現代日本語書き言葉均衡コーパスのUniversal Dependencies

現代日本語書き言葉均衡コーパスの Universal Dependencies

大村　舞 (国立国語研究所コーパス開発センター) *
浅原　正幸 (国立国語研究所コーパス開発センター)

Universal Dependencies Annotation for 'Balanced Corpus of Contemporary Written Japanese'

Mai Omura (National Institute for Japanese Language and Linguistics)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)

要旨

自然言語処理の分野では多言語かつ言語横断的な言語研究が盛んに取り組まれている。その言語横断的な言語研究の取り組みとして Universal Dependencies(UD) がある。UD では品詞や係り受け構造の標準・スキーマを定め、多言語のコーパスを提供している。本論文では、日本語コーパスである現代日本語書き言葉均衡コーパス (BCCWJ) を UD のスキーマへと変換したコーパスについて紹介をする。BCCWJ では日本語における文節単位の係り受け情報がすでに付与されている。この係り受け構造を基にして UD へと変換するプログラムの開発を行った。しかし、文節単位は UD の単語単位には沿っていない。そのため、BCCWJ で提供されている短単位と長単位というふたつの言語単位を単語の単位をして認定したコーパスを構築する。短単位と長単位について UD のスキーマに当てはめた場合、どのような係り受け構造ができるのかを示す。

1. はじめに

Universal Dependencies[(1)](以下 UD) は、多言語で一貫した構文構造とタグセットを定義し、言語間での共通した依存構造タグ付きコーパスを提供することを目的とした活動、あるいはそのコーパスのことを指している。我々は UD の日本語版を設計する活動として、品詞体系、ラベル付き依存構造の定義の策定、その github 上での文書化と、参照用のコーパスの作成に着手している。本稿ではこの UD 日本語版設計の活動の一環として、現代日本語書き言葉均衡コーパス（以下 BCCWJ）(Maekawa et al. 2014) に基いた日本語 UD コーパスについて紹介する。

既存の日本語依存構造タグ付きコーパスとして、京都大学テキストコーパス (Kurohashi and Nagao 2003)・日本語係り受けコーパス (Mori et al. 2014) などが存在する。また、UD 基準の依存構造タグ付きコーパスとして、日本語句構造ツリーバンク (田中ほか 2014, Tanaka and Nagata 2013) を変換した日本語版 UD コーパス UD Japanese KTC (Tanaka et al. 2016) が

* mai-om@ninjal.ac.jp
(1) Universal Dependencies v2 http://universaldependencies.org/

公開されている。このコーパスは日本語句構造ツリーバンクにある形態素や句構造などのアノテーションを用いて変換されたものである。本データは UniDic の短単位を単語の単位として採用している。そのほか Wikipedia 由来の UD Japanese (Japanese 無印) や、パラレルコーパス由来の UD Japanese PUD (Japanese-PUD) があるが人手による修正が行われていない。(2)

本論文では、BCCWJ を UD の体系へと変換したコーパスを紹介する。BCCWJ には、短単位・長単位の形態論情報だけでなく文節単位の依存構造・並列構造アノテーションである BCCWJ-DepPara (Asahara and Matsumoto 2016) や述語項構造情報アノテーションである BCCWJ-PAS (植田ほか 2015) が提供されている。これらの情報に対する変換プログラムを作成することで，Universal Dependencies の議論に即した木構造の変換に対応することができる。以下では、現在行っている体系の概要について解説する。

2. BCCWJ の Universal Dependencies 化

2.1 Universal Dependencies の構成

UD では、語順が自由な言語も含めて言語横断的に共通化した体系を確立するために、句構造を考慮せず、すべての構文構造を単語間の依存関係と関係のラベルで表現する。異なる言語間で依存構造解析器の性能比較を行うだけでなく、言語学的に類型論的な分析が可能にすべく言語横断的な設計を目指している。図 1 のように内容語間の依存構造を中心とした表現を用いる。現在のアノテーション体系は version 2.0 は、Google Universal Part-of-speech Tags (Petrov et al. 2012) を基にして表 1 のような 17 種類の品詞ラベル Universal PoS tags が定義されている。さらに Universal Stanford Dependencies(Marie-Catherine et al. 2014) を基にして表 2 のような 37 種類の係り受けのラベル Universal dependency relations が定義されている。

