日本語 Universal Dependencies の試案

金山 博 †
hkana@jp.ibm.com
宮尾 祐介 *
yusuke@nii.ac.jp
田中 貴秋 ††
tanaka.takaaki@lab.ntt.co.jp
森 信介 ‡
forest@i.kyoto-u.ac.jp
浅原 正幸 ‡‡
masayu-a@ninjal.ac.jp
植松 すみれ *
uematsu@nii.ac.jp
† 日本アイ・ビー・エム株式会社 東京基礎研究所
* 情報・システム研究機構 国立情報学研究所
†† NTT コミュニケーション科学基礎研究所
‡ 京都大学学術情報メディアセンター ‡‡ 人間文化研究機構 国立国語研究所

1 はじめに

言語横断的な係り受け構造を設計する試みとして、Universal Dependencies (UD) [1, 5] と、そこで使われる単語の品詞体系の Universal PoS tags [7] が注目を浴びている。著者らは、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た。本稿では、その最初の報告として、定義の原案とこれまでの主な論点、また既存のコーパスを UD の形式に変換する試みの状況について述べる。

2 Universal Dependencies

Universal Dependencies (UD)[1, 5] は、構文解析の後段の処理の共通化や、他の言語のコーパスを用いた言語横断的な学習、言語間の定量的な比較などを可能にするための土台を目指して、多言語で一貫した構文構造とタグセットを定義するという活動である。

表現や言語資源作成を単純化するため、またくだけた文や特殊な構造に対して頑健にするために、句構造 (constituent) 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方針 (lexicalism) を取っており、単語の品詞体系として Google Universal Part-of-speech Tags [7] を、係り受けのラベルとして Universal Stanford Dependencies[2] を基にして設計されている。以下で、現在 github 上 [1] で公開されている UD の定義について記述する。

2.1 単語と品詞体系

Universal PoS version 2.0 では、全言語の品詞を集約するための体系として、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、言語ごとに個別に定義する属性値 (features) を持たせて、情報が失われないようにしている。それらの属性も、言語間で類似した現象に対して共通した表現を与えることを目指して、各言語の議論を github 上で参照できるようにしている。

英語の場合、Penn Treebank のタグとの対応付けをもって品詞タグの定義としている*1。ほとんどは直感的であるが、以下のものは注意を要する。

表 1 Universal PoS 2.0 の 17 種の品詞タグセット。

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

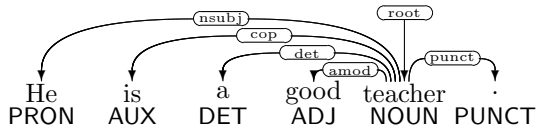
ADP Penn Treebank の IN（前置詞等）のうち従属接続詞を除いたもの、TO（to）のうち前置詞となるもの。
PART 所有格の ''s', 否定の 'n't', 不定詞の 'to' のみ。
CONJ 'and', 'or' などの等位接続詞。
SCONJ 'when', 'since' など副詞的に使われる従属接続詞、'that' などの補文標識と、関係代名詞の 'that'。
AUX 'can' などのモダリティの助動詞や be 動詞および、受動態を示す 'get'。

2.2 依存関係とラベル

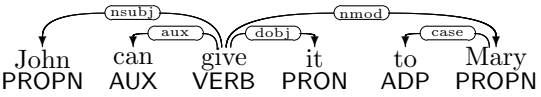
UD では、上記の品詞タグが付与された二単語の間に方向を持った依存関係を付与して、構文構造を記述する。文の主辞 (root) 以外の語はいずれかの語を修飾する形とするため、文全体は木構造となる。依存関係に付与するラベルとして、USD で定義された 42 種のラベルを一部改変した 40 種を用いる。

主語と動詞の呼応などの文法的な対応関係を厳密に考慮せずに、内容語相互の関係を重視することにより、言語間の構造の違いを吸収している。その典型的な例としてコピュラの扱いが挙げられる。英語の例文 (1) にあるように、UD では teacher を主辞 (root) とし、he と teacher の間に直接の依存関係を付与している。be 動詞を主辞、he と teacher はそれぞれ be の主語と補語だと捉える従来の考え方と異なるが、これにより、欧米語特有の be 動詞に特化しない依存構造を得ることができる。

