UD Japanese-CEJC とその評価

大村 舞
国立国語研究所

若狭 絢
国立国語研究所

松田 寛
Megagon Labs

浅原 正幸
国立国語研究所

概要

Universal Dependencies に基づく言語資源の構築が各言語で進められている。日本語においても Universal Dependencies に準拠する言語資源が構築されてきたが、すべて書き言葉に基づくものであった。本研究では、日本語 Universal Dependencies の新しい言語資源として、『日本語日常会話コーパス』に基づく UD Japanese-CEJC を構築したので報告する。既存の書き言葉の日本語 Universal Dependencies の言語資源と同様に、国語研短単位形態論情報・国語研長単位形態論情報・文節係り受けに基づく変換規則によりデータの構築を行った。さらに、さまざまな条件により解析器を構築し、評価を行ったので報告する。

1 はじめに

Universal Dependencies（UD） [1] は言語横断的な依存構造木に基づくツリーバンクを構築するプロジェクトであり、2023 年 1 月現在 100 以上の言語において 200 近くのツリーバンクが構築されている。その研究対象は、書き言葉から話し言葉にシフトしており、Dobrovoljc ら [2] により最近の研究動向がまとめられている。また、日本語 UD チームが、ツリーバンクやパーサの構築の興味のある研究者で組織され、過去にさまざまな日本語 UD リソース [3] の構築を進めてきた。しかし、これまで構築してきた日本語 UD リソースは書き言葉を対象としており、話し言葉を対象とした日本語リソースはなかった。

2022 年に国立国語研究所によって「日本語日常会話コーパス」（Corpus of Everyday Japanese Conversation: CEJC） [4] が構築された。CEJC は、性別・年齢などに基づいた均衡性を考慮された協力者に収録を依頼し、収録されたコーパスであり、転記・短単位形態論情報・発話単位などが付与されている。また、同データに対して、長単位形態論情報・文節係り受けの情報 [5] も付与されている。これらのアノテーション情報が整備されたことにより、Omura ら [6, 7] が提案する文節係り受けから UD への変換規則を用いることで、CEJC から国語研短単位・長単位それぞれの日本語 UD リソースを構築することが可能となった。

本研究では、話し言葉向けに若干改変した変換規則を CEJC に適用し、日本語話し言葉に基づく UD リソース **UD Japanese-CEJC** を新たに構築した。A.1 節の**表 1** に、Dobrovoljc ら [2] がまとめた話し言葉 UD リソースのさまざまな観点項目に沿った UD Japanese-CEJC の特徴をまとめている。音声データのみならず、2 種類のカメラ（通常のカメラと 360 度カメラ）による映像データともアラインメントをとった UD リソースは世界的にも類を見ない。UD Japanese-CEJC は、現時点で映像を含む世界最大規模の話し言葉ツリーバンクとなるだろう。

本稿では、2023 年 5 月に公開予定の UD Japanese-CEJC の構築とともに、さまざまな UD パーサについて評価したので報告する。具体的には、データ変換における話し言葉に特有のフィラーや言いよどみの扱いや、書き言葉-話し言葉間のパーサの分野適応について報告する。

2 UD Japanese-CEJC の構築

2.1 『日本語日常会話コーパス』と文節係り受けアノテーション

『日本語日常会話コーパス』（CEJC） [4] は、200 時間、 577 の（2 人以上の）会話、1675 発話者による大規模な話し言葉のコーパスである。データは、節末やポーズに基づいて「発話単位」 [8] と呼ばれる単位に分割されている。さらに音声・映像データを転記したうえで、転記データに対し、国語研短単位形態論情報が付与されている。CEJC データの短単位に基づく形態素数は約 240 万形態素である。

このうち 20 時間分のデータはコアと呼ばれ、さまざまなアノテーションが人手で付与されている。

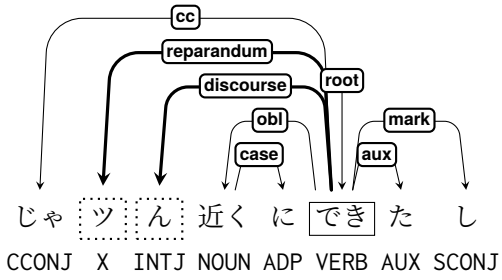

図 1 UD Japanese-CEJC の発話（T011_007）に対する UD の例。「ツ」と「ん」が言いよどみとフィラーである

コアには国語研長単位形態論情報や文節係り受け情報 [5] も整備されている。国語研長単位形態論情報と文節境界は、一度長単位解析器 Comainu [9] で解析したものに対し、人手で修正しながら付与した。さらに、文節係り受け解析器 CaboCha [10] で解析したうえで、BCCWJ-DepPara [11] 互換の文節係り受けアノテーションを、人手修正もしつつ付与した。