2.2 BCCWJ の Universal Dependencies 化

『現代日本語書き言葉均衡コーパス』(BCCWJ)(Maekawa et al. (2014)) は、書籍全般、雑誌全般、新聞、白書、ブログ、 ネット掲示板、教科書、法律などのジャンルにまたがって 1 億 430 万語のデータを格納したコーパスであり、現在、日本語について入手可能な唯一の均衡コーパスである。このうちコアデータである 1980 サンプル・57256 文には形態論情報が付与されており、文節依存構造・並列構造・述語項構造が付与されている。既存のアノテーションに基づき、変換プログラムを構築することで、UD 本体の基準の変更や日本国内での議論に対応することができる。

以下では単位認定・品詞割り当て・依存構造ラベル割り当て・ファイルフォーマットについて説明する。

(2) UD 基準の依存構造タグ付きコーパス (Japanese 無印, Japanese-KTC, Japanese-PUD) は `http://universaldependencies.org/` にて配布されている。

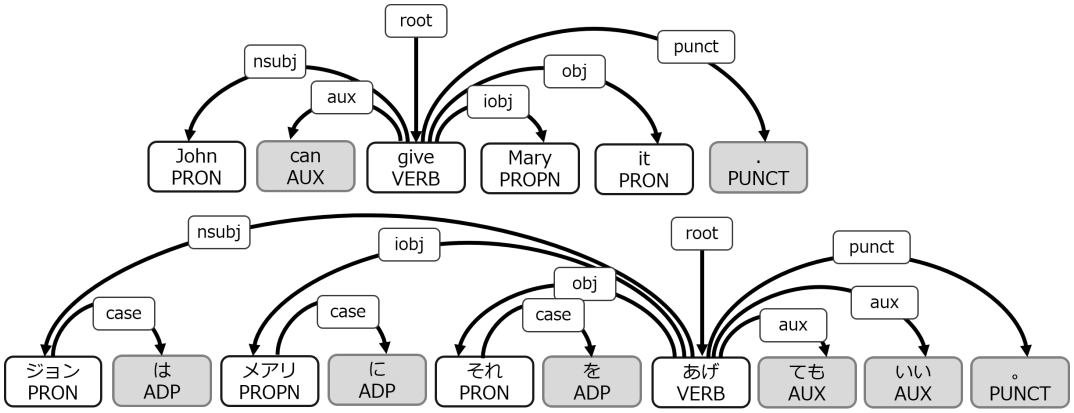


図 1　Universal Stanford Dependencies のイメージ．上が英文、下が日本語．助動詞や格助詞など、英語と日本語の違いがあっても、内容語の関係は保たれている

表 1　Universal PoS version 2.0 一覧

NOUN	名詞
PROPN	固有名詞
VERB	動詞
ADJ	形容詞
ADV	副詞
INTJ	間投詞
PRON	代名詞
NUM	数詞
DET	限定詞
ADP	接置詞
AUX	助動詞
PART	接辞
CONJ	接続詞
SCONJ	従属接続詞
PUNCT	句読点
SYM	記号
X	その他

2.3　単語認定

日本語は英語とは異なり、単語に分割されていない。そのためまず単語の認定について決める必要がある。基本的に BCCWJ で用いられている単語単位をベースとして UD を構築する。

BCCWJ のすべてのサンプルは**短単位・長単位**という言語単位に基づいて形態素解析されている。**短単位**は日本語の形態的側面に着目した規定した単位であり、語種ごとに規定した最

表 2　37 種類の係り受けのラベル Universal dependency relations 一覧

	格要素	節	修飾語	機能語
核となる要素	nsubj obj iobj	csubj ccomp xcomp		
その他の要素	obl vocative expl dislocated	advcl	advmod discourse	aux cop mark
名詞関連	nmod appos nummod	acl	amod	det clf case
並列	複数の単語	いろいろ	特殊なケース	その他
conj cc	fixed flat compound	parataxis list	orphan goeswith reparandum	punct root dep

短単位	魚 NOUN	フライ NOUN	を ADP	食べ VERB	た AUX	か PART	も ADP	しれ VERB	ない AUX	ペルシャ PROPN	猫 NOUN
長単位	魚フライ NOUN		を ADP	食べ VERB	た AUX	かもしれない AUX				ペルシャ猫 NOUN	
文節	魚フライを			食べたかもしれない						ペルシャ猫	

図 2　長単位・短単位・文節のイメージ

小単位の線形結合に基づき定義されている。**長単位**は日本語の構文的な機能に着目して規定した単位であり、文節の構成要素ともなっている。このうち文節単位の依存構造アノテーションとして BCCWJ-DepPara(Asahara and Matsumoto 2016) がある。短単位・長単位・文節は図 2 のように短単位＜長単位＜文節という階層関係が成り立っている。

