ. In the spoken data, words are shorter per a tree (see Table 4). Consequently, the DEPREL *root* is the largest element within the spoken data. Because PUNCT does not appear in the spoken data, the DEPREL *punct* is zero.

	CEJC	GSD	BCCWJ
ADJ	3.69%	1.98%	2.14%
ADP	13.61%	21.62%	20.03%
ADV	6.74%	1.22%	1.51%
AUX	13.24%	10.93%	9.74%
CCONJ	1.64%	0.42%	0.41%
DET	0.56%	0.51%	0.48%
INTJ	10.74%	0.01%	0.07%
NOUN	14.86%	30.05%	29.24%
NUM	1.67%	2.67%	3.11%
PART	8.49%	0.65%	1.18%
PRON	3.77%	0.57%	0.90%
PROPN	1.39%	3.69%	2.87%
PUNCT	0.00%	9.93%	11.69%
SCONJ	6.68%	4.13%	4.49%
SYM	0.00%	0.67%	1.53%
VERB	9.86%	10.96%	10.57%
X	3.05%	0.00%	0.03%

Table 5: The distribution of UPOS labels in UD_Japanese-CEJC, GSD and BCCWJ (SUW)

4 Parser Evaluation

We conducted experiments to assess the reproducibility and parsability of the CEJC corpus. Through a comparison between CEJC and GSD, we illustrate the distinctions between spoken and written Japanese in terms of UD annotation.

4.1 Corpus

To evaluate parsing, we used the following UD Japanese v.2.11 [6] corpora: GSD, CEJC, and their combination (CEJC+GSD). Although SUW and LUW UDs are present, we only considered SUW to examine differences between the spoken and written corpora. The GSD was split among train, dev, and test sets by original UD corpus. The UD CEJC was divided between training, development, and testing sets according to a 8:1:1 ratio based on conversation form as provided by the CEJC: chat, consultations, and meetings. Table 7 shows the distribution of UDs in the experiment. The models were constructed with the sentence (tree) boundary as given, as it is easy to imagine that the utterance units and written sentences are clearly different in Table 4. In particular, CEJC explicitly lacks

[5]Because the SUW are encapsulated in the LUW, there is no significant difference in distribution. Therefore, only SUW are listed.

[6]These UD Japanese is also in development as of November 2022. This version conforms to the latest UD guidelines.

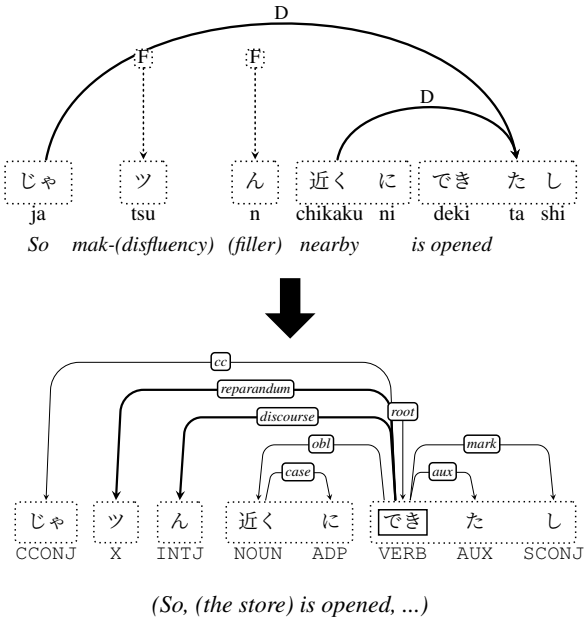

Figure 3: Sample construction of UD_Japanese-CEJC（T011_007）. The upper figure represents Bunsetsu-dependencies and the lower figure shows the UD conversion. The dotted box denotes the Bunsetsu boundary, and the Bunsetsu dependency edge label 'D' is an ordinal dependency relation, where 'F' indicates that no relation is present.

punctuation, making it difficult to identify speech breaks.

4.2 Parser Model

We used spaCy v3.4 (Honnibal et al., 2020), along with spacy-transformers v1.2 as a parsing model framework. spaCy is a trainable network that features a component pipeline for sentence analysis and word tokenisation, part-of-speech tags, dependencies, and named entities. Furthermore, spaCy can use pre-trained transformers (Wolf et al., 2020) such as BERT (Devlin et al., 2019), and allows loss gradients to be shared between the transformers-based pre-training model and analysis component.

A significant distinction between CEJC and other written word treebanks lies in the presence of specific word characteristics, particularly fillers and reparanda. To address this feature, we propose two models: *the two-stage analysis model* and *the simultaneous analysis model*. We assessed the effectiveness of these models in accurately capturing the relationship between fillers and reparanda.