書き言葉と話し言葉の大きな違いは、依存構造木の単位にある。書き言葉の場合には、句点などの文境界補助記号を手がかりにした文単位で依存構造木を構築されることが多い。しかし、話し言葉の場合には明示的な句点はないため、CEJC では節末もしくはポーズに基づいて定義される「発話単位」 [8] を認定されており、文節係り受けは発話単位に基づいて構築している。さらに、話し言葉には、書き言葉にはあまり見られないフィラー・言いよどみや述語省略などが出現することも特徴的である。

2.2 UD 体系への変換

本研究では、前節で説明した CEJC に付与された文節係り受けの情報を入力として、Omura ら [6] の変換ルールに基づき変換し、UD Japanese-CEJC を構築した。ただし、この変換ルールは UD Japanese-GSD、PUD、BCCWJ といった書き言葉 UD に対して用いていたものである。 そのため、話し言葉に対応するためにも、いくつかの拡張をおこなった。

まず、依存構造木の基本単位を発話単位とした。さらに、フィラー・言いよどみなど、係り先が規定できないものについては、最後の root へ係るようにした [1)]。さらに、述語省略などにより係り先が規定できない場合も、最後の root へ係るようにした。

図 1 に構築した UD の例を示す。図の場合、言いよどみの「ツ」やフィラーの「ん」などは一意に root の「でき」に係る形として変換している。このような、本質的に係り先が決められない文節・単語が出現することは、依存構造解析への困難さにつながると考えられ、後述のパーサ解析の結果からも伺える。

1) UD のガイドラインに従う場合の、複数の root が許容されていないために起きる処理である。

2.3 UD Japanese-CEJC の基礎統計

表 1 UD Japanese-CEJC（話し言葉）と UD Japanese-GSD（書き言葉）の基礎統計

Corpora	Unit	Trees	Tokens	Avg.	Max.
CEJC	短単位	59,319	256,885	4.3	84
	長単位	59,319	231,774	3.9	75
GSD	短単位	8,100	193,654	23.9	136
	長単位	8,100	150,243	18.5	108

表 1 に、今回構築した話し言葉の UD Japanese-CEJC と、比較のため、従前より整備している書き言葉の UD Japanese-GSD の基礎統計を示す。Trees の列は CEJC においては発話単位の数、GSD においては文の数を表す。Tokens の列は UD における単語/形態素の数を示す。Avg. の列は、係り受け木単位の平均単語/形態素数を示す。Max. は、係り受け木単位の最大単語/形態素数を示す。より詳細な UD の統計は A.2 節の表に示している。

3 パーサの解析評価

本節では UD Japanese-CEJC の解析可能性について検討する。UD Japanese-CEJC と UD Japanese-GSD を用いてパーサモデルを構築し UD 解析を行うことで、話し言葉と書き言葉の解析可能性の違いについて検証する。

3.1 実験設定

パーサの評価のため、UD Japanese v2.11 コーパスのうち、GSD、CEJC と、この 2 つを合わせたデータ CEJC+GSD を用いて比較を行う。GSD データは、既に train/dev/test に分割されているものを使用する。CEJC データについては、CEJC で区分されている「会話」形式に基づき、train/dev/test の比率が 8:1:1 になるように、均衡性を保ちながら新たに分割を行ったものを使用した。実験では、各コーパスの発話単位境界・文境界を与えたうえで構築した。**表 1** が示唆するように、発話単位と文境界認定自体には差異があり、比較が困難であるからである。実験では UDPipe および spaCy の 2 つのフレームワークを使用して依存構造解析を行い、精度を比較した。 UD には短単位と長単位のものがあるが、本稿では、短

