沖縄語のUniversal Dependenciesツリーバンクコーパスの構築

沖縄語の Universal Dependencies ツリーバンクコーパスの構築

宮川 創 [1]　金山 博 [2]　田口 智大 [3]　當山 奈那 [4]
[1] 国立国語研究所 研究系　[2] 日本アイ・ビー・エム株式会社 東京基礎研究所
[3]University of Notre Dame　[4] 琉球大学 人文社会学部
so-miyagawa@ninjal.ac.jp　hkana@jp.ibm.com
ctaguchi@nd.edu　tohyama@hs.u-ryukyu.ac.jp

概要

沖縄語は日琉語族のうちの琉球諸語に属する言語で，消滅の危機に瀕しているが，ローリソース言語の中では比較的テキスト資料は多い．そこで，国立国語研究所の語彙資源『沖縄語辞典』および，談話テキスト資源『方言談話資料』をもとに，沖縄語の Universal Dependencies ツリーバンクを構築している．本稿ではその過程で検討した，表記法，語分割，品詞や係り受けタグの選択について報告する．

1　はじめに

沖縄語は，日琉諸語のうちの琉球諸語の中の北琉球諸語に属し，沖縄島中南部を中心に使用されてきた言語である．しかし，その使用は急速に衰退しており，UNESCO によって，消滅の危機にある言語と認定された [1]．それにもかかわらず，沖縄語の中でも，琉球国時代の政治の中心地であった首里の言語は，日琉諸語のなかでは，日本語のつぎに古い書記記録をもち，書かれたテキストは比較的多い．Universal Dependencies [2] (以下，UD) は，世界の諸言語の係り受け構造を記述するオープンコミュニティのプロジェクトである．世界中の消滅危機言語や少数言語や古典言語のツリーバンク開発も盛んであり，UD の最新版 ver. 2.11 は，世界の 138 の言語のツリーバンクを有している．消滅の危機に瀕する言語ならびに少数言語としては，ブラジルの少数民族の諸言語や，オーストラリアの先住民族の諸言語などが含まれている．しかし，日本語派の諸語と琉球語派の諸語からなる日琉語族の諸言語で，UD ツリーバンクを有しているのは現代日本語と古典日本語のみである．そこで，本稿は，ローリソース言語かつ消滅危機言語の言語資源開発の一環としての沖縄語の UD ツリーバンクの設計について，現在までの作業の進展も交えながら論じる．語彙情報は国立国語研究所の『沖縄語辞典』 [3] に基づき，コーパスのテキストは同研究所の『方言談話資料』の第 6・8・9・10 巻 [4, 5, 6, 7] に収録されている沖縄語首里方言の談話テキスト資料を用いている．いずれの資料も，国立国語研究所のリポジトリ上で CC BY 4.0 ライセンスで配布されている．琉球諸語のタグ付きコーパスは現在存在せず，本稿が述べる沖縄語 UD ツリーバンクコーパスが琉球諸語で初の言語学的タグ付きコーパスとなる見込みである．

2　沖縄語の言語学的特徴

沖縄語は日琉語族に属しており，日本語とは親縁関係にあり，統語論・形態論的特徴もその多くを日本語と共有している．例えば，次の図 1 は，「私は (わんねー) 彼 (あり) に (んかい) 酒 (さき) を 与えた (くぃたん) 」という意味の例文であるが，基本語順は SOV であり，格助詞 (んかい) が名詞の後に来ている．この図では，例文の上には単語の依存先 (HEAD) を表す矢印と 5 節で述べるそれぞれの依存/係り受け関係 (DEPREL) タグが，例文の下には単語毎に 4 節で述べる Universal Part-of-Speech (UPOS) とグロス (語釈) が示されている．