The two-stage analysis model comprises two models: a component that detects and removes the

	CEJC	GSD	BCCWJ
acl	2.11%	3.61%	3.62%
advcl	3.87%	3.72%	3.85%
advmod	4.73%	1.18%	1.43%
amod	0.10%	0.23%	0.25%
appos	0.00%	0.00%	0.00%
aux	9.10%	8.90%	7.56%
case	12.72%	21.33%	19.65%
cc	1.59%	0.42%	0.41%
ccomp	0.34%	0.20%	0.22%
compound	3.97%	14.19%	14.67%
cop	1.98%	1.26%	1.20%
csubj	0.09%	0.08%	0.11%
csubj:outer	0.00%	0.00%	0.00%
dep	1.00%	0.04%	0.99%
det	0.54%	0.51%	0.48%
discourse	2.72%	0.01%	0.03%
dislocated	0.00%	0.00%	0.00%
fixed	4.15%	4.45%	4.26%
mark	14.20%	4.06%	5.04%
nmod	2.87%	6.70%	6.92%
nsubj	2.51%	4.02%	3.69%
nsubj:outer	0.00%	0.23%	0.18%
nummod	0.98%	1.45%	1.16%
obj	0.48%	2.74%	2.62%
obl	5.64%	6.55%	5.41%
punct	0.00%	9.93%	11.69%
reparandum	1.21%	0.00%	0.00%
root	23.09%	4.18%	4.55%

Table 6: Distributions of DEPREL labels in UD_Japanese-CEJC, GSD and BCCWJ (SUW)

	train		dev		test	
	trees	tokens	trees	tokens	trees	tokens
GSD	7,050	168,333	507	12,287	543	13,034
CEJC	36,997	157,227	9,837	43,378	12,485	56280
CEJC-	34,105	149,614	9,057	41,055	11,437	53,627

Table 7: The train/dev/test distribution of UD corpus (GSD/CEJC)

span fillers and reparanda, and a component that subsequently analyzes the parsing tree. Following the method described in (Asahara and Matsumoto, 2003) in regards to the spans of fillers and reparanda detecting named entities, the model was trained via spaCy, whereas the other model was trained by eliminating fillers and reparanda (CEJC-). While the model has two components, the accuracy of parsing results is only evaluated using the correct trees in the absence of fillers and reparanda (CEJC-) as seen in (Table 8), as it is difficult to map removed words as fillers and reparanda and others as original text data.

The simultaneous analysis model includes fillers and reparanda simultaneously. SpaCy can share a transformer's information among multiple analytical components and perform simultaneous

learning. The pipeline components of spaCy were organized in the order of transformers, morphologizer analysis, parser analysis, and NER analysis. The ner analysis is used to detect fillers and reparanda equivalently to the two-stage analysis model.

As a transformer pre-trained model on spaCy, we used cl-tohoku/bert-japanese [7], a BERT model trained on the Japanese version of Wikipedia with words tokenized by MeCab (Kudo et al., 2004) and split into subwords by the WordPiece algorithm. The parser component of spaCy is based on the Non-Monotonic Arc-Eager Transition System with extensions to Projectivization/Deprojectivization by Lifting of Nivre (Nivre and Nilsson, 2005) to handle intersecting contexts.

4.3 Parsing Results

Table 8 presents the tokenisation, tagging, lemmatisation, and dependency parsing results obtained by the two spaCy models. **Tokens**, **UPOS**, **XPOS**, and **Lemma** are reproducible and expressed by their F_1 scores. **UAS** (Unlabeled Attachment Score) and **LAS** (Labelled Attachment Score) are standard evaluation metrics in dependency parsing results. These results were output by the evaluation scripts of CoNLL 2018 shared tasks (Zeman et al., 2018).

When the training and testing data are different (e.g. train/dev GSD and test CEJC, or train/dev CEJC and test GSD), tokenisation (**Tokens**) and POS tagging (**UPOS** and **XPOS**) exhibit poor performance. This is because there are differences in vocabulary and distributions of POS and DEPREL. During tokenisation, spoken utterances have significantly different delimiters compared to those observed in written sentences, as the former include fillers, disfluencies, and repairs. It is also difficult to tokenize without spaces, as required by Japanese. POS tagging presents similar challenges. Although the major POSs of the CEJC are `INTJ`, `CCONJ`, and `PRON` (e.g. first personal pronoun, second personal pronoun), the POS `INTJ` is very rare in GSD. Consequently, the assignment of `INTJ` requires training data from the CEJC. Overall, the combined training data (train/dev: CEJC+GSD) achieved the best performance for both GSD and CEJC tokenisation and tagging.