*1 例えば、Universal PoS の ADJ は、Penn Treebank の JJ(形容詞), JJR(比較級), JJS(最上級) と定義されている。

(1) 

コピュラと同様に、助動詞や前置詞も内容語の子とする。例 (2) では、John は助動詞 can ではなく本動詞 give の主語として nsubj で結ばれ、前置詞 to を伴う Mary は nmod で直接 give に係り、to は Mary の子となる。これにより、述語との位置関係や前置詞で格を表現する英語、後置詞で格を示す日本語、名詞が格変化するロシア語の間で、内容語間の依存関係を統一させることができる。

(2) 

3 日本語 UD の定義

空白による明示的な単語の境界を持たない日本語においては、何をもって語の単位とするかは自明ではない。日本語 UD を考えるにあたって、そもそも単語とは何であるかといった議論を避けるとともに、既存の辞書やコーパスを UD の形式に自動変換できることを目指し、UniDic [3, 10] の語彙項目、すなわち BCCWJ コーパス [4] の短単位を単語とする方針とした。

3.1 日本語の品詞体系

英語版の Universal PoS を Penn Treebank の品詞体系とのマッピングをもって定義しているのに倣い、日本語版は UniDic との対応をもとに品詞を定義する。

UniDic は辞書項目の集合であるため、例えば「勉強」という語であれば「名詞-サ変可能」と、文脈によってサ変動詞になりうる名詞であるという品詞となっている。以下の定義では、これを NOUN と VERB に振り分けるように、純粋な UniDic の品詞とのマッピングだけでなく、文脈を考慮した条件が加わることがあることに注意されたい。この是非は 4 節で議論する。

なお、以下では、UniDic の品詞大分類と中（小）分類をまとめて「助詞-格助詞」のように記述する。‘ ’ で囲まれた語は例を示し、下線がある場合にはその部分が該当する語である。

NOUN　名詞-普通名詞 ‘パン’（但し VERB, ADJ として使われるものを除く）。
PROPN　名詞-固有名詞 ‘北海道’。
VERB　動詞（但し非自立となるものを除く）‘食べ’・名詞-サ変可能で動詞の語尾が付いたもの ‘食事 する’。
ADJ　形容詞 ‘大きい’（但し非自立となるものを除く）・形状詞[*2] ‘豊か’・連体詞（但し DET を除く）‘大きな’・名詞-形状詞可能 で形状詞の語尾が付く場合 ‘自由 な’。
ADV　副詞 ‘ゆっくり’。
INTJ　感動詞 ‘あっ’。
PRON　代名詞 ‘私’。
NUM　名詞-数詞 ‘5’。
AUX　助動詞 ‘た’・動詞/形容詞のうち非自立のもの ‘して いる’, ‘食べ にくい’。
CONJ　接続詞 ‘または’・助詞-接続助詞のうち、等位接続詞として用いるもの ‘と’。
SCONJ　接続詞・助詞-接続助詞 ‘て’（CONJ となるものを除く）・準体助詞 ‘行く の が’。
DET　連体詞の一部 ‘この’ ‘その’ ‘あんな’ ‘どんな’。
ADP　助詞-格助詞 ‘が’/副助詞 ‘しか’/係助詞 ‘こそ’。
PART　助詞-終助詞 ‘か’・接尾辞 ‘衝撃 的 だ’。
PUNCT　補助記号-句点/読点/括弧開/括弧閉。
SYM　記号・補助記号のうち PUNCT 以外のもの。
X　空白。

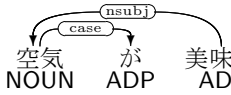
[*2] 他の体系で「形容動詞」「ナ形容詞」と呼ばれるもの。

3.2 日本語の依存構造とラベル

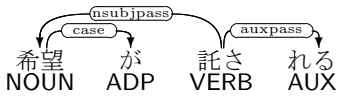
品詞と同様に、UD のラベルに基づいて、対応する日本語の現象と例を記述していく。但し、必ずしも網羅性をもった定義ではなく、まだ未解決の問題も多い。UD その他で使われている記法に基づき、矢印の起点側が主辞、終点側が修飾語を示す。