表 2 GSD（書き言葉）と CEJC（話し言葉）についてのパーサ解析精度。いずれも短単位での評価である。

FW	train/dev	test	**Token**	**UPOS**	**XPOS**	**Lemmas**	**UAS**	**LAS**
UDPipe								
	GSD	GSD	96.18%	93.94%	93.11%	94.59%	85.06%	83.55%
	GSD	CEJC	81.60%	65.50%	64.11%	69.58%	61.11%	57.01%
	CEJC	GSD	81.57%	75.96%	65.61%	78.63%	47.60%	44.38%
	CEJC	CEJC	94.58%	91.48%	91.09%	92.46%	84.44%	82.19%
	CEJC+GSD	GSD	96.38%	94.45%	93.76%	94.95%	85.48%	84.02%
	CEJC+GSD	CEJC	95.98%	93.16%	93.04%	93.94%	86.55%	84.44%
	CEJC–	GSD	71.43%	66.60%	61.80%	68.78%	42.44%	40.34%
	CEJC–	CEJC–	94.95%	92.11%	91.75%	93.08%	85.29%	82.99%
spaCy two-stage analysis model (eliminating gold fillers and reparandums)								
	CEJC–	GSD	98.15%	84.54%	-	-	80.58%	71.97%
	CEJC–	CEJC–	96.38%	94.45%	-	-	89.71%	87.54%
spaCy simultaneous analysis model (including fillers and reparandums)								
	GSD	GSD	98.15%	97.05%	-	-	91.75%	90.87%
	GSD	CEJC	95.40%	78.53%	-	-	80.94%	75.03%
	CEJC	GSD	98.15%	84.33%	-	-	79.61%	70.54%
	CEJC	CEJC	95.40%	93.38%	-	-	88.27%	86.33%
	CEJC+GSD	GSD	98.15%	97.17%	-	-	91.52%	90.59%
	CEJC+GSD	CEJC	95.40%	93.46%	-	-	88.45%	86.68%

単位に限定した結果を示している。

UDPipe の実験では、UDPipe（v1.2.0）により構築したモデルで依存構造解析を行い、解析精度を評価した。UDPipe [12] は GRU やニューラルネットワークなどに基づいて実装されており、単語分割、品詞付与、依存構造解析をパイプラインで解析するモデルを構築できる。UDPipe では依存構造解析で単語埋め込みを使用できるため、258 億語規模のウェブテキストで学習された NWJC2vec [13] を使い、300 次元の Skip-gram モデルを使用した。

spaCy [2] の実験では、Transformers ベースの事前学習モデル[3] と解析コンポーネントとの間で損失勾配を共有できる spacy-transformers を使用して、次の 2 つのアプローチついて解析精度を評価した。

二段階解析モデル フィラー・言いよどみを除去してから UD 品詞付与と依存構造解析を行うモデルである。浅原 [14] にならい、フィラー・言いよどみのスパンを固有表現抽出器（ner）で学習して該当スパンを入力テキストから除去した上で、UD 品詞付与（morphologizer）と依存構造解析（parser）を別のパイプラインで処理する。なお、UD 品詞付与・依存構造解析の精度評価では、正解ラベルを用いてフィラー・言いよどみを除去した CEJC–を用いた [4]。

同時解析モデル フィラー・言いよどみを含めて入力全体を一度に処理するモデルである。単一の spaCy パイプライン上に transformers・morphologizer・parser・ner の順にコンポーネントを配置した。spaCy の parser コンポーネントは、Non-Monotonic Arc-Eager Transition System [15] をベースとして、交差文脈を扱うために Nivre [16] の Lifting による Projectivization/Deprojectivization の拡張が施されており [5]、フィラー・言いよどみから root への係り先が周辺文脈と交差する場合にも対応可能である。実際にフィラー・言いよどみの係り関係をどの程度の精度で扱うことができるか評価を行う。

なお、spaCy を用いた実験は時間の都合上、XPOS と Lemmas の体系変換は行わなかった。

3.2 結果

表 2 にパーサの評価結果を示す。**Tokens**、**UPOS**、**XPOS**、**Lemma** は、それぞれの列の再現可能性を F_1 スコアで示す。**Tokens** はわかち書き、**UPOS** は UD POS、**XPOS** は UniDic 品詞、**Lemma** は UniDic 語彙素である。依存構造関係は **UAS**（Unlabeled Attachment Score）と **LAS**（Labelled Attachment Score）という標準的な評価指標を用いた。いずれも CoNLL 2018 Shared Tasks [17] の評価プログラムに基づく。

訓練データ（train/dev）とテストデータ（test）で、

2） spaCy v3.4.3 および spacy-transformers v1.1.8 を使用
3） bert-base-japanese-v2 を使用
4） 入力テキストが異なる状態での精度評価が困難なため。
5） https://spacy.io/api/example#get_aligned_parse

表 3　二段階解析モデルと同時解析モデルのフィラー・言いよどみの解析精度

Category	Occurence	spaCy two-stage analysis model	spaCy simultaneous analysis model	
	train / dev / test	Token P / R / F	Token P / R / F	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%