BCCWJ 版 UD では、Tanaka et al. (2016) に倣い、BCCWJ の品詞体系である短単位を基本単位として採用する。ただし、以降の節で説明する通り、UD や他の言語と基準を合わせるためには長単位に付属している用法の情報も必要となる。過去の研究では短単位を基準として調査されているものが多く、長単位について議論されているものは少ない。このため、短単位ベースと長単位ベースのコーパスどちらも準備する予定である。以降は短単位を基本として議論するものの、長単位に基づく変換規則についても言及する。

2.4　Universal PoS tags への変換

UD では全言語の品詞を集約するための体系として Universal PoS version 2.0 を採用している。Universal PoS version 2.0 では、表 1 に示す 17 種の品詞が定義されている。品詞の細分類や、性数・時制・格など文法的属性に関するものは、`FEATS` や `MISC` など列に言語ごとの

個別に定義する属性値 (features) を持たせることで情報が失われないようにしている。

日本語版の UD では UniDic(伝ほか 2007) と Universal PoS tags との対応表を構築することで UD の品詞を定義する。BCCWJ は、短単位では**語彙主義**的な可能性に基づく品詞体系を採用している。例えば「名詞-普通名詞-副詞可能」は「名詞」用法も「副詞」用法もある語彙であることを意味する。長単位では文脈に基づいてこの用法の曖昧性を解消する**用法主義**に基づく品詞を規定している。BCCWJ には長単位形態論情報として「用法」の情報が付与されている。

Universal PoS tags への変換は基本的には語彙主義に基づく品詞で対応付けを行う。UniDic は短単位の品詞体系であり、辞書的、語彙的な品詞情報を規定している。そのため短単位においては、UniDic をそのまま品詞体系として用いることが可能である。

ただし、いくつかの単語に関しては、用法主義に基づく品詞体系を用いる。例えば、サ変名詞や形状詞の場合は語彙主義に基づく品詞体系ではなく、文脈に基づいて用法の曖昧性を解消する用法主義に基づく品詞を用いる。用法主義に基づく品詞のほうが、他の言語との対応がとりやすいという利点があるということと、語尾の有無などにより揺れが少なく `VERB`, `ADJ` とする条件を規定し易いからである。以上を踏まえて Universal PoS tags と UniDic の対応を表 3 に示す。この議論は Tanaka et al. (2016) でも言及されている。

2.5 係り受け構造の変換

BCCWJ-DepPara (Asahara and Matsumoto 2016) には BCCWJ に対する文節係り受け情報が含まれている。文節単位係り受けを単語単位の係り受けに変換するために、文節内の主辞を決定し、文節内の他要素に関してはすべて文節内の主辞に係けるようにする。文節内の主辞は内容語と機能語が分かれる内容語の最右の語を採用する。具体的には CaboCha (工藤・松本 2002) に実装されている 文節の UniDic 主辞決定規則 (`selector.cpp`) をもとに若干変更したうえで実装した。図 3 に主辞決定規則を示す。

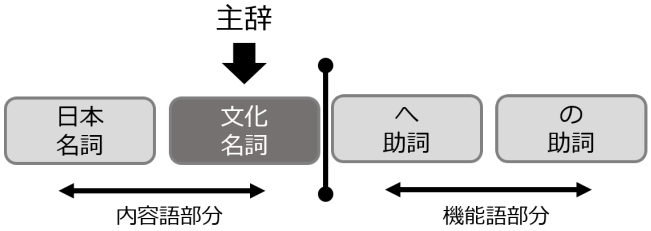
しかし、BCCWJ-DepPara は係り先情報は含んでいるものの、Universal dependency relations に対応する係り受けの統語的な用法 (ラベル `nsubj, obj, iobj` など) を含んでいない。そこで、BCCWJ-PAS (植田ほか 2015) の述語項構造情報と係り受け関係がある単語対の品詞情報などから Universal dependency relations ラベルを決定して割り当てている。表 4 に Universal dependency relations 割り当ての規則を示す。

現在の規則は節であるか否かの判別を行っておらず、`csubj, advcl, acl` などの節関連のラベルについては解決できていない。今後、節であるか否かについて基準を作成することで、解決をはかる。

また BCCWJ-DepPara には並列構造情報が含まれているが、今回の変換規則については `cc`, `conj` などの並列構造関連の規則がまだ定義できていない。これについても今後検討していく。

