Vol.2026-CH-140 No.2
2026/2/1

日本漢文Universal Dependenciesの開発

安岡孝一 (京都大学)

1 はじめに

筆者が班長を務める京都大学人文科学研究所共同研究班「古典中国語コーパスの応用研究」(2023 年 4 月〜2026 年 3 月．班員: Christian Wittern，池田巧，李媛，劉冠偉，久保旭，守岡知彦，山崎直樹，二階堂善弘，鈴木慎吾，師茂樹，藤田一乗) では，古典中国語 (漢文) の周辺言語に対し，Universal Dependencies (UD) [1] の適用に挑戦している．対象言語の一つが日本漢文であり,『日本書紀』を手始めに『日本靈異記』『御成敗式目』『狄島夜話記』『日本樂府』『日本漢詩』の UD コーパスを開発中である．しかしながら，これらの UD コーパスを古典中国語 UD [2] に押し込めるのは無理があり，コードスイッチングを含めた様々な手法を追加する必要が生じている．

本稿では，日本漢文 UD が古典中国語 UD からどうハミ出てしまうのか，どのような解決法が適切なのか，模索する．なお，本稿における日本漢文 UD コーパスの開発は，科学研究費補助金基盤研究 (B)23K28379『古典漢文依存文法コーパスから日本漢文コーパスへの展開』の研究助成を受けている．また，コーパス管理システムの統合的開発と，それに伴うコーパス作成作業は，国文研プロジェクト型共同研究 (重点課題研究)『日本漢文 Universal Dependencies に向けた OCR・品詞付与・係り受け解析システムの統合的開発』の成果である．

2 Universal Dependenciesの概要

UD は，書写言語における品詞・形態素属性・依存構造 (係り受け関係) を，言語に関わらず記述する手法である．句構造を考慮せずに係り受け関係を記述することで，言語横断性を高めており，全ての文法構造を単語間のリンクで記述するのが特徴である．

依存構造解析それ自体は，Tesnière の構造的統語論 [3] に源を発し，Мельчук の有向グラフ記述 [4] によって，一応の完成を見た手法である．その最大の特長は，いわゆる動詞中心主義によって言語横断的な記述が可能だという点にあり，Мельчук 依存文法をコンピュータ向けに洗練した UD においても，言語に関わらない記述，という特長が前面に押し出されている．UD における文法構造記述は，句構造を考慮せず，全てを単語間のリンクとして表現する．これは，Мельчук の有向グラフ記述が，単語間のリンクという形態を取っていたからであり，そういう割り切りの結果として，言語横断的な文法構造記述を可能としているのである．

UD 依存構造コーパスの交換用フォーマットとして，CoNLL-U と呼ばれるタブ区切りテキスト (文字コードは UTF-8) が規定されている．CoNLL-U の各行は各単語に対応しており，以下に示す 10 個のタブ区切りフィールドで構成される．

1. ID: 単語ごとに付与されたインデックスで，文ごとに 1 から始まる整数．縮約語に対しては，単語の範囲を示すのも可．
2. FORM: 語，または，句読記号．
3. LEMMA: 基底形，語幹．
4. UPOS: UD で規定された言語普遍な品詞タグ[a]．
5. XPOS: 言語固有の品詞タグ．
6. FEATS: UD で規定された言語普遍な形態素属性のリスト．言語固有の拡張も可．
7. HEAD: 当該の単語の係り受け元 ID．係り受け元が無い場合は 0 とする．
8. DEPREL: UD で規定された言語普遍な係り受けタグ (表 1)．HEAD が 0 の場合は root とする．言語固有の拡張も可．
9. DEPS: 複数の係り受け元を持つ場合，全ての HEAD:DEPREL ペア．
10. MISC: その他のアノテーション．

ID・FORM・LEMMA は，単語そのものに関するフィールドである．UPOS・XPOS・FEATS は，単語の品詞と形態素属性に関するフィールドである．HEAD・DEPREL・DEPS は，単語の係り受けに関するフィールドである．

UD における係り受け関係は，単語間の有向グラフを HEAD と DEPREL で記述する．HEAD は，その単語に入る有向枝のリンク元 ID を示しており，DEPREL は，その有向枝における係り受けタグである．ただし，HEAD が 0 の場合，その枝に入るリンク元は存在しない．リンクの本数は単語の個数に等

[a] ADJ・ADP・ADV・AUX・CCONJ・DET・INTJ・NOUN・NUM・PART・PRON・PROPN・PUNCT・SCONJ・SYM・VERB・X の 17 種類．

表 1: UD 係り受けタグ (DEPREL)