●述語の要素

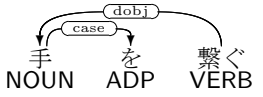
nsubj　主格で述語に係る名詞句。

(3) 

nsubjpass　主格で受身の助動詞を伴う用言に係る名詞句。

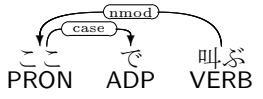
(4) 

dobj　目的格で述語に係る名詞句。

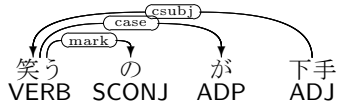
(5) 

iobj　格助詞「に」を伴うなどして述語に係る名詞句。

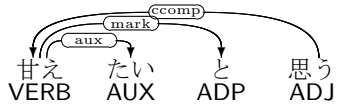
nmod　これまでに示した以外の格の名詞句や、時相名詞により用言を修飾する場合。

(6) 

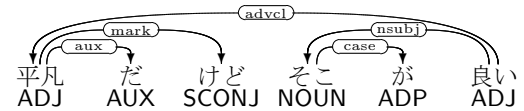
csubj　主語になる名詞節。準体助詞を伴う用言句が主語となる場合。

(7) 

ccomp　補文。

(8) 

advcl　副詞節。主に接続助詞を伴って用言を修飾する節。

(9) 

advmod　副詞による修飾。

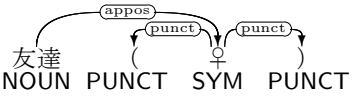
neg　否定語の付与。

(10) 

●名詞の修飾

nummod　数量の指定。

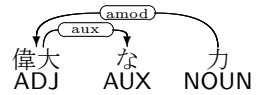
appos　同格の表現。

(11) 

acl　連体修飾節。但し amod に該当する場合を除く。このほか「てからの」「ながらの」などの接続表現。

(12) 

amod　形容詞・形状詞・連体詞（DET 以外）が格を伴わずに名詞を修飾する場合。

(13) 

det　DET による修飾。

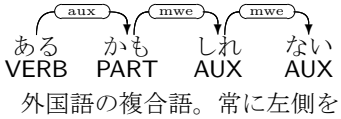
●複合語

compound　名詞と名詞・動詞と動詞の複合。

(14) 

name　固有名詞の複合語。

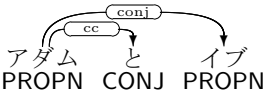
mwe　機能表現の複合語。

(15) 

foreign　外国語の複合語。常に左側を主辞とする。

●並列

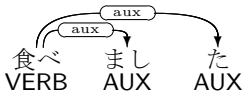
conj　並列構造。左側の要素を主辞とする。

(16) 

cc　等位接続詞。構造は (16) を参照。

●その他の要素

aux　用言に付く助動詞や、非自立の補助用言。「か」などの終助詞を含む。

(17) 

cop　コピュラの「だ」が付く場合。

mark　従属接続詞、接続助詞、補文標識の「と」「か」などが付く場合。例 (7)(8) などを参照。

case　助詞による格の表示。(3) などの例を参照。

punct　句読点。

その他　xcomp, expl など日本語に該当する事象が無いものや、goeswith, vocative, list, remnant など言語非依存の特殊なラベルについては割愛する。

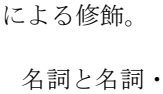
4　議論のポイント

3 節に示した日本語の品詞や依存構造を考える際に、議論となった点を挙げる。一部はまだ議論が続いており、先述の定義も今後変更される可能性がある。

サ変動詞と形状詞　3.1 節の冒頭で触れたサ変動詞や形状詞の扱いは、品詞の規定の方法論と深く関連する。一つの立場は、文脈によらず語彙自体が持つ品詞を表示する**語彙主義**で、単純に UniDic を形態素解析器 MeCab に適用した際の出力などがこれに相当する。