話し言葉・書き言葉の差異がある条件（例「train/dev GSD と test CEJC」もしくは「train/dev CEJC と test GSD」）では、わかち書き (**Tokens**) と品詞タグ付け (**UPOS** と **XPOS**) の性能が悪くなった。これは、話し言葉と書き言葉で、語彙・品詞の分布が異なるからである。例えば、CEJC において INTJ・CCONJ・PRON（一人称代名詞・二人称代名詞など）の品詞が多いが、GSD においてはあまり出てこない。結果として、CEJC+GSD の両方のデータを訓練データに用いたモデルがもっとも性能が良いことがわかった。

3.2.1　フィラーや言いよどみによる影響

話し言葉には、フィラーや言いよどみなどがあり、本質的にわかち書きが難しい。**表 2** より、CEJC からフィラーや言いよどみを取り除いて学習と評価を行った二段階解析モデル（CEJC–の行）の性能が他のモデルより高いことは、そのひとつの裏付けと言える。また実用上フィラーや言いよどみが不要の場合には、二段階解析モデルが好まれるであろう。

フィラーや言いよどみに限定して解析精度を評価した結果を**表 3** に示す。トークン認定精度は二段階解析モデルよりも同時解析モデルの方が精度がやや高い傾向にあった。これは、依存構造解析とフィラー・言いよどみ判定を同時に学習することによる効果と考えられる。ただし、全体で評価した場合と比べて、フィラーや言いよどみのトークン化精度はどちらも 6 ポイント以上低下している。同時解析モデルの UD 品詞付与精度と依存構造解析精度においても、トークン化での劣化分に相当する精度低下が見られる。

3.2.2　依存構造木の長さと句読点の有無

表 1 でわかる通り、CEJC の依存構造木の長さは GSD の依存構造木の長さに比べて短い。 それでも、CEJC の依存関係の同定は品詞同定と同様に難しい結果となった。書き言葉の依存構造解析においては、句読点が長距離の依存関係の曖昧性解消に有効であるが、話し言葉には句読点が付与されていないために、長距離の依存関係同定に弱い傾向があると考えられる。さらに、フィラー・言いよどみ・述語省略など依存関係が不定なものの扱いが困難であると考えられる。CEJC–の項を確認すると、UDPipe の CEJC–のモデルは GSD の結果が 71.43%と悪くなっている。CEJC を用いたモデルは話し言葉の短さに基づき構築されているため、GSD のような書き言葉を生成するのが困難になっている。spaCy の場合、トークン化に形態素辞書を使用していること、事前学習モデルと同時学習を行っていること、などの効果で精度が安定していると考えられる。

4　おわりに

本稿では、『日本語日常会話コーパス』に基づいた新しい UD 日本語リソース UD Japanese-CEJC を紹介した。これは、日本語においてはじめての話し言葉の UD リソースであり、音声・映像アライメントも可能な世界最大規模の話し言葉 UD である。

さらに、構築した UD Japanese-CEJC（話し言葉）と UD Japanese-GSD （書き言葉）を用いた UDPipe と spaCy に基づくパーサを構築し、解析可能性の検証を行った。話し言葉においてフィラー・言いよどみ・言い直しなどがわかち書きや品詞タグ付けに影響を及すほか、依存関係同定も書き言葉と異なるふるまいを確認した。spaCy に基づく実験においては、固有表現抽出器を用いたフィラー・言いよどみ検出器を用い、フィラー・言いよどみを取り除いた状態で再度依存構造解析を行う二段階解析モデルを検討した。フィラーや言いよどみを予め除く二段階解析モデルは依存構造解析に有効であることを確認した。

今後の展開として、UD Japanese-CEJC と UD Japanese-BCCWJ との相互比較を検討している。また、UD Japanese-CEJC v2.11（Jan. 2023）[6] は、国立国語研究所により有償版 CEJC 契約者向けに「中納言」ダウンロードサイトから配布されているが、次回の v2.12（May. 2023 リリース）からは、Universal Dependencies の公式サイトでも表層形をスタンドオフした形式で配布する予定である。

実験で使用した spaCy 同時解析モデルは、Megagon Labs の GitHub リポジトリから公開を予定している。

6）長単位形態論情報・係り受けアノテーション自体のライセンスは CC BY 4.0 だが、表層形・短単位形態論情報の利用には有償版 CEJC の契約が必要である。

謝辞

本研究は、株式会社リクルート・国立国語研究所共同研究「日本語版 Universal Dependencies に基づく日本語依存構造解析モデルの研究開発」、国立国語研究所共同研究プロジェクト「実証的な理論・対照言語学の推進：アノテーションデータを用いた実証的計算心理言語学」、JSPS 科研費 JP18H00521, JP19K13195, JP22H00663, JP22K18483, の助成を受けたものです。

参考文献

[1] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal Dependencies. **Computational Linguistics**, Vol. 47, No. 2, pp. 255–308, June 2021.