2.6 UD のファイル形式：CoNLL-U フォーマット

UD では、ファイル形式として CoNLL-U 形式が採用されている。表 5 のような 10 列で構成された、タブ区切りのテキストファイルとなっている。



機能語部分と内容語部分が分かれるように
主辞を決定する。

手順:

文節内の単語を見る

1．その単語の品詞が/助詞|助動詞|接尾辞，形容詞的|接尾辞，形状詞的|接尾辞，動詞的/にマッチする
->その前の単語が主辞

2．その単語の品詞が/助詞|助動詞|接尾辞，形容詞的|接尾辞，形状詞的|接尾辞，動詞的/にマッチしない
->次の単語を見る

3．最後の単語である -> 前の単語が主辞

図 3　文節の主辞決定規則

```
# sent_id = OC01_00001-1
# text = 詰め将棋の本を買ってきました。
1    詰め    tsumeru VERB    動詞-一般    _    2    compound    _    bpos="CONT"
2    将棋    shougi  NOUN    名詞-普通名詞-一般    _    4    nmod    _    bpos="SEM_HEAD"
3    の    no    ADP 助詞-格助詞    _    2    case    _    bpos="SYN_HEAD"
4    本    hon NOUN    名詞-普通名詞-一般    _    6    dobj    _    bpos="SEM_HEAD"
5    を    wo    ADP 助詞-格助詞    _    4    case    _    bpos="SYN_HEAD"
6    買っ    kau VERB    動詞-一般    _    8    advcl    _    bpos="SEM_HEAD"
7    て    te    SCONJ    助詞-接続助詞    _    6    mark    _    bpos="SYN_HEAD"
8    き    kuru    VERB    動詞-非自立可能    _    0    root    _    bpos="ROOT"
9    まし    masu    AUX 助動詞    _    8    aux    _    bpos="FUNC"
10    た    ta    AUX 助動詞    _    8    aux    _    bpos="SYN_HEAD"
11    。    PUNCT    補助記号-句点    _    8    punct    _    bpos="CONT"

# sent_id = OC01_00001-2
# text = 駒と盤は持っていません。
1    駒    koma    NOUN    名詞-普通名詞-一般    _    3    nmod    _    bpos="SEM_HEAD"
2    と    to    ADP 助詞-格助詞    _    1    case    _    bpos="SYN_HEAD"
3    盤    ban NOUN    名詞-普通名詞-一般    _    5    dobj    _    bpos="SEM_HEAD"
4    は    ha    ADP 助詞-係助詞    _    3    case    _    bpos="SYN_HEAD"
5    持っ    motsu    VERB    動詞-一般    _    0    root    _    bpos="ROOT"
6    て    te    SCONJ    助詞-接続助詞    _    5    mark    _    bpos="FUNC"
7    い    iru VERB    動詞-非自立可能    _    5    aux    _    bpos="FUNC"
8    ませ    masu    AUX 助動詞    _    5    aux    _    bpos="FUNC"
9    ん    zu    NEG 助動詞    _    5    aux    _    bpos="SYN_HEAD"
10    。    PUNCT    補助記号-句点    _    5    punct    _    bpos="CONT"
....
```

図 4　BCCWJ の UD サンプル (OC01_00001)．上記のようにタブ区切りのテキストファイルになる．MISC 列には文節情報などを付与する。

実際のサンプルを図 4 に示す。これは短単位で解析した結果であり、本稿執筆時点での開発段階のものである。実際には MISC 列などに、日本語特有の言語情報を付与することで、言語の特徴を用いた言語解析の研究などに利用できるようにする。データセットは http://universaldependencies.org/ にて公開する。

3. おわりに

本稿では日本語コーパスである現代日本語書き言葉均衡コーパス (BCCWJ) を UD の体系へと変換したコーパスについて紹介した。

本稿執筆時点では、短単位ベースを元にした UD の変換まで完了している。今後は長単位ベースのコーパスも実装し、短単位・長単位ベースの日本語 UD データを公開予定である。

謝　辞

本研究（の一部）は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度) の成果である。

文　献

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

Sadao Kurohashi, and Makoto Nagao (2003). *Building a Japanese Parsed Corpus – while Improving the Parsing System.*, Chap. 14 pp. 249–260.: Kluwer Academic Publishers.

Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada (2014). “A Japanese Word Dependency Corpus.” *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC'2014)*, pp. 753–758.