BCCWJ コーパスにおいては、短単位で語彙主義的な品詞規定を行い、長単位では文脈に基づいて用法の曖昧性を解消する**用法主義**に基づく品詞を規定するという、二重形態素解析を行っている。BCCWJ の人手によるアノテーションや長単位解析器 Comainu などを用いた場合は、文脈に応じて、サ変名詞や形状詞語幹を動詞や形容詞と判定することができる。

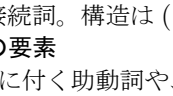
日本語 UD では、語彙主義に則りサ変名詞を常に名詞とする (18a) ではなく、用法主義に基づく (18b) の形式とすることを考えている[*3]。これは、後者のほうが実態に近い構造が得られ、他の言語との対応が取りやすいという利点があることと、語尾の有無などにより揺れが少なく VERB, ADJ とする条件を規定できるからである[*4]。

(18a) 

(18b)

非自立　UniDic において、「走っている」「来てほしい」などの補助用言は、「動詞 (形容詞)-非自立可能」のように、その辞書項目が自立語的にも非自立語的にも利用可能であることを示す品詞ラベルが用いられる。二重形態素解析の結果があれば、当該語が自立語か非自立語かが識別され、一意に AUX か否かが規定できる場合がある。AUX を区別したほうが依存構造が単純化されるという利点がある一方、否定のスコープなど一部の構文構造が失われるという問題点がある。

品詞を変える接辞　UniDic の接尾辞は、形容詞を名詞化する「さ」なら「接尾辞-名詞的」、名詞を形容詞化する「っぽい」であれば「接尾辞-形容詞的」のように、構成する長単位（複合語）の品詞を規定する情報を持つ。UD の表現で表現するにあたって、サ変名詞と同様、以下の二通りが考えられよう。

(19a) 

(19b)

このように語形成力の高い接尾辞については、(19b) のように長単位の品詞を継承する[*5]ことはせず、(19a) のように短単位の品詞を保存する。そして、接尾辞には品詞ラベル PART を付与する。(18b) の長単位の品詞を継承する場合との整合性こそ失われるものの、英語の to 不定詞が名詞や副詞として振る舞う場合にも品詞が VERB のままであるなど、他の言語の表現形式とも概ね符合する。

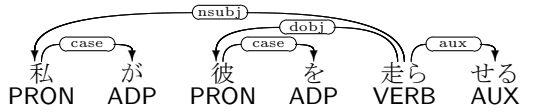
節の定義　英語での形容詞の限定的用法にはラベル amod が用いられるが、該当する日本語の現象は、用言の連体形を用いた体言の修飾である。従って、英語での関係節（acl）に相当する連体節との区別が明確でない。現状の案では、(12) と (13) のように、acl の特別な場合として、形容詞・形状詞が格を伴わずに名詞を修飾する場合に amod のラベルを付与している。

[*3] 活用語尾を含めた「勉強し」を VERB とすべきであるという意見もあるが、語の単位を UniDic の短単位で統一するという原則を崩さないことにした。

[*4] 但し、体言止めの扱いなどはまだ議論中である。

[*5] この場合、語彙項目と結びつかない品詞が無尽蔵に生成されてしまう。

態の扱い　現状の UD では、nsubjpass などのラベルに見られるように、受動態の格の交替を判別できるようにしているが、日本語の場合は使役などでも格の交替が起こりうる。例 (20) の場合、「せる」が内容語でないため、「私が」「彼を」の両方が「走ら」を修飾する構造となる[*6]。使役の格を厳密に区別するためには、別途ラベルの導入が必要となってしまう。

(20)


格のラベルと助詞の分類　主語や目的語が文法的に明らかな英語と異なり、日本語は格助詞などが格を示唆するが、助詞が省略されたり、係助詞・副助詞が用いられる場合など、何が nsubj や dobj になるかは自明ではないことがある。現状ではコーパスの中で付与された格の情報に頼ったラベルの定義をしたり、ADP と PART を助詞の種類のみで分類したりしているが、他の言語と同等の処理を目指すうえで、まだ検討の余地がある。