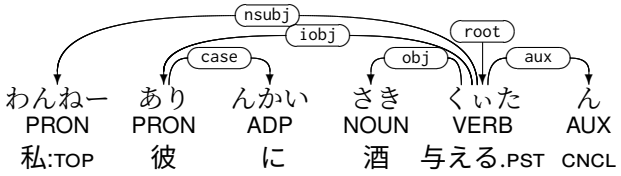

図 1　「私は彼に酒を与えた」を意味する例文[1)]

形態素同士の融合が，沖縄語では日本語よりも頻繁に見られる．例えば，図 1 の ‘わんねー’ (代名詞 ‘わー’「私」＋主題の副助詞 ‘や’「は」) のような代名詞/名詞と副助詞の融合である．しかし，これらの問題に踏み込む前に表記法の問題を解決する必要がある．沖縄語は正書法が確立しておらず，漢字仮

1)　TOP は主題，PST は過去，CNCL は終止形接尾辞を表す．

名混じり表記，仮名表記，アルファベット表記など様々な表記が存在するためである．

3 沖縄語の表記の問題

漢字仮名混じり表記は，琉球王国時代から，一般的によく用いられているが，様々な変種が存在する．特に漢字表記は，意味や個人的な解釈に基づくことが多く，揺れが多く見られる．沖縄語は，日本語と漢字の読みが異なることが多いことから，近年では，仮名のみの表記をすることも多くなっている．ルビなしの漢字のみで正確な読みを全て推測するのが一般的に困難であることから，本研究ではまず仮名表記のみでの沖縄語の UD 化を試みる．沖縄語の仮名表記には以下の種類がある．

1. 伝統的仮名表記：『おもろさうし』[8]，組踊[9]，琉歌，ベッテルハイム訳聖書など．キヤウンと書いて，[tɕoːn]「来ている」[2] と発音されるなど，発音と綴りとの乖離が大きい．
2. 現代的な仮名表記：沖縄県のしまくとぅば正書法検討委員会による「しまくとぅば普及推進行動計画 (しまくとぅば県民運動) 」の公式表記や，西岡敏・仲原穣著，伊狩典子・中島由美協力『沖縄語の入門』[11] の表記，琉球諸語統一的表記法 [12]，船津好明の沖縄文字表記 [13]，沖縄語普及協議会の表記などがあるが，声門閉鎖音の対立などの表記の仕方が異なる．ひらがな/カタカナを用いるかでもばらつきがある．
   (a)『沖縄語の入門』[11] の表記：日本語しか知らない読者に対しても分かりやすい仮名表記になっている．例：ウゥトゥ /wutu/「夫」vs. ウトゥ /ʔutu/「音」，ワー /waa/「私」vs. ッワー /ʔwaa/「豚」．
   (b) 沖縄県のしまくとぅば正書法検討委員会 [10]：ほぼ西岡・仲原と同じだが，一部 Unicode にはない上付き文字を使用している．例：ᵘウトゥ または ’ウトゥ /wutu/「夫」vs. ウトゥ /ʔutu/「音」．
   (c) 小川晋史他『琉球のことばの書き方 琉球諸語統一的表記法』[12] の表記法：「しま書体」フォントの使用が前提で，Unicode 未登録の文字を多数使用している．
   (d) 船津好明 [13] による現代的なかな表記＋独自の「沖縄文字」(ひらがなの合字)：純粋なモーラベースだが Unicode にない．

沖縄語 UD で使用するテキストデータのために現在作成している『沖縄語辞典』[3] の多言語翻訳版 TEI XML ファイルを整備するにあたって，以下の表記法に統一することを決定した．この辞典の基である島袋盛敏の稿本は，カタカナ表記であったが，国立国語研究所の地方言語研究室により言語学的なアルファベット表記に置き換えられた．しかし，この表記は専門的であるため，一般人には容易ではない．そこで，一般に用いられている表記法について，調査を行い，特に表記が分かれる発音の違いを整理した．そして，様々な仮名表記を相互に変換可能にした上で「しまくとぅば県民運動」の表記に準じながら，Unicode では表せない上付き文字を下付き文字に置き換え，コンピュータでの使用の容易さを第一に考えたかな表記の標準化を行った．なお，親しみやすさと入力しやすさを考慮して仮名はカタカナではなくひらがなを用いた．この標準化されたかな表記に『沖縄語辞典』[3] の表記を自動的に変換し，それを基に UD を構築している．