田中貴秋・永田昌明・松崎拓也・宮尾祐介・植松すみれ (2014).「統語情報と意味情報を統合した日本語句構造ツリーバンクの構築」　言語処理学会第 20 回年次大会発表論文集, pp. 737–740.

Takaaki Tanaka, and Masaaki Nagata (2013). “Constructing a Practical Constituent Parser from a Japanese Treebank with Function Labels.” *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL'2013)*, pp. 108–118.

Takaaki Tanaka, Yusuke Miyao, Masayuki Asahara, Sumire Uematsu, Hiroshi Kanayama, Shinsuke Mori, and Yuji Matsumoto (2016). “Universal Dependencies for Japanese.” *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC2016)*, pp. 1651–1658.

Masayuki Asahara, and Yuji Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annotation Treebank on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション」　第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214.

Slav Petrov, Dipanjan Das, and Ryan McDonald (2012). “A universal part-of-speech tagset.” *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC2016)*, pp. 2089–2096.

De Marneffe Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning (2014). “Universal Stanford depen-

dencies: A cross-linguistic typology." *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC'2014)*, pp. 4585–4592.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」 日本語科学, pp. 101–123.

工藤拓・松本裕治 (2002).「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, 43:6, pp. 1834–1842.

表 3　Unidic → UD PoS tags 変換規則

短単位品詞	短単位基本形	長単位用法	UD PoS
^形容詞-非自立可能		形容詞-一般	ADJ
^形容詞-非自立可能		助動詞	AUX
^形容詞			ADJ
^連体詞	^[こそあど此其彼] の		DET
^連体詞	^[こそあど此其彼]		PRON
^形状詞-一般			ADJ
^形状詞-タリ			ADJ
^形状詞-助動詞語幹			AUX
^副詞			ADV
^感動詞			INTJ
^名詞-普通名詞-一般			NOUN
^名詞-普通名詞-サ変可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-サ変可能		動詞-一般	VERB
^名詞-普通名詞-形状詞可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-形状詞可能		形状詞-一般	ADJ
^名詞-普通名詞-副詞可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-副詞可能		副詞	ADV
^名詞-普通名詞-サ変形状詞可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-サ変形状詞可能		形状詞-一般	ADJ
^名詞-普通名詞-サ変形状詞可能		動詞-一般	VERB
^名詞-普通名詞-助数詞可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-助数詞可能		名詞-数詞	NUM
名詞-数詞			NUM
^名詞-助動詞語幹			AUX
^名詞-固有名詞			PROPN
^動詞-非自立可能		動詞-一般	VERB
^動詞-非自立可能		助動詞	AUX
^動詞			VERB
^助動-[格係副] 助詞			ADP
^助動詞			AUX
^接続助詞	て		SCONJ
^接続助?詞			CCONJ
^連体詞			ADJ
^助詞-準体助詞			SCONJ
^助詞-[^格接副]			PART
^代名詞			PRON
^補助記号-(句点読点—括弧)—			PUNCT
^補助記号			SYM
^記号			SYM
^空白			X
^接頭辞			NOUN
^接尾辞			PART

表 4　UD 係り受け変換規則

係り元の UD 品詞	係り元の UniDic 品詞	係り先の品詞	bccwj-pas ラベル	UD ラベル
NUM				nummod
CCONJ				cc
ADV				advmod
ADJ				amod
INTJ				discourse
PROPN				name
NOUN or PRON				nmod
VERB		VERB		advcl
VERB		ADJ		advcl
VERB		NOUN		acl
VERB		PRON		acl
VERB		NUM		acl
			bccwj-pas:ni	iobj
			bccwj-pas:o	obj
			bccwj-pas:ga	nsubj
	助詞-[格副係] 助詞			case
SCONJ				mark
VERB				aux
PART				aux
PUNCT				punct
NUM				nummod
NEG				neg
NOUN		NOUN		compound
NOUN		NUM		compound
	サ変	NOUN		compound
VERB		NOUN		compound
X				dep
SYM				dep

表 5　CoNLL-U 形式の各列の説明

列	フィールド名	説明
1	ID	1-origin の ID (ROOT が 0)
2	FORM	書字形出現形
3	LEMMA	語彙素読みをローマ字にしたもの
4	UPOSTAG	品詞 Universal POS
5	XPOSTAG	品詞 BCCWJ 短単位品詞
6	FEATS	その他品詞情報 （“—”で OR を表現、順不同）
7	HEAD	係り先 ID
8	DEPREL	係り受け関係
9	DEPS	Secondary Dependency (List, Head-deprel pairs)
10	MISC	その他 （文節内の主辞情報）