	Nominals	Clauses	Modifier Words	Function Words
Core arguments	nsubj 主語 obj 目的語 iobj 間接目的語	csubj 節主語 ccomp 節目的語 xcomp 節補語		
Non-core dependents	obl 斜格補語 vocative 呼称語 expl 形式語 dislocated 外置語	advcl 連用修飾節	advmod 連用修飾語 discourse 談話要素	aux 動詞補助成分 cop 繋辞 mark 標識
Nominal dependents	nmod 体言による連体修飾語 appos 同格 nummod 数量による修飾語	acl 連体修飾節	amod 用言による連体修飾語	det 決定詞 clf 類別詞 case 格表示

Coordination	MWE	Loose	Special	Other
conj 接続 cc 接続詞	fixed 固着 flat 並列 compound 複合	list 細目 parataxis 隣接表現	orphan 親なし goeswith 泣き別れ reparandum 言い損じ	punct 句読点 root 親 dep 未定義

しく，各リンクのリンク先は，全て互いに異なっている．すなわち，各単語から出るリンクは複数有り得るが，各単語に入るリンクは 1 つだけである．なお，リンクはループしない．

UD の係り受けリンクは，Мельчук 依存文法の後裔であり，いわゆる動詞中心主義である．動詞をリンク元として，主語や目的語へとリンクする．修飾関係においては，被修飾語から修飾語へとリンクする．ただし，側置詞 (前置詞や後置詞) を体言の修飾語だとみなす点 [5] が，Мельчук とは異なっている．ちなみに，コピュラ文においては，補語をリンク元として，主語へとリンクする．

3 日本漢文 UD コーパスの開発

われわれは現在，表 2 に示す日本漢文 UD コーパスを開発中である．具体的な開発手順としては，各画像から NDL 古典籍 OCR-Lite [6] や Qwen3-VL[b] で本文テキストを起こし，テキストを整形・訂正した後に SuPar-Kanbun [2] を用いて仮の古典中国語 UD を作成し，deplacy [7] 上の古典中国語 UD エディターで編集をおこなう，という流れが確立しつつある．

3.1 『日本書紀』UD コーパス

国立国会図書館蔵『日本書紀』慶長 4 年刊本 (請求記号 WA7-251)[c]・慶長 15 年古活字本 (請求記号 WA7-120)[d] をもとに，『日本書紀』UD コーパスの開発を継続しておこなっている [8, 9]．本文の解釈は，日本古典文學大系 [10, 11] に依っている．

3.2 『日本靈異記』UD コーパス

日本古典全集『日本靈異記』[e] をもとに，景戒『日本靈異記』UD コーパスの開発をおこなっている．本文の解釈は，新日本古典文学大系 [12] に依っている．

3.3 『御成敗式目』UD コーパス

新校羣書類從『御成敗式條』[f] をもとに，『御成敗式目』UD コーパスの開発をおこなっている．本文の解釈は，『御成敗式目ハンドブック』[13] に依っている．

3.4 『狄島夜話記』UD コーパス

東北大学附属図書館蔵『狄島夜話記』(請求記号 丙 C/3-13/4) をもとに，釋天嶺『狄島夜話記』UD コーパスの開発をおこなっている [14]．ただし『狄島夜話記』は，その大部分が『松島夜話』[15] に収録されていることから，将来的には『松島夜話』UD コーパスの開発に移行する予定である．

[b] https://arxiv.org/abs/2511.21631
[c] https://dl.ndl.go.jp/pid/1288454
[d] https://dl.ndl.go.jp/pid/2607065
[e] https://dl.ndl.go.jp/pid/1019515
[f] https://dl.ndl.go.jp/pid/1879793/1/204

表 2: 開発中の日本漢文 UD コーパス

『日本書紀』	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/Nihonshoki.html
『日本靈異記』	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/Ryoiki.html
『御成敗式目』	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/Goseibai.html
『狄島夜話記』	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/Ezogashima.html
『日本樂府』	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/Nihongafu.html
『日本漢詩』上	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/NihonKanshi1.html
下	https://corpus.kanji.zinbun.kyoto-u.ac.jp/db-machine/~yasuoka/kyodokenkyu/ud-kanbun/conllusvg/NihonKanshi2.html

3.5 『日本樂府』UD コーパス

国文学研究資料館蔵『日本樂府』(請求記号 ナ8-53)[g]をもとに，頼山陽『日本樂府』UD コーパスの開発をおこなっている．本文の解釈は，福山天蔭 [16] に依っている．

3.6 『日本漢詩』UD コーパス

新釈漢文大系『日本漢詩』[17, 18] をもとに,『日本漢詩』UD コーパスの開発をおこなっている．ただし，頼山陽「蒙古來」が『日本樂府』と重複しており，他の日本漢文 UD コーパスとの関係を調整する必要がある．

