形態論情報付き日本語 Universal Dependencies

田口智大 [1]　宮川創 [2]
[1] University of Notre Dame　[2] 国立国語研究所 研究系
ctaguchi@nd.edu　so-miyagawa@ninjal.ac.jp

概要

本論文では，日本語の形態論を反映した，日本語 Universal Dependencies（UD）の新しいトークン化の基準と形態的素性のアノテーションを提案する．現在のバージョンの日本語 UD v2.11 では，文は形態素単位でトークン化されており，各トークンには形態的素性がほとんど付与されていない．しかしながら，日本語は形態的屈折を欠いた孤立語ではなく，形態的変化を持った膠着語である．この現状を考慮して，本稿では日本語 UD のトークン化の基準を見直し，形態素単位ではなく語単位のトークン化を提案する．そして，各トークンに含まれた形態論的情報を表すために，UD 共通の形態的素性のアノテーションを提案する．

1　はじめに

本稿では，日本語 Universal Dependencies（UD）ツリーバンクの形態論情報付与の必要性について論じ，新たなトークンの基準を提案する．UD は，2015 年に開始された多言語ツリーバンクのプロジェクトであり，あらゆる言語に対して共通のアノテーション基準で品詞タグ，形態論情報，統語的依存関係を記述することを理念としている [1, 2]．目下のところ最新のリリースであるバージョン 2.11 では，130 以上の言語のツリーバンクが公開されており，Universal Dependencies を用いた通言語的な言語処理研究も行われ始めている [3]．特に，日本語では，分かち書き・品詞ラベリング・依存構造ラベリングを同時に行う解析器である GiNZA が UD を用いて開発されている [4]．しかしながら，現在までの日本語 UD におけるトークンと形態論情報については，UD 共通の基準とは異なるアノテーション基準が採用されていることが報告されている [5, 6, 7]．第一に，日本語 UD のトークン化の基準は『現代日本語書き言葉均衡コーパス』[8] におけるトークン化の基準と同じであることが明記されており，概ね一般言語学における形態素の概念に等しい．日本語は，用言の形態的変化を有する膠着語であるが，現行の日本語 UD の方針では，あらゆる動詞接辞および形容詞接辞が独立したトークンとして扱われている．第二に，日本語 UD では，一部の例外を除いて，形態的素性が一切付与されていない．UD では，形態的素性の記載について，形態的屈折を持つ場合は必須となっていることから，統合的（synthetic）言語の性格を持つ日本語の形態論が反映されていない．そこで，本稿では，UD 共通の基準を鑑み，他言語での事例と照らし合わせながら，日本語の形態的変化を考慮した新たなトークン化基準と形態論情報の付与の試案を提示する．

2　UD と日本語 UD

2.1　UD の基本方針

UD では，語彙主義（lexicalism）に立脚し，各トークンは統語的語（syntactic word）であり，形態的屈折を含んだ表層形を単位として FORM に記載される．．そして，表層形から形態的屈折を除いた基本的な形（レンマ，辞書形，引用形などとも呼ばれる）が LEMMA として記載される．各トークンに含まれる形態論的な屈折は，形態的素性の集合として FEATS 欄に記載される．形態的素性を集合として扱うことで，形態素境界が不明瞭なケースや，一つの形態素が二つ以上の形態的素性を担う統合的（synthetic）なケースにも対応することができる設計となっている．例えば，膠着語であるフィンランド語の *työssä*（仕事中）という語は，名詞の *työ*（仕事）に単数内格（inessive singular）接尾辞の*-ssä* が付いたものであるが，UD ではまとめて一つのトークンとして扱い，LEMMA として työ，FEATS として Case=Ine|Number=Sing とアノテーションすることで，形態的屈折を表現している．そのほかに，UD では，各トークンの品詞タグ（UPOS），依存先のトークン ID（HEAD），依存関係タグ（DEPREL）が付与さ