[7] https://github.com/cl-tohoku/bert-japanese/

Results of filter and reparandum detection are shown in Table 9. The simultaneous analysis model tended to be slightly more accurate than the two-step analysis model. This is thought to be an effect of learning-dependent structure analysis, as well as the simultaneous identification of fillers and reparanda. However, compared to the overall evaluation (in Table 8), the accuracy of tokenisation, POS tagging, and dependency analysis for both fillers and reparanda decreased by more than 6 points.

The dependency attachment (**UAS** and **LAS**) of the CEJC is also difficult, and even the CEJC tree length (avg. 4.3) is shorter than that of the GSD tree (avg. 23.9). GSD also encompasses punctuation in written texts, which helps determine the roots of trees and resolve long-distance dependencies. In contrast, the CEJC does not include punctuation in the transcription, making it difficult to determine the roots of trees as well as presenting challenges with respect to fillers and disfluencies.

5 Conclusions

This study introduces a novel UD Japanese resource derived from the Corpus of Everyday Japanese Conversation (CEJC), representing the first spoken language resource in the UD Japanese framework. The UD resource was built upon transcriptions of audio files from individual speakers, accompanied by two types of video recordings (standard camera and omnidirectional 360-degree camera). Whereas previous efforts have been limited in their incorporation of text-to-video alignment, this study presents a substantial treebank with video, surpassing existing UD resources in this aspect. In the future, we plan to primarily expand the annotation based on audio information; e.g., overlap markers similar to those used in the UD French Rhapsodie (Kahane et al., 2021a).

Parser evaluations were conducted to compare the performance of the parser on UD_Japanese-CEJC (spoken) and GSD (written) datasets. The findings clearly demonstrate the challenges associated with parsing spoken Japanese using a model trained on written corpora. The presence of fillers, disfluencies, and repairs significantly impacted tokenisation and POS tagging accuracy, highlighting the unique characteristics of spoken language that must be accounted for to improve parsing performance.

The UD version of CEJC is currently available

train/dev	test	**Token**	**UPOS**	**XPOS**	**Lemmas**	**UAS**	**LAS**
spaCy two-stage analysis model (eliminating gold fillers and reparandums)							
CEJC–	GSD	98.15%	84.54%	96.96%	94.38%	80.58%	71.97%
CEJC–	CEJC–	96.38%	94.45%	92.33%	86.33%	89.71%	87.54%
spaCy simultaneous analysis model (including fillers and reparandums)							
GSD	GSD	98.14%	97.04%	96.96%	94.38%	91.72%	90.84%
GSD	CEJC	81.16%	84.33%	89.32%	84.92%	80.74%	74.71%
CEJC	GSD	98.14%	84.31%	96.96%	94.38%	79.58%	70.52%
CEJC	CEJC	95.44%	93.39%	89.32%	84.92%	88.19%	84.51%
CEJC+GSD	GSD	98.14%	97.16%	96.96%	95.64%	91.49%	90.56%
CEJC+GSD	CEJC	95.55 %	93.47%	93.47%	89.32%	88.38%	86.57%

Table 8: Results of tokenisation, tagging, lemmatisation and dependency parsing using CEJC and GSD.

Category	Occurence train / dev / test	**Two-stage analysis model Token P / R / F**	**Simultaneous analysis model Token P / R / F**	**Simultaneous analysis model UPOS / UAS / LAS**
Filler	1,736 / 524 / 559	88.6% / 87.3% / 87.9%	86.9% / 90.4% / 88.6%	87.7% / 82.4% / 82.0%
Reparandum	2,122 / 741 / 793	90.5% / 86.0% / 88.2%	88.4% / 87.4% / 87.9%	87.9% / 83.7% / 83.2%

Table 9: Results of accuracy detection for fillers and reparanda analyses.

to CEJC subscribers through the dedicated download site on the CEJC platform. Additionally, the UD will be made available on the Universal Dependencies site and the UD Japanese Github repository [8][9] in a standoff format. This wider distribution will enable researchers to access and utilize the UD Japanese CEJC data for various linguistic analyses and applications. The spaCy models employed in the conducted experiments will be made publicly available in Github repository [10]. These resources will allow researchers and practitioners to utilize the models for their own natural language processing tasks and further contribute to the advancement of linguistic research in the field of Japanese spoken language processing.

Acknowledgements

This work is supported by JSPS KAKENHI 19K13195, a collaborative research project with Recruit Co. Ltd., and a NINJAL Collaborative Research Project 'Evidence-based Computational Psycholinguistics Using Annotation Data'.

[8] https://github.com/udjapanese/UD_Japanese-CEJCSUW

[9] https://github.com/udjapanese/UD_Japanese-CEJCLUW

[10] https://github.com/megagonlabs/UD_Japanese-GSD/releases/tag/nlp2023

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. European Language Resources Association.