4 古典中国語 UD との差異

前節で「古典中国語 UD エディターで編集をおこなう，という流れが確立しつつある」と述べたものの，この「流れ」は残念ながら「確立」していない．われわれが製作中の日本漢文 UD コーパスは，いくつかの点で古典中国語 UD からハミ出てしまっており，それが「確立」を阻んでいるのだ．

4.1 接頭辞の「御」

「可有御成敗」の「御」は，接頭辞とみなすことにした (図 1)．ただし，このような接頭辞の「御」は，古典中国語 UD には例がない．古典中国語 UD における「御」は動詞であり，接頭辞の「御」は日本漢文に特有だと考えられる．その一方,『日本書紀』には動詞の「御」も見られる (図 2) ため，日本漢文 UD における「御」の扱いは複雑にならざるを得ない．

4.2 主語の「者」と条件節の「者」

「吾兒宮首者即脚摩乳手摩乳也」の「者」は，主語を明示するために用いられており，日本語の係助詞「は」に近い．これを古典中国語 UD で記述するなら,「者」に nsubj を繋ぐべき (図 3) である．

一方,「若以平心射者則當無恙」の「者」は，条件節の末尾を示しており，日本語の接続助詞「ば」に近い．これに対して，われわれは「者」を mark でぶらさげる (図 4) ことにした．主語の「者」と条件節の「者」[19] で扱いが違うが，このあたりしか落とし所が無いように思われる [8].

4.3 動詞の直後の「之」

動詞の直後にあらわれる「之」のうち，たとえば「皇子歸而宿之」の「之」については，議論 [20] がある.「宿」という動詞の目的語ではなく,「宿」が動詞であることをはっきりさせるために，形式目的語あるいは句末助詞として「之」が付けられている，ということらしい．形式目的語ならば expl で，句末助詞ならば discourse:sp で繋ぐのが，古典中国語 UD の流儀である．

この「皇子歸而宿之」に対し，われわれは「之」を PRON とみなし，とりあえず obj で繋ぐことにした (図 5)．これらの多くは動詞の目的語と見ても,「之」が何を指しているのか必ずしも明確でないものの，日本漢文 UD としては問題が無さそうだからである [8]. まあ，作業の都合上と言ってもいい．

4.4 形容詞の「赤」

「踰垣走者足跡赤」は「垣を踰えて走る者は足跡赤し」と読み下されている [16]．いわゆる二重主語構文 [21] であり,「赤」は日本語の形容詞とみなせる．しかしながら，われわれは古典中国語 UD にお

