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

Paper URL

Mai Omura, NINJAL, Japan
Aya Wakasa, Tohoku University
Hiroshi Matsuda, Megagon Labs, Tokyo, Recruit Co., Ltd
Masayuki Asahara, NINJAL, Japan

NINJAL National Institute for Japanese Language and Linguistics

Sigdial Poster Session 2

Overview & Contributions

Please Check UD Japanese-CEJC !

SUW LUW

Introduction of Universal Dependencies (UD) for Corpus of Everyday Japanese Conversation

- Construction of UD Japanese for CEJC
- Comparisons with other corpora or written language corpora based on features, statistics, and parsing

Offers Universal Dependencies Annotation for CEJC

Contributes to the advancement of linguistic research in the field of Japanese spoken language processing

Design of UD Japanese-CEJC (UD CEJC)

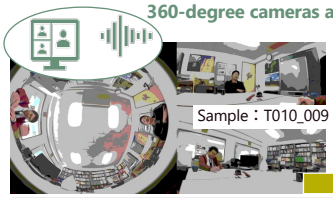
Corpus of Everyday Japanese Conversation (CEJC)

360-degree cameras and audio

[Koiso et al., 2022]

A conversational corpus comprising 200 hours data

- Real-life activities occurring naturally in everyday settings.
- Provides audio, text transcripts, and video data.
- Encompasses a wide range of annotations.




```
…
47,55.191,57.95,IC01,(F あの)(0.107)先生の: (F あの) 日程基準がないんです (T (U よ))。
48,58.067,58.609,IC02,あー。
…
```

Transcription (utterance)

```
…
T010_009,55.191,57.950,I,IC01,日程,日程,名詞-普通名詞-一般
T010_009,55.191,57.950,I,IC01,基準,基準,名詞-普通名詞-一般
T010_009,55.191,57.950,I,IC01,が,が,助詞-格助詞
T010_009,55.191,57.950,I,IC01,ない,無い,形容詞-非自立可能
T010_009,55.191,57.950,I,IC01,ん,の,助詞-準体助詞
T010_009,55.191,57.950,I,IC01,です,です,助動詞
T010_009,55.191,57.950,I,IC01,よ,よ,助詞-終助詞
T010_009,58.067,58.609,B,IC02,あー,ああ,感動詞-一般
T010_009,58.845,60.310,B,IC01,たぶん,多分,副詞
T010_009,58.845,60.310,I,IC01,四,四,名詞-数詞
T010_009,58.845,60.310,I,IC01,年,年,名詞-普通名詞-助数詞可能
T010_009,58.845,60.310,I,IC01,生,生,接尾辞-名詞的-一般
…
```

dialog_id,start_time,end_time,utterance_bi,speaker_id,word information, ..

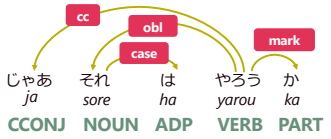
Short unit word (SUW) POS information (Word segmentation and POS)

Conversion of Universal Dependencies of CEJC

Semi-automatic manually fixed annotation

CEJC + Long unit word (LUW) POS Information + Japanese **Bunsetsu Dependances** Annotation

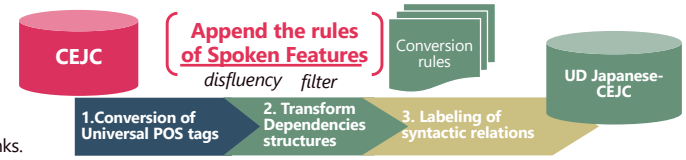
Rule-based Conversion



OK, I will try to do that.

Enabling Multimodal Annotation via UD and Spoken Corpus Collaboration

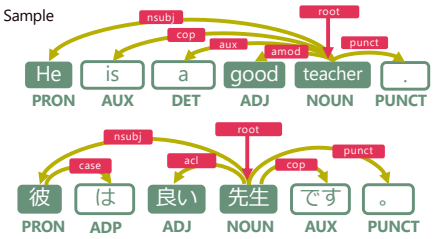
Based on Omura et al., 2018



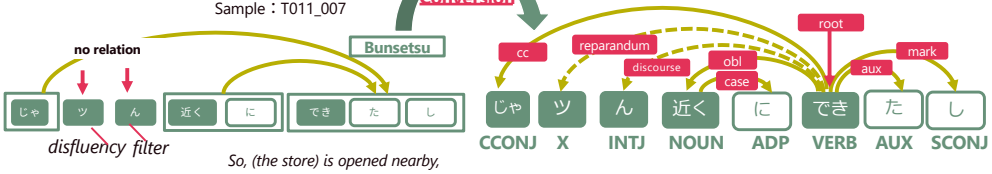
Universal Dependencies (UD)

- A project and standardization framework that constructs cross-lingual dependency treebanks.
- Almost 200 treebanks constructed in over 100 languages.
- UD represents the syntactic relationships between words within a sentence.