4 沖縄語の品詞と UPOS

UD では，Google の Universal Part-of-Speech[14] を継承した，Universal Part-of-Speech (UPOS; 付録 A の表 3 参照) が，品詞タグとして用いられる．UPOS は，UD の標準形式の CoNLL-U 形式のテーブルの第 4 番目の列に書かれる．UPOS は，すべての言語に適用できるよう，比較的少なく，普遍的な品詞に絞ったものである．日本語の助詞は，格助詞・副助詞は ADP，終助詞は PART，接続助詞は SCONJ など，用途によって UPOS が分かれる．さらに，UD の語の表現は原文の全体をカバーするため，句読点などのパラ言語的記号も PUNCT として UPOS を与えられる．現在，日本語には 8 つの UD ツリーバンクがある．沖縄語は，日本語と同じく日琉諸語であり，品詞のカテゴリーが日本語と大きく異なることがないため，日本語の品詞と UPOS の対応を使えるところがほとんどである．しかし，沖縄語は，より言語学的な品詞分類が成されることが多く，文法家により異なる品詞体系がなされるため，注意が必要である．沖縄語は，言語類型論的に見れば，日本語共通語に比べ，形態素同士の融合 (fusion) の度合いが高い．例えば，‘ちゅーん’「来る」の肯定現在進行形は ‘ちょーん’「来ている」であるが，否定現在形は ‘くーん’「来ない」，肯定過去形は ‘ちゃん’「来

2) しまくとぅば正書法検討委員会の表記法 [10] では ‘ちょーん’.

た」である．接語・接辞の区分けが難しい問題もあり，特に名詞および代名詞に接続する助詞と動詞に接続する接尾辞・接語で頻繁にみられる．

国立国語研究所『沖縄語辞典』(1963) の名詞，自動詞，他動詞，形容詞，連詞，副詞，連体詞，接続詞，感動詞，助詞の 10 の品詞は，沖縄語 UD の UPOS に以下のように対応づけられる．

- NOUN 普通名詞：‘ぼーづぃ’「坊主」など，および，助数詞：‘ちゅ’ (‘ぴちゅ’「一人」‘たちゅ’「二人」などの) など．
- PROPN 固有名詞：‘なーふぁ’「那覇」など．
- VERB 動詞 (辞書形から末尾の ‘ん’ を除去したもの; 5.1 節で詳述する)：‘ゆぬ’ (‘ゆぬん’「読む」から) など．
- ADJ 形容詞 (辞書形から末尾の ‘ん’ を除去したもの) * ‘まぎさ’「大きい」など，および，DET を除く連体詞：‘いるんな’「色んな」など．
- ADV 副詞：‘しかしか’「イライラと」，‘しぷーとぅ’「びっしょり濡れて」など．
- INTJ 間投詞：‘うね’「おや」，‘あひゃんがれー’「やけくそなときに言う言葉」など．
- PRON 代名詞：‘わー’「私」，‘わったい’「私たち二人 (双数)」，‘うんじゅ’「貴方」など．
- NUM 数詞：漢語数詞：‘さん’「三」など，および，助数詞を伴う非借用語数詞 (助数詞は名詞として分離; 5.2 節を参照)：‘た’「二」など．
- AUX 動詞の屈折および派生接尾辞/接中辞 (音節文字を使用する関係上，子音語幹につく挿入母音は動詞側に含まれる)：‘ん’ (叙述・終止の接尾辞)，‘りー’ (受身の接中辞)，および，連詞：‘や’「だ」など，付録 E 参照．
- CCONJ 等位接続詞：‘また’「また」，‘とぅ’「と」など．
- SCONJ 準体助詞 ‘し’．[3)]
- DET 連体詞の一部：‘くぬ’「この」‘あぬ’「あの」‘あふぃな’「あんな」など．
- ADP 格助詞：‘が’ (主格)，‘ぬ’ (主格・属格)，‘んかい’ (与格)，副助詞：‘ん’ (添加)，‘や’ (主題)，係助詞：‘どぅ’ (焦点) など．
- PART 様々な終助詞：疑問の終助詞 ‘が’ と ‘み’，確認などの終助詞 ‘やー’ など．
- PUNCT 句点，読点，括弧開，括弧閉など．
- SYM 記号・補助記号のうち PUNCT 以外のもの．
- X フィラー：‘あぬー’「あのー」など．