[g] https://kokusho.nijl.ac.jp/biblio/200004383

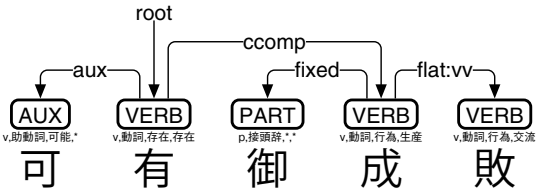

図 1: 『御成敗式目』二十六「可有御成敗」の UD 案

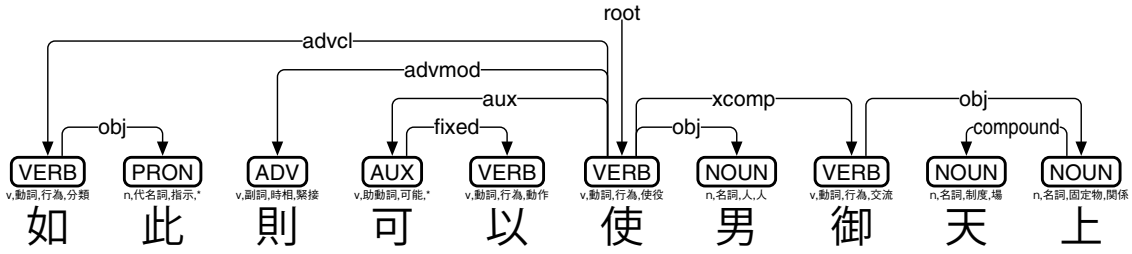

図 2: 『日本書紀』巻一「如此則可以使男御天上」の UD 案

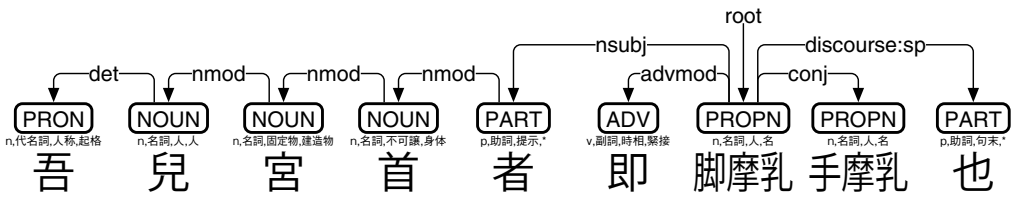

図 3: 『日本書紀』巻一「吾兒宮首者即脚摩乳手摩乳也」の UD 案

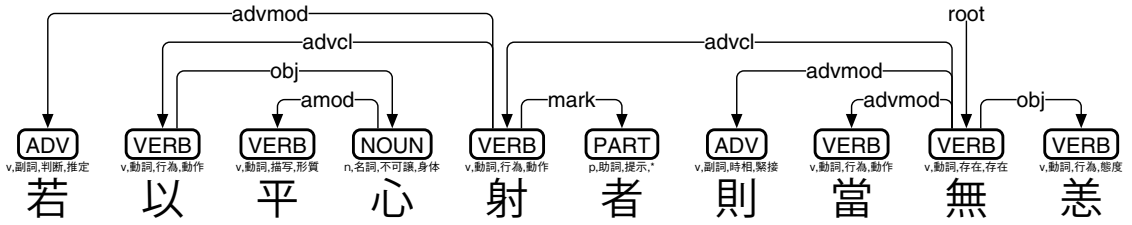

図 4: 『日本書紀』巻二「若以平心射者則當無恙」の UD 案

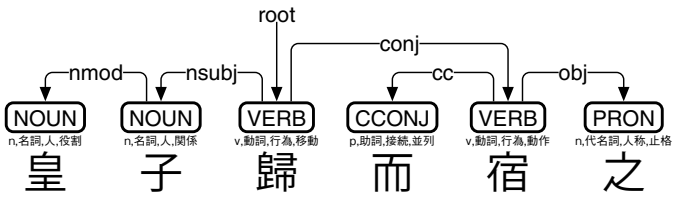

図 5: 『日本書紀』巻二十六「皇子歸而宿之」の UD 案

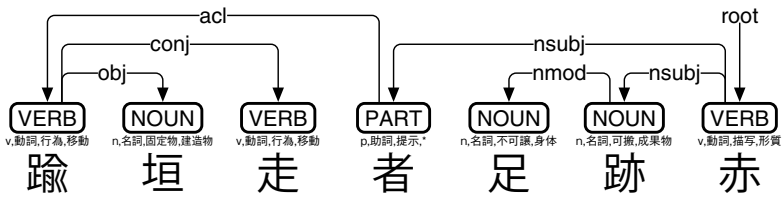

図 6: 『日本樂府』四十七「蹤垣走者足跡赤」の UD 案

いて，形容詞という品詞を廃止 [22] しており，ここは VERB とした (図 6).

5 コードスイッチング

ひらがな成立以前に書かれた『日本書紀』『日本靈異記』においては，漢字音を用いて上代日本語を直接記述する記法と，上代日本語を古典中国語 (漢文) に翻訳して記述する記法が混用しており，書写言語でのコードスイッチング [23] が起こっている．このようなコードスイッチングを日本漢文 UD で扱うべく，CoNLL-U の MISC に Lang=ojp [h] タグを導入した．たとえば「背揮此云志理幣提爾布倶」の後半は，「しりへでにふく」という上代日本語が漢字音を用いて直接記述されている．この部分については Lang=ojp タグを付与した上で，国語研短単位 [24] に基づく日本語 UD [25] を適用した (図 7).

また『日本書紀』巻十九には，新羅の闘将の言葉として「久須尼自利」が現れる．末音添記 [26] の可能性を考えつつ，「굿 내 싈」とみなして Lang=oko [i] タグを導入してみた (図 8) が，正直なところ全く自信がない．

『狄島夜話記』は，基本的に古典中国語で書かれているものの，漢字音を用いて書かれたアイヌ語が紛れ込んでいる [14]．たとえば「因崇之曰久岨部加茂伊」は，後半が「kor pe kamuy」の音写である．これについても，Lang=ain [j] タグを導入し，アイヌ語 UD [27] を適用した (図 9).