れることが最低限の条件となっており，各文について形態統語論の基本的な情報を網羅する構造となっている．

2.2 日本語 UD の概要

現在のバージョン 2.11 の日本語 UD では，GSD，GSDLUW，PUD，PUDLUW，BCCWJ，BCCWJLUW，Modern，KTC の計八つのツリーバンクが公開されている．このうち，GSD，PUD，BCCWJ については，トークナイゼーションの基準として，『現代日本語書き言葉近郊コーパス』において語の基準とされている短単位（Short Unit Word）と長単位（Long Unit Word）のそれぞれを採用した二つの版を公開しており，後者はツリーバンク名に LUW が付加されている．短単位は「現代語において意味を持つ最小の単位」[1]と定義されており，言語学における形態素の定義と概ね一致する．短単位では，複合語を構成する複数の形態素をそれぞれ別々のトークンとして扱っている一方で，長単位では複合語を一つのトークンとして扱っている．これらの基準は，接辞（affix）や接語（clitic）を独立したトークンとして扱うという点で，日本語の学校文法（橋本文法）における「単語」の概念と類似している．しかしながら，動詞の活用接尾辞を別トークンとして扱う点においては，UD の他の膠着語におけるアノテーション基準とは異なっている．

さらに，これらの日本語 UD ツリーバンクでは，否定（Polarity=Neg）を除く一切の形態的素性が付与されておらず，空欄として残されている．日本語には，極性のほか，テンス・アスペクト，使役・受動などの態，敬語（honorifics）などの動詞の活用が存在するが，これらの形態素は別トークンとして扱われ，形態的素性としての Tense，Aspect，Voice，Polite などのキーは現状割り当てられていない．このように，日本語 UD では，トークン化の基準と形態論のアノテーションという二点において，UD の一般的方針とは異なる設計となっている．

2.3 他の膠着語での扱い

以上に述べた問題点について，解決案を提示する前に，UD の他の膠着語では動詞の活用や名詞の格標示がどのようにアノテーションされているのかを具体例とともに概観する．朝鮮語では，KAIST，GSD，PUD の三つのツリーバンクが公開されているが，これらのツリーバンクのアノテーション方針は統一されていない．いずれのツリーバンクにおいても，トークン化の基準は朝鮮語の正書法におけるスペースをもとにしており，動詞活用の接辞や名詞の格接語は主要部のトークンの一部として扱われている．形態的素性については，KAIST および GSD では一切記載されていない一方で，PUD では，表 1 に示す通り，動詞活用および名詞に付属する形態素が FEATS 列に表されている[2]．テュルク諸語では，いずれのツリーバンクでも動詞の活用や名詞の格変化は主要部の一部としてトークナイズされ，その情報は素性として FEATS に表されている．例として，トルコ語のアノテーション事例を表 2 に示す．最後に，部分的に膠着語的な性質を持つヒンディー語では，表 3 に示すように，格（後置詞）と動詞の屈折を形態的素性として明記しているが，格は後置詞として独立したトークンを構成している．

表 1 UD Korean-PUD における形態的素性アノテーションの例（文番号 n01010042 より引用）

FORM	LEMMA	UPOS	FEATS
때가	때	NOUN	Case=Nom|Polite=Form
있었다고	있었다	ADJ	Mood=Ind|Tense=Past|VerbForm=Fin
말했다	_	NOUN	_

表 2 UD Turkish-GB における形態的素性アノテーションの例（文番号 GK05-0003 より引用）．形態的素性は紙幅のため一部省略している．

FORM	LEMMA	UPOS	FEATS
nereye	nere	PRON	Case=Dat|Number=Sing|PronType=Int
gidiyorsun	git	VERB	Aspect=Prog|Evident=Fh|Mood=Ind|...

表 3 UD Hindi-PUD における形態的素性アノテーションの例（文番号 n01001011 より引用）．形態的素性は紙幅のため一部省略している．

FORM	LEMMA	UPOS	FEATS
पोस्त	_	NOUN	Animacy=Inan|Case=Acc|Gender=Fem|...
में	_	ADP	Case=Loc
लिखा	_	VERB	Aspect=Perf|Gender=Masc|Mood=Ind|...