3) 例：‘ゆぬし’「読むの」．日本語 UD の場合は動詞のテ形 ‘て’ (CCONJ となるものを除く) も SCONJ に入れられる場合が多いが，沖縄語では ‘て’ に相当するものは動詞語幹と融合することが多いため，単独の形態素としての抽出は困難である．

5 語分割と係り受けの問題点

依存/係り受け関係 (DEPREL) タグに関しては，基本的に統語論が近い現代日本語の UD のもの ([15] や [16] など) を踏襲しているため，ここでは個々の事例を詳述することは割愛する．しかしながら，沖縄語特有の問題，特に，語幹交替と融合語の問題などは依存関係タグや UPOS でどう扱うかを議論する必要がある．以下で，それらの問題点を挙げ，その解決策を考察する．

5.1 動詞と形容詞の語幹と活用接辞

日本語よりも屈折度が高く語幹が交替する沖縄語の動詞は，日本語 UD にはない対応をしなければならない．沖縄語の動詞の活用形には，I 型から XIV 型までの 14 の型があり ([3] の p.59)，さらに，I 型には下位分類が 4 つ，XIV 型には下位分類が 2 つある．特に注意すべきは第 III〜XIV 型動詞であり，これらは語幹の末尾子音が変化する．例えば，以下の ‘ゆぬん’ /junuɴ/「読む」の場合は，基本語幹 jum- の m が n や d に変化する．

- jum-語幹：‘ゆまん’/jumaɴ/「読まない」，‘ゆみぶしゃん’/jumibusjaɴ/「読みたい」，‘ゆめー’/jumee/「読めよ」，‘ゆむな’/jumuna/「読むな」など．
- jun-語幹：‘ゆぬん’/junuɴ/「読む」，‘ゆなびーん’/junabiiɴ/「読みます」，‘ゆぬら’/junura/「読むだろうか」など．
- jud-語幹：‘ゆだん’/judan/「読んだ」，‘ゆどーちゅん’/judoocuɴ/「読んでおく」，‘ゆでぃー’/judii/「読んだか」，‘ゆでぃ’/judi/「読んで」など．

UD_Japanese-BCCWJ[17] などでは，「読む」なら「読ま」「読み」「読む」「読め」などを異形態として登録しており，その後に来る丁寧，否定，アスペクト，ムードなどの動詞接尾辞は，AUX として登録している．沖縄語でも，動詞語幹＋接尾辞の組合わせを全ての動詞で登録するよりは，音節文字ベースでの動詞の語幹を全て登録して，接尾辞の残りは，助動詞として登録する方が登録数が少なくて済む．同方法を沖縄語で行う場合，付録 B の表 4 で示している通り，‘ゆぬん’「読む」には，‘ゆぬ’，‘ゆな’，‘ゆま’，‘ゆみ’，‘ゆだ’，‘ゆどー’，‘ゆでぃー’，‘ゆ

でぃ’，‘ゆでー’ の 9 種類の異形態 (あるいは異トークン) の登録が必要である[4]．更に沖縄語には約 10 種類の不規則動詞が存在し，日本語共通語には見られない補充法[5]がいくつか見られる．これらの出現しうる不規則な表層形を全て登録する．