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

[3] Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura, and Yugo Murawaki. Universal Dependencies version 2 for Japanese. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation**, Miyazaki, Japan, May 2018. European Language Resources Association.

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

[5] 浅原正幸, 若狭絢. 『日本語日常会話コーパス』に対する係り受け情報アノテーション. 言語処理学会 第 28 回年次大会 発表論文集, pp. 1699–1703, 3 月 2022.

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

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

[8] Yasuharu Den, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida. 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**, Valletta, Malta, May 2010. European Language Resources Association.

[9] 小澤俊介, 内元清貴, 伝康晴. 長単位解析器の異なる品詞体系への適用. 自然言語処理, Vol. 21, No. 2, pp. 379–401, 2014.

[10] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.

[11] Masayuki Asahara and Yuji Matsumoto. BCCWJ-DepPara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In **Proceedings of the 12th Workshop on Asian Language Resources**, pp. 49–58, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[12] Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing ud 2.0 with UDpipe. In **Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies**, pp. 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[13] Masayuki Asahara. NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'. **Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication**, Vol. 24, pp. 7–22, January 2018.

[14] Masayuki Asahara and Yuji Matsumoto. Filler and disfluency identification based on morphological analysis and chunking. In **Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition**, pp. 163–166, Tokyo, Japan, April 2003. ISCA.

[15] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

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

[17] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 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**, pp. 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics.

A 付録

A.1 UD Japanese-CEJC の特徴

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

表 1 CEJC 転記ファイルの (cf. Dobrovoljc 論文 [2], Table 2)

表 1 に CEJC 転記ファイルの特徴を Dobrovoljc の論文 [2] の観点に基づいて示す。Language variety は日本語共通語話者（関東近県在住）のために多様性がない。また、日本語には Capitalization の慣習がない。さらに、CEJC は転記規則として punctuation を利用していない。他言語の UD リソースにない特色として、映像データとのアラインメントがある。

A.2 UD Japanese-CEJC の統計

UPOS	GSD(SUW)	CEJC(SUW)
ADJ	3839	9469
ADP	41864	34958
ADV	2364	17321
AUX	21158	34021
CCONJ	819	4223
DET	987	1442
INTJ	13	27578
NOUN	58184	38181
NUM	5163	4291
PART	1259	21813
PRON	1108	9684
PROPN	7141	3573
PUNCT	19233	0
SCONJ	7995	17172
SYM	1301	0
VERB	21226	25329
X	0	7830

表 2 UPOS の分布（GSD（短単位）と CEJC（短単位））

表 2 に GSD（短単位）と CEJC（短単位）の UPOS（品詞）ラベルの分布を示す。話し言葉には句読点や記号を転記に用いないために PUNCT と SYM が全く出現しない。また、話し言葉は、CCONJ と INTJ が書き言葉より大きな割合で出現するが、PRON は省略される傾向がみられる。X は、笑・泣・歌などといった書き言葉では見られず、いずれにも該当しない表現に付与している。

DEPREL	GSD(SUW)	CEJC(SUW)
acl	6998	5421
advcl	7197	9950
advmod	2280	12163
amod	445	250
aux	17235	23382
case	41307	32669
cc	819	4096
ccomp	390	881
compound	27489	10200
cop	2441	5085
csubj	157	234
dep	76	2573
det	987	1393
discourse	16	7000
dislocated	0	0
fixed	8620	10648
mark	7860	36481
nmod	12970	7362
nsubj	8242	6450
nummod	2800	2522
obj	5309	1224
obl	12683	14482
reparandum	0	3100
punct	19233	0
root	8100	59319

表 3 DEPREL の分布（GSD（短単位）と CEJC（短単位））

表 3 に GSD（短単位）と CEJC（短単位）の DEPREL（依存構造関係）ラベルの分布を示す。話し言葉においては、係り先が不定のものを文末の root ノードを係り先とするために root が大きな割合となる。また、reparandum は言いよどみ、discourse はフィラーを意味する。UPOS PUNCT が話し言葉には出現しないために、DEPREL ‘punct’ も出現しない。