3 試案

前節で見たように，UD では，動詞屈折や格変化を有する言語は，形態素単位ではなく統語的語の単位でトークン化を行い，形態的変化の情報は素性として FEATS に記載する方針である．実際に，日本語と同様に膠着語的性質を持つ言語の UD では，動詞の屈折的形態素は素性として表され，さらに言語やツリーバンクによっては，名詞の格標示も素性の一

1） https://clrd.ninjal.ac.jp/bccwj/morphology.html

2） 動詞 말했다が NOUN のタグを付与されていることや，時制などが表されてないことなど，今後の発展の余地の残るアノテーションである．

表 4 UD における形態的素性と日本語の格の対応

格	素性	例
主格（ガ格）	Case=Nom	猫**が**
属格（ノ格）	Case=Gen	猫**の**
与格（ニ格）	Case=Dat	猫**に**
対格（ヲ格）	Case=Acc	猫**を**
方向格（ヘ格）	Case=Lat	猫**へ**
奪格（カラ格）	Case=Abl	猫**から**
処格（デ格）	Case=Loc	猫**で**
共格（ト格）	Case=Com	猫**と**
比較格（ヨリ格）	Case=Cmp	猫**より**

部として扱われている．本章では，日本語 UD において，どのようなトークン化の基準が適切なのか，および形態論的素性はどのように表されるべきかを，いくつかの論点に分けて検討する．

3.1 格助詞は形態的格変化か

国文法では，「が」「を」「に」「へ」「で」「から」「より」「まで」「と」「の」の十個の助詞が格助詞として挙げられている．格は，述語とその項の関係を表す文法的な役割であり，膠着語では一般的に接辞や接語によって表現される．UD では，名詞の格は Case という素性キーによって表される．UD の枠組みの中では，日本語のこれらの格を素性として表すことは可能である．表 4 はこの対応をまとめたものである．

しかしながら，日本語の格が名詞の形態論の一部を構成する接辞（affix）であるか，より形態統語的に独立性の高い接語（clitic）であるかどうかは自明ではなく，議論の余地の残る問題である．そもそも，接辞と接語を明確に区別するような通言語的に一般的な定義を与えることは，不可能であることが指摘されている [9]．日本語を扱う統語論の研究では，格助詞を含む諸々の助詞や助動詞に関して，明示的に接辞と接語の区別を行なっていないものが多く，その区別は未だ定まっていない．宮岡（2002）は，日本語の格助詞が名詞の屈折形であるとは見做せないという理由から，格助詞は接語であるという解釈を提示している [10]．実際に，日本語の格助詞には例外的な形態的パラダイムがなく，名詞以外とも結合することから，接語としての性質を持っているといえる．一般的に，UD では接語を個別のトークンとして扱うため，これらの格助詞を接語として解釈する場合，格助詞は別のトークンとして扱われる必要がある．したがって，現段階では，格助詞の扱いとして以下の二通りが考えられる．一つ目は，朝鮮語 UD（表 1）やトルコ語 UD（表 2）のように，格助詞を名詞の派生接尾辞として扱い，形態的素性として表す方法である．二つ目は，ヒンディー語 UD（表 3）のように，格助詞を接語（あるいは後置詞）として扱い，別のトークンとする一方で，格助詞自体に格の形態的素性を付与する方法である．いずれの方法を採用するかは，慎重な議論と裏付けが必要であるが，少なくとも形態的素性の付与は可能である．

その他の助詞（副助詞，接続助詞，係助詞，並立助詞，終助詞，間投助詞，準体助詞）は述語との文法的関係を表すものではないため，格には含めない．特に，主題を表す助詞は，Korean-PUD では主格として扱われているが，日本語の「は」は述語との関係を指定するものではなく，情報構造上での役割を持つため，格に当たらない．しかし，形態統語的に主題を表す要素として，適切な素性が今後 UD 標準のガイドラインに整備される可能性はある．

3.2 用言の形態論

日本語における用言（動詞・形容詞・形容動詞）の形態的変化に関わる接辞は，国文法において助動詞と助詞に分類されるものに含まれる．国立国語研究所の分類では，文語的表現も含め，「う」「ごとき」「させる」「ざる」「しめる」「せる」「そうだ・そうです」「た・た」「だ」「たい」「たる」「です」「ない」「なる」「ぬ・ん」「ふうだ・ふうです」「べし」「まい」「ます」「みたいだ・みたいです」「よう」「ようだ・ようです」「らしい」「られる」「る」「れる」「ん・む」が助動詞として挙げられている [11]．これらのうち，連体形・終止形に接続する「ごとき」「そうだ・そうです[3)]」「ふうだ・ふうです」「みたいだ・みたいです」「ようだ・ようです」「らしい」は，動詞以外の品詞にも接続することができ，形態統語的自由度が高いため，現行の UD 通り助動詞 AUX として扱う．また，「だ」「たる」「です」「なる」は動詞に接続する要素ではないため，ここでの議論から除外する．他方で，連用形や仮定形に接続する助詞「ながら」「つつ」「たり」「たら」「て」「ば」は，形態素配列（morphotactics）の自由度が低く，アスペクトなどの文法的役割を担っているため，ここでは接尾辞として扱う．以上の形態素と UD における素性の対応を表 5 に示す．