沖縄語では形容詞も活用する．14 の型がある動詞と比べ，形容詞は 2 つの型しかなく規則的であるが，動詞と同様，音節ベースの各々の活用形を全て登録する．語幹を除く活用接辞 (接尾辞および接中辞) はそれぞれ AUX にし，語幹 (VERB) と活用接辞 (AUX) の係り受け関係は aux にする[6]．

5.2 数詞と助数詞の扱い

UD_Japanese-BCCWJ などでは，「一人」など非借用語数詞＋助数詞の組み合わせを NOUN に入れているが，これは，国語研の短単位がこれらを名詞として扱っているためである．沖縄語では，「一つ」が ‘てぃーち’ であるのに対して，「一人」が ‘ちゅい’ であるように，数詞の「1」にあたる部分が ‘てぃー’ と ‘ちゅ’ で異なる形式が用いられることもあるものの，多くの数詞＋助数詞は，数詞の部分に促音や長音が挿入されるだけの単純な異形態である場合が多い．そこで，数詞＋助数詞に関しても，不規則な形式の数詞が出るものも含めて[7]，非借用語数詞の標準形のレンマを定め，促音や長音が挿入された形を数詞の異形態とし，助数詞を名詞として，数詞と助数詞の依存関係を clf (classifier) とする[8]．

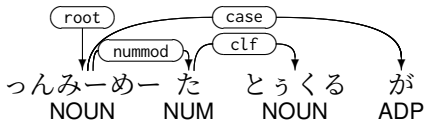

図 2 「お姉様御両名が」を意味する例文

5.3 名詞/代名詞/助詞と副助詞の融合

名詞，あるいは，代名詞がその後に続く副助詞と融合するのも沖縄語の特徴である．特に一部の名詞・代名詞・助詞は，副助詞の ‘や’ (日本語の主題の副助詞「は」に相当) が後続する場合，音融合を起こす．一部の代名詞は副助詞 ‘ん’ (日本語の副助詞「も」に相当) が後続するときに特別な形を取る．

表 1 代名詞/名詞と副助詞の融合の例

融合語	代名詞/名詞	副助詞
‘わんねー’「私は」	‘わー’「私」	‘や’「は」
‘しぐとー’「仕事は」	‘しぐとぅ’「仕事」	‘や’「は」
‘わんにん’「私も」	‘わー’「私」	‘ん’「も」

これらの母音融合は，語幹の子音や母音が変化する動詞の「テ形」と比べ，比較的単純である．よって，フランス語の du (前置詞 de と定冠詞 le の融合) のように，複数語トークン (Multi-Word Token; MWT [18]) を用いて，CoNNL-U の省略形である表 2 の ‘んでー’ (接続助詞 ‘んでぃ’ と副助詞 ‘や’ の融合) のように表す．

表 2 接続助詞と副助詞の融合語を表した例

ID	表層	レンマ	UPOS	依存関係	
1	うふぁなし	うふぁなし	NOUN	4	obj
2-3	**んでー**	_	_	_	_
2	んでぃ	んでぃ	SCONJ	1	mark
3	や	や	ADP	1	case
4	さ	す	VERB	0	root
5	な	な	AUX	4	aux
6	やー	やー	PART	4	discourse

6 おわりに

以上，『沖縄語辞典』[3] の語彙情報と『方言談話資料』[4, 5, 6, 7] の談話テキスト資料を用いて現在開発中の沖縄語 UD ツリーバンクコーパスについて，UPOS などのタグの設定と沖縄語特有の問題点について述べた．現在，本稿の方法に従って，ツリーバンク構築を行っており，GitHub 上に UD リポジトリ[9]の一つとして UD_Okinawan を作成し，公開していく予定である．今後解決していくべき問題としては，補充法を含む不規則に変化する不規則動詞をトークン化していく際に，それぞれの形態がどのような文法機能を有しているか明白にする必要があることである．この際，UD の Universal Features を用いて，その形態のテンス，アスペクト，ムード，極性 (肯定/否定) など文法機能を示すことで解決できる[10]．このように沖縄語の UD ツリーバンクを構築し，日本語との異同，さらに世界の諸言語との違いを量的な分析で明示できるように公開する．こうして，本沖縄語 UD ツリーバンクを琉球諸語の UD ツリーバンクの嚆矢とする．