5　日本語 UD コーパスの構築

日本語の UD が付与された言語資源を構築する第一歩としては、既存の構文構造コーパスを変換することが有望であると考えられる。構文構造のコーパスには, BCCWJ[4] の短単位による単語の単位で係り受けを付与したコーパス [6] 、句構造のツリーバンク [8, 9, 11] などがある。

UD 自体は単語単位の係り受け構造であるが、節の単位を考慮した係り受けラベル (acl, advcl, csubj, ccomp など) があることや、並列構造のスコープを考える必要があるため、句構造のツリーバンクからの変換を考えることは、有効であると考えられる。また、格関係のラベル (nsubj, dobj など) を区別する必要があるため、変換元のコーパスには、述語項構造の情報が含まれていることが望ましい。格関係を含め内容語間の係り受け関係を捉える上では、格助詞や助動詞相当の機能表現の複合語（に対して、かもしれない など）をまとめて扱った方が都合が良いと考えられ、BCCWJ の長単位を導入することによりこれらを一つの機能語として扱い、変換することも有用であると考えられる。

現在は、日本語句構造ツリーバンク [8, 11] を利用し、形態素や句構造などのアノテーションを参照することで、日本語 UD へ変換するプログラムの開発を行っている。この変換プログラムおよび自動変換した日本語 UD コーパスは、今後ウェブ上で公開する。また、BCCWJ についても、形態素、係り受け、述語項構造等のアノテーションを利用し、自動変換を行うプログラムを開発する予定である。

一方で、UD への自動変換に必要な全ての情報を網羅した言語資源は、現時点ではほとんど存在しないと考えられる。既存コーパスのアノテーションを活用した自動変換をベースにしつつ、4 節で挙げたような課題に対応するため、新たなアノテーションスキームに関するさらなる議論が必要である。

[*6] 英語では使役の make などが別の本動詞となるため、この点が考慮されていないと思われる。

6　まとめと今後の展望

本稿では、Universal Dependencies 日本語版の定義の試案を示した。現段階では、UD の原則的定義に従いつつ日本語の既存の言語資源との対応関係を意識した案となっている。これまでは日本語の言語資源は日本独自の基準で研究されてきた面があり、既存の言語資源と UD の定義では対応関係が自明でない点が多い。本稿で示した案においても一貫性や変換可能性に課題が残っており、今後の議論により日本語の独自性と言語横断性を両立した体系の構築を目指す。

この活動により、言語横断的な研究や新たな応用への取り組みが加速されることが期待される。例えば、複数言語の UD コーパスを利用した構文解析器の構築、対訳コーパスに対して付与したアノテーションの比較、UD の言語横断性を利用した機械翻訳、などの応用が考えられる。一方、4 節で挙げた課題を見ても、UD 全体の定義が英語に特化していると思われる点もある。UD を日本語へローカライズすることにとどまらず、他の言語のコミュニティと連携しながら UD 全体の発展に向けて情報発信していくことが必要となろう。

参考文献

[1] Universal Dependencies contributors. Universal dependencies. https://universaldependencies.github.io/docs/, 2014.

[2] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pp. 4585–4592, 2014.

[3] Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. A proper approach to Japanese morphological analysis: Dictionary, model, and evaluation. In *Proceedings of LREC*, 2008.

[4] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48, pp. 345–371, 2014.

[5] Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal dependency annotation for multilingual parsing. In *ACL (2)*, pp. 92–97, 2013.

[6] Shinsuke Mori, Hideki Ogura, and Tetsuro Sasada. A Japanese word dependency corpus. In *Proceedings of LREC*, pp. 753–758, 2014.

[7] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of LREC*, 2012.

[8] Takaaki Tanaka and Masaaki Nagata. Constructing a practical constituent parser from a Japanese treebank with function labels. In *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages*.

[9] 吉本啓, 周振, 小菅智也, 大友瑠璃子, Alastair Butler. 日本語ツリーバンクのアノテーション方針. 第 19 回言語処理学会年次大会予稿集.

[10] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 小西光, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版 (上)(下). 人間文化研究機構国立国語研究所, 2011.

[11] 田中貴秋, 永田昌明, 松崎拓也, 宮尾祐介, 植松すみれ. 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. 第 20 回言語処理学会年次大会, pp. 737–740, 2014.