Sample



Sample：T011_007



Uniform part-of-speech labels UPOS
Uniform dependencies labels DEPREL

- **The existing UD Japanese corpus [Omura et al., 2021; Omura et al., 2018; Asahara et al., 2018]**
  - **UD_Japanese-BCCWJ(LUW)**
  - **UD_Japanese-GSD(LUW)**
  - **UD_Japanese-PUC(LUW)**

All of them are written UD

UD CEJC Features

Sound file ID	Yes
Text-sound alignment	Yes
Speaker ID	Yes
Language variety	No (only common Japanese)
Standard orthography	Yes
Capitalization	Not applicable
Speaker overlap	Yes
Final punctuation	Not applicable
Other punctuation	Not applicable
Incomplete words	Yes
Fillers	Yes
Silent pauses	Yes
Incidents	Yes

Highlights of UD CEJC

Text-video alignment	Yes
Dialog act	Yes (ISO-24617-2) **[Iseki et al., 2019]**
Incidents	Partial

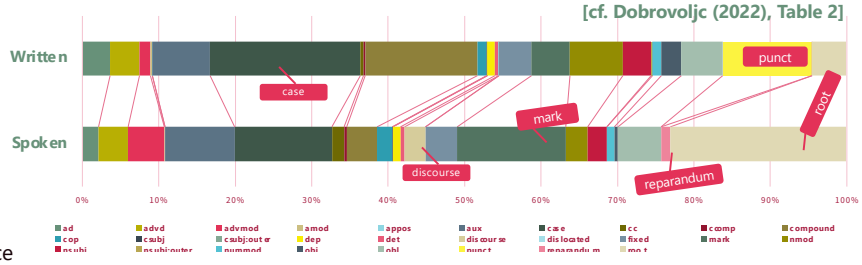
Comparative Analysis of UD CEJC with Other Corpora

Statistics of UD Japanese CEJC

Corpus	Unit	Sents	Words	Avg	Bunsetsu
CEJC	SUW	59,319	256,885	4.3	136,071
	LUW	59,319	231,774	3.9	136,071
GSD (Written)	SUW	8,100	193,654	23.9	65,966
	LUW	8,100	150,243	18.5	65,966
BCCWJ (Written)	SUW	57,109	1,253,903	21.9	425,751
	LUW	57,109	995,632	17.4	425,751

Sentence v.s. utterance

[cf. Dobrovoljc (2022), Table 2]



Distributions of DEPREL Labels

Comparative Experimental Results with Syntactic Analysis

Parsing Models

- **Two-stage analysis model**

  spaCy v3 Ver. 3.4

  Detects and removes span fillers and reparanda. Subsequently analyses the parsing tree.
- **Simultaneous analysis model**

  Simultaneous inclusion of filters and reparanda.

Model URL

Results

- Best performance achieved for both GSD and CEJC. Challenges exist in analyzing fillers and disfluencies.
- The dependency attachment in CEJC is challenging.

FW	Train/Dev	Test	Token	UPOS	XPOS	UAS	LAS
Two-stage analysis model	CEJC–	GSD	98.15%	84.54%	96.96%	**80.58%**	**71.97%**
	CEJC–	CEJC–	96.38%	94.45%	92.33%	**89.71%**	**87.54%**
Simultaneous analysis model	GSD	GSD	98.15%	97.05%	96.96%	91.75%	90.87%
	GSD	CEJC	95.40%	78.53%	89.32%	80.94%	75.03%
	CEJC	GSD	98.15%	84.33%	96.96%	79.61%	70.54%
	CEJC	CEJC	95.40%	93.38%	89.32%	88.27%	86.33%
	CEJC+GSD	GSD	98.15%	97.17%	96.96%	**91.52%**	**90.59%**
	CEJC+GSD	CEJC	95.40%	93.46%	93.47%	**88.45%**	**86.68%**