4) 更に，Universal Feature でその文法機能を記すべきである．付録 F の表 5 を参照.
5) 英語の go に対する went など元々別の語彙がパラダイムに組み込まれた形.
6) 具体例は付録 E の図 4 および 5 を見よ.
7) ‘てぃーち’ の ‘てぃー’ と ‘ぴちゅ’ の ‘ぴ’ をレンマ化し{ぴ}とした．‘ぴ’ を選んだのはこの形式が出現することが多いからである.
8) 他の例はを付録 D の図 3 を見よ.
9) https://github.com/UniversalDependencies (閲覧日 2022-01-11)
10) 付録 F の表 5 を参照.

謝辞

本研究は JSPS 科研費 基盤研究 (B) JP19H01265「多言語による日本語学用語辞典および日琉諸語の用例に対するグロス規範の作成」, 挑戦的研究 (萌芽) JP21K18376 「フィールドデータのアーカイブに向けた問題点の整理と解決策」, 人間文化研究機構 共同研究プロジェクト (2022 年度〜) 共創先導プロジェクト 共創促進研究「学術知デジタルライブラリの構築」, 国立国語研究所 共同研究プロジェクト (2022 年度〜) 機関拠点型基幹研究「開かれた言語資源による日本語の実証的・応用的研究」基幹型「消滅危機言語の保存研究」, 国立国語研究所 共同研究プロジェクト (2022 年度〜) 広領域連携型基幹研究「異分野融合による総合書物学の拡張的研究」国語研ユニット「古辞書類に基づく語彙資源の拡張と語彙・表記の史的変遷」, 国立国語研究所 共同研究プロジェクト (2022 年度〜) 機関拠点型基幹研究 「開かれた言語資源による日本語の実証的・応用的研究」基幹型「多様な語彙資源を統合した研究活用基盤の共創 – 統括班」, 東京外国語大学アジア・アフリカ言語文化研究所共同利用・共同研究課題「理論言語学と言語類型論と計量言語学の対話にもとづく言語変化・変異メカニズムの探求」, 大学共同利用機関情報・システム研究機構データサイエンス共同利用基盤施設 2022 年度公募型共同研究 ROIS-DS-JOINT「日琉諸語の言語類型アトラス LAJaR の開発と分析」の助成を受けたものです.

参考文献

[1] 新垣友子. 琉球における言語研究と課題. 沖縄大学地域研究所（編）, 琉球諸語の復興, 沖縄大学地域研究所叢書, pp. 13–29. 芙蓉書房出版, 東京, 2013.

[2] 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 12th Language Resources and Evaluation Conference (LREC 2020)**, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[3] 国立国語研究所. 沖繩語辞典, 国立国語研究所資料集, 第 5 巻. 財務省印刷局, 東京, 1963.

[4] 国立国語研究所. 方言談話資料 (6) – 鳥取・愛媛・宮崎・沖縄 –, 国立国語研究所資料集, 第 10 巻. 国立国語研究所, 1978.

[5] 国立国語研究所. 方言談話資料 (8) – 老年層と若年層との会話 – 群馬・奈良・鳥取・島根・愛媛・高知・長崎・沖縄, 国立国語研究所資料集, 第 10 巻. 国立国語研究所, 1985.

[6] 国立国語研究所. 方言談話資料 (9) – 場面設定の対話 – 青森・群馬・千葉・新潟・長野・静岡・愛知・福井・奈良・鳥取・島根・愛媛・高知・長崎・沖縄, 国立国語研究所資料集, 第 10 巻. 国立国語研究所, 1986.