Masayuki Asahara and Yuji Matsumoto. 2003. Filler and disfluency identification based on morphological analysis and chunking. In *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003)*, pages 163–166, Tokyo, Japan. ISCA.

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 Language Resources*, pages 49–58, Osaka, Japan. The COLING 2016 Organizing Committee.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Sasha Calhoun, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman, and David Beaver. 2010. The NXT-format Switchboard Corpus: a rich resource for investigating the syntax, se-

mantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419.

Bernard Caron, Marine Courtin, Kim Gerdes, and Sylvain Kahane. 2019. A surface-syntactic UD treebank for Naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24, Paris, France. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 2010. Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 2103–2110, Valletta, Malta.

Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. 2008. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 1019–1024, Marrakech, Morocco. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaja Dobrovoljc. 2022. Spoken Language Treebanks in Universal Dependencies: an Overview. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1798–1806, Marseille, France. European Language Resources Association.

Kaja Dobrovoljc and Joakim Nivre. 2016. The Universal Dependencies Treebank of Spoken Slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1566–1573, Portorož, Slovenia. European Language Resources Association (ELRA).

Itsuko Fujimura, Shoju Chiba, and Mieko Ohso. 2012. Lexical and grammatical features of spoken and written Japanese in contrast: exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference : Speech and Corpora*, pages 393–398, Belo Horizonte, Brazil. Firenze University Press.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *Proceedings of 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*, volume 1, pages 517–520 vol.1.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yuriko Iseki, Keisuke Kadota, and Yasuharu Den. 2019. Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *Proceedings of 2019 22nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques*, pages 1–6, Cebu, Philippines. IEEE.

Sylvain Kahane, Bernard Caron, Emmett Strickland, and Kim Gerdes. 2021a. Annotation guidelines of UD and SUD treebanks for spoken corpora: A proposal. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories*, pages 35–47, Sofia, Bulgaria. Association for Computational Linguistics.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021b. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Hanae Koiso, Haruka Amatani, Yasuharu Den, Yuriko Iseki, Yuichi Ishimoto, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka, Yasuyuki Usuda, and Yuka Watanabe. 2022. Design and evaluation of the Corpus of Everyday Japanese Conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5587–5594, Marseille, France. European Language Resources Association.

Shunsuke Kozawa, Kiyotaka Uchimoto, and Yasuharu Den. 2014. Adaptation of Long-Unit-Word analysis system to different part-of-speech tagset [in Japanese]. *Journal of Natural Language Processing*, 21(2):379–401.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, pages 1—7. Association for Computational Linguistics.

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*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Kikuo Maekawa. 2003. Corpus of Spontaneous Japanese : Its design and evaluation. In *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, pages 7–12, Tokyo, Japan. ISCA.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguti, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language resources and evaluation*, 48(2):345–371.

Mitchell P Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. *Linguistic Data Consortium, Philadelphia*, 14.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan. Association for Computational Linguistics.

Mai Omura and Masayuki Asahara. 2018. UD-Japanese BCCWJ: Universal Dependencies annotation for the Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 117–125, Brussels, Belgium. Association for Computational Linguistics.

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2021. Word delimitation issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies*, pages 142–150, Sofia, Bulgaria. Association for Computational Linguistics.

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg, and Janne Bondi Johannessen. 2018. The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 1651–1658, Portorož, Slovenia. European Language Resources Association.

Takaaki Tanaka and Masaaki Nagata. 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 Languages*, pages 108–118, Seattle, Washington, USA. Association for Computational Linguistics.

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online). Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Tak-sum Wong, Kim Gerdes, Herman Leung, and John Lee. 2017. Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 266–275, Pisa,Italy. Linköping University Electronic Press.

Adam Yaari, Jan DeWitt, Henry Hu, Bennett Stankovits, Sue Felshin, Yevgeni Berzak, Helena Aparicio, Boris Katz, Ignacio Cases, and Andrei Barbu. 2022. The Aligned Multimodal Movie Treebank: An audio, video, dependency-parse treebank. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9531–9539, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoshihiro Yamazaki, Yuya Chiba, Takashi Nose, and Akinori Ito. 2020. Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness. In *Proceedings*

of the Twelfth Language Resources and Evaluation Conference, pages 443–448, Marseille, France. European Language Resources Association.

Ishimoto Yuichi and Ohsuga Tomoko. 2018. Spontaneous speech resources in Japan. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC 2018) Special Speech Sessions*, pages 1–5.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proceedings of 20th Annual Conference of the International Speech Communication Association (Interspeech 2019)*, pages 1526–1530, Graz, Austria. ISCA.