3) 連用形に接続する「そう」は，将然相を文法的に表す接辞として，動詞の活用に含める．

表 5　日本語 UD における動詞活用の形態的素性のアノテーションの一案．VerbForm=Exem は現在の UD には登録されていない新規のカテゴリであることに注意されたい．

形式	素性	形成規則	例
終止・連体形非過去	Tense=Pres|VerbForm=Fin	終止形・連体形	書く
否定	Polarity=Neg	未然形 + *-nai*	書かない
受動	Voice=Pass	未然形 + *-(ra)re-*	書かれる
可能	Mood=Pot	下一段活用	書ける
使役	Voice=Cau	未然形 + *-(sa)se-*	書かせる
意志	Mood=Opt	未然形 + *-(y)ou*	書こう
否定意志	Mood=Opt|Polarity=Neg	未然／終止形 + *-mai*	書かまい／書くまい
丁寧	Polite=Form	連用形 + *-masu*	書きます
進行副動詞 (1)	Aspect=Prog|VerbForm=Conv	連用形 + *-nagara*	書きながら
進行副動詞 (2)	Aspect=Prog|VerbForm=Conv	連用形 + *-tutu*	書きつつ
将然相	Aspect=Prosp	連用形 + *-sou*	書きそう
例示	VerbForm=Exem	連用形 + *-tari*	書いたり
過去	Tense=Past	連用形 + *-ta*	書いた
過去条件	Mood=Cnd|Tense=Past	連用形 + *-tara*	書いたら
副動詞	VerbForm=Conv	連用形 + *-te*	書いて
希望	Mood=Des	連用形 + *-tai*	書きたい
不定詞	VerbForm=Inf	連用形 + *-Ø*	書き
条件	Mood=Cnd	仮定形 + *-ba*	書けば
命令	Mood=Imp	命令形	書け

用言に形態的素性を付与するにあたって，以下の論点が生じる．第一に，日本語では形容詞や形容動詞も屈折するため，提案された形態的素性は用言全体に適用されるが，現行の UD の枠組みでは，動詞の VerbForm にあたる素性が形容詞には存在しない．解決策としては，朝鮮語の Korean-PUD で実験的に用いられている Form という素性キーを適用するか，VerbForm を用言全体に一般化するかのどちらかが考えられる．二点目は，連体形の形態的素性の扱いである．現代日本語では，動詞・形容詞の変化において終止形と連体形の区別が失われているが，形容動詞では連体形の区別が残っている．そのため，終止形と連体形をどちらも VerbForm=Fin として扱い，形容動詞の連体形のみに別の素性を与えるか，終止形のみを VerbForm=Fin として扱い，連体形全体に別の素性を与えるか，のどちらかの選択肢に統一する必要がある．なお，連体形と終止形の区別を持つ朝鮮語の UD では，連体形には Form=Adn という素性が与えられている．三点目は，助動詞「た・だ」の形態的素性が過去時制 Tense=Past なのか，完了アスペクト Aspect=Perf なのかという点である．この点に関する見解の一致は見られていないが，日本語 UD の一貫性のために，どちらかに統一する必要がある．

3.3　提案の実装に向けて

ここまで，日本語 UD における名詞と動詞の形態論を考慮した，新たなトークンの基準と形態論情報のアノテーションを提案した．現行の日本語 UD でのトークン化と本稿の提案を比較したものを付録 A に示す．本稿で形態的屈折および派生として扱った形態素は，国文法の品詞カテゴリである助詞・助動詞の一部に対応している．したがって，現行の日本語 UD では，言語依存の品詞タグを任意に記載することのできる XPOS 列に国文法の品詞カテゴリが明記されていることから，これを参照してトークンの融合および形態論情報の追加を自動で処理することが可能であろう．