[7] 国立国語研究所. 方言談話資料 (10) – 場面設定の対話 その 2 – 青森・群馬・千葉・新潟・長野・静岡・愛知・福井・奈良・鳥取・島根・愛媛・高知・長崎・沖縄, 国立国語研究所資料集, 第 10 巻. 国立国語研究所, 1987.

[8] 波照間永吉. おもろさうし. 琉球文学大系 / 名桜大学『琉球文学大系』編集刊行委員会編纂, No. 1. ゆまに書房, 東京, 2022.

[9] 当間一郎. 組踊写本の研究. 第一書房, 東京, 1999.

[10] 沖縄県文化観光スポーツ部. 沖縄県における「しまくとぅば」の表記について, 2022. https://www.pref.okinawa.lg.jp/site/bunka-sports/bunka/shinko/documents/02_shimakutubahyouki.pdf (2023-01-11 閲覧).

[11] 西岡敏, 仲原穣, 伊狩典子, 中島由美. 沖縄語の入門 : たのしいウチナーグチ. 白水社, 改訂版, 2006.

[12] 小川晋史, 重野裕美, 新永悠人, 又吉里美, 當山奈那, Thomas Pellard, 林由華, 下地理則, 下地賀代子, 中川奈津子, Christopher Davis, 麻生玲子, 山田真寛. 琉球のことばの書き方 : 琉球諸語統一的表記法. くろしお出版, 2015.

[13] 船津好明, 中松竹雄. 沖縄口 (うちなーぐち) さびら : 沖縄語を話しましょう. 琉球新報社, 2010.

[14] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. **arXiv preprint arXiv:1104.2086**, 2011.

[15] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇有吾, 松本裕治. Universal Dependencies 日本語コーパス. 自然言語処理, Vol. 26, No. 1, pp. 3–36, 03 2019.

[16] 金山博, 宮尾祐介, 浅原正幸, 田中貴秋, 植松すみれ. 日本語 Universal Dependencies の試案. 言語処理学会第 21 回年次大会 発表論文集, pp. 505–508, 2015.

[17] 大村舞, 浅原正幸. UD Japanese-BCCWJ の構築と分析. Vol. 3, pp. 161–175, 2018.

[18] 金山博, 大湖卓也. UD_English-EWT とのつきあい方 . 言語処理学会 第 28 回年次大会 発表論文集, pp. 2013–2017, 2022.

付録 (Appendix)

A. UPOS 一覧

表 3　UD ver. 2 の Universal POS tags

開かれたクラス	閉じたクラス	他
ADJ 形容詞	ADP 接置詞	PUNCT 句読点
ADV 副詞	AUX 助動詞	SYM 記号
INTJ 間投詞	CCONJ 並列接続詞	X その他
NOUN 名詞	DET 限定詞	
PROPN 固有名詞	NUM 数詞	
VERB 動詞	PART 不変化詞[11]	
	PRON 代名詞	
	SCONJ 従属接続詞	

B. 動詞活用形トークナイズの例

表 4　XII 型動詞 ‘ゆぬん’「読む」のいくつかの例

動詞形と訳	沖縄語 UD での分析
‘ゆぬん’「読む」	‘ゆ ぬ’ (VERB; {ゆ ぬ}[12]) ＋ ‘ん’ (AUX)
‘ゆなびーん’「読みます」	‘ゆ な’ (VERB; {ゆ ぬ}) ＋ ‘びー’ (AUX) ＋ ‘ん’ (AUX)
‘ゆまん’「読まない」	‘ゆ ま’ (VERB; {ゆぬ}) ＋ ‘ん’ (AUX)
‘ゆみぶしゃん’「読みたい」	‘ゆ み’ (VERB; {ゆ ぬ}) ＋ ぶしゃ (AUX) ＋ ‘ん’ (AUX)
‘ゆめー’「読めよ」	‘ゆめー’ (VERB; {ゆぬ})
‘ゆむな’「読むな」	‘ゆ む’ (VERB; {ゆ ぬ}) +‘な’ (AUX)
‘ゆだん’「読んだ」	‘ゆだ’ (VERB; {ゆぬ}) ＋ ‘ん’ (AUX)
‘ゆどーちゅん’「読んでおく」	‘ゆ ど ー’ (VERB; {ゆ ぬ}) ＋ ‘ちゅ’ (AUX) ＋ ‘ん’ (AUX)
‘ゆでぃー’「読んだか」	‘ゆでぃー’ (VERB; {ゆぬ})
‘ゆでーん’「読んである」	‘ゆ で ー’ (VERB; {ゆ ぬ}) ＋ ‘ん’ (AUX)
‘ゆでぃ’「読んで」	‘ゆでぃ’ (VERB; {ゆぬ})