4　終わりに

本稿では，UD の共通ガイドラインに立脚し，日本語 UD に形態的素性の情報を組み込むことの必要性を主張し，その具体的なアノテーションの試案を提示した．このように，UD 全体との統一性を高めることで，UD を用いた通言語的な研究や応用が可能となる．また，今後の開発が期待される，沖縄語，国頭語，奄美語，宮古語，八重山語，与那国語，八丈語などの，日本語以外の日琉諸語の UD ツリーバンクの開発のための土台となることが期待される．

謝辞

本論文を作成するにあたって，京都大学の村脇有吾氏およびテュービンゲン大学の Çağrı Çöltekin 氏より建設的な助言をいただき，ここに感謝を表します．
This material is based upon work supported by the National Science Foundation under Grant No. BCS-2109709. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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, 07 2021.

[2] Daniel Zeman, et al. Universal dependencies 2.11, 2022. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[3] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[4] 松田寛, 大村舞, 浅原正幸. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会 第 25 回年次大会 発表論文集 (2019 年 3 月), 2019.

[5] Gregory Pringle. Thoughts on the Universal Dependencies proposal for Japanese: The problem of the word as a linguistic unit. http://www.cjvlang.com/Spicks/udjapanese.html, Date accessed: November 10, 2022, 2016.

[6] Yugo Murawaki. On the definition of japanese word, 2019.

[7] Çağrı Çöltekin and Taraka Rama. What do complexity measures measure? correlating and validating corpus-based measures of morphological complexity. **Linguistics Vanguard**, 2022.

[8] 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. **Lang. Resour. Eval.**, Vol. 48, No. 2, p. 345–371, jun 2014.

[9] Martin Haspelmath. Defining vs. diagnosing linguistic categories: A case study of clitic phenomena, 2015.

[10] 宮岡伯人. 語とはなにか：エスキモー語から日本語をみる. 三省堂, 2002.

[11] 国立国語研究所. 現代語の助詞・助動詞：用法と実例. 国立国語研究所報告, Vol. 3, , 1951.

[12] 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 (UDW 2018)**, pp. 117–125, Brussels, Belgium, November 2018. Association for Computational Linguistics.

A　トークン化基準の比較

表 6　短単位、長単位、ならびに文節単位でのトークンと、形態的素性を加味した試案に基づいたトークンの比較。例文は大村ほか（2018）[12] より引用した。

短単位	魚	フライ	を	食べ	た	か	も	しれ	ない	ペルシャ	猫
	NOUN	NOUN	ADP	VERB	AUX	PART	ADP	VERB	AUX	NOUN	NOUN
長単位	魚フライ		を	食べ	た	かもしれない				ペルシャ猫	
	NOUN		ADP	VERB	AUX	AUX				NOUN	
文節	魚フライを			食べたかもしれない						ペルシャ猫	
短単位＋用言活用	魚	フライ	を	食べた		か	も	しれない		ペルシャ	猫
	NOUN	NOUN	ADP	VERB		PART	ADP	VERB		NOUN	NOUN
	_	_	Case=Acc	Tense=Past VerbForm=Fin		_	_	Polarity=Neg Tense=Pres VerbForm=Fin		_	_
短単位＋用言活用＋名詞格	魚	フライを		食べた		か	も	しれない		ペルシャ	猫
	NOUN	NOUN		VERB		PART	ADP	VERB		NOUN	NOUN
	_	Case=Acc		Tense=Past VerbForm=Fin		_	_	Polarity=Neg Tense=Pres VerbForm=Fin		_	_
長単位＋用言活用	魚フライ		を	食べた		かもしれない				ペルシャ猫	
	NOUN		ADP	VERB		AUX				NOUN	
	_		Case=Acc	Tense=Past VerbForm=Fin		Tense=Pres VerbForm=Fin				_	
長単位＋用言活用＋名詞格	魚フライを			食べた		かもしれない				ペルシャ猫	
	NOUN			VERB		AUX				NOUN	
	Case=Acc			Tense=Past VerbForm=Fin		Tense=Pres VerbForm=Fin				_	