C. AUX の実例

- 動詞接尾辞 (辞書形から末尾の ‘ん’ を除いた形)
  - ‘ん’ (叙述・終止の接尾辞)
  - ‘る’ (連体の接尾辞)
  - ‘ぎゆ’/‘あぎゆ’ (進行の接中辞)
  - ‘びー’/‘いびー’ (丁寧の接中辞)
  - ‘みしぇー’/‘んしぇー’ (尊敬の接中辞)
  - ‘ぶしゃ’ (欲求の接中辞)
  - ‘しゅ’/‘あしゅ’ (使役の接中辞)
  - ‘ゆーしゅ’ (可能の接中辞)
  - ‘みゆ’/‘いみゆ’ (使役の接中辞)
  - ‘りゆ’/‘ありゆ’ (受身の接中辞)
  - ‘ぎさ’/‘いぎさ’ (推定・伝聞の接中辞) ,
  - ‘らー’/‘いらー’ (仮定の接尾辞) など.
- 連詞　(辞書形から末尾の ‘ん’ を除いた形式)[13]
  - ‘や’「だ」　(否定形 ‘あら’)
  - ‘でーび’「です」
  - ‘ぐとー’「のようだ」など.

D. 数詞＋助数詞の書き方

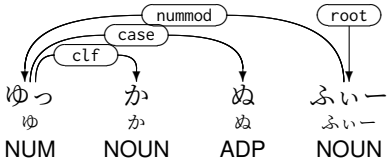

図 3　「四日の日」を意味する例文

E. 動詞の接尾辞の助動詞としての扱い

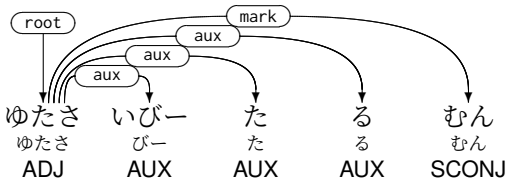

図 4　「宜しかったでしょうに」を意味する例文

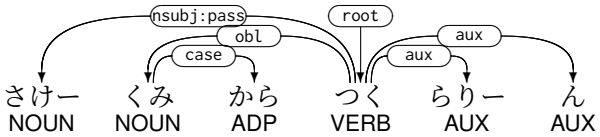

図 5　「酒は米から作られる」を意味する例文

F. 不規則動詞の Universal Feature の記述例

表 5　‘ちゅーん’「来る」, ‘くーん’「来ない」, ‘ちょーん’「来ている」, ‘ちゃん’「来ない」[14]

‘ちゅー’	Tense=Pres|Polarity=Pos
‘くー’	Tense=Pres|Polarity=Neg
‘ちょー’	Tense=Pres|Aspect=Prog|Polarity=Pos
‘ちゃ’	Tense=Pres|Polarity=Neg

11)　[16] および [15] では,「接辞」と訳されている.
12)　レンマを{ }で囲って表す.
13)　必ず, 活用接尾辞 (AUX) を伴う. 英語の be のような「verbal copula」の UPOS は AUX になる.
14)　レンマは{ちゅー}であり, これら VERB には, ‘ん’ などの AUX が必ず後続.