ただし「菩薩」「夜叉」「頭陀」など梵語由来の音写語に対しては，Lang=san [k] タグを付与しなかった．また，地名や人名などの固有名詞についても，基本的に Lang タグを付与していない．

6 おわりに

われわれが開発中の日本漢文 UD に関し，現時点での進捗状況といくつかの問題点を述べた．われわれとしては，日本漢文 UD をいずれリリース[l] したいと考えているものの，その際には，Classical Chinese ブロックに入れるか，Old Japanese ブロックに入れるか，カレル大学その他の機関と調整が必要である．あるいは，Lang タグごとに複数ブロックに泣き別れ，というケースも有り得るが，それはわれわれの望む形ではない．頭の痛いところである．

[h] https://iso639-3.sil.org/code/ojp
[i] https://iso639-3.sil.org/code/oko
[j] https://iso639-3.sil.org/code/ain
[k] https://iso639-3.sil.org/code/san
[l] https://universaldependencies.org

参考文献

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

[2] 安岡孝一, ウィッテルン クリスティアン, 守岡知彦, 池田巧, 山崎直樹, 二階堂善弘, 鈴木慎吾, 師茂樹, 藤田一乗: 古典中国語（漢文）Universal Dependencies とその応用, 情報処理学会論文誌, Vol.63, No.2 (2022 年 2 月), pp.355-363.

[3] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).

[4] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

[5] Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text Processing and Computational Linguistics (April 2015), pp.3-16.

[6] 青池亨: CPU 環境で高速に動作する軽量 OCR「NDL 古典籍 OCR-Lite」の開発, 人文科学とコンピュータシンポジウム「じんもんこん 2024」論文集 (2024 年 12 月), pp.181-186.

[7] 安岡孝一: Universal Dependencies にもとづく多言語係り受け可視化ツール deplacy, 人文科学とコンピュータシンポジウム「じんもんこん 2020」論文集 (2020 年 12 月), pp.95-100.

[8] 安岡孝一, ウィッテルン クリスティアン, 池田巧, 藤田一乗, 守岡知彦, 山崎直樹, 二階堂善弘, 鈴木慎吾, 師茂樹: 『日本書紀』Universal Dependencies への挑戦, 人文科学とコンピュータシンポジウム「じんもんこん 2023」論文集 (2023 年 12 月), pp.169-176.

[9] 安岡孝一: 『日本書紀』におけるコードスイッチングについて, 日本漢字学会第 7 回研究大会予稿集 (2024 年 12 月), pp.31-54.

```
# text = 背揮此云志理幣提爾布倶
1  背      背    NOUN  n, 名詞, 不可譲, 身体          _  2  nsubj     _  SpaceAfter=No
2  揮      揮    VERB  v, 動詞, 行為, 動作            _  4  csubj     _  SpaceAfter=No
3  此      此    PRON  n, 代名詞, 指示,*              _  4  expl      _  SpaceAfter=No
4  云      云    VERB  v, 動詞, 行為, 伝達            _  0  root      _  SpaceAfter=No
5  志理幣  後方  NOUN  名詞-普通名詞-一般             _  6  compound  _  Lang=ojp|SpaceAfter=No
6  提      手    NOUN  名詞-普通名詞-助数詞可能       _  8  obl       _  Lang=ojp|SpaceAfter=No
7  爾      に    ADP   助詞-格助詞                    _  6  case      _  Lang=ojp|SpaceAfter=No
8  布倶    吹く  VERB  動詞-一般                      _  4  ccomp     _  Lang=ojp|SpaceAfter=No
```

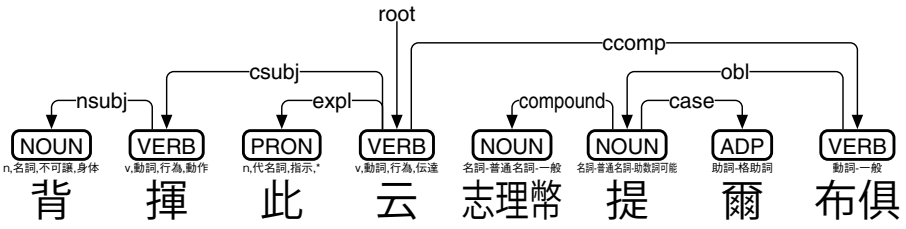

図 7: 『日本書紀』巻二「背揮此云志理幣提爾布倶」の CoNLL-U・UD 案

```
# text = 久須尼自利
1  久須  긋  NOUN  _  _  3  vocative  _  Lang=oko|SpaceAfter=No
2  尼    내  PRON  _  _  3  det       _  Lang=oko|SpaceAfter=No
3  自利  실  NOUN  _  _  0  root      _  Lang=oko|SpaceAfter=No
```

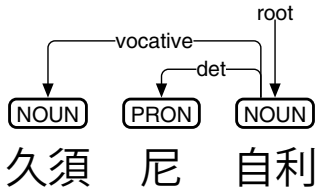

図 8: 『日本書紀』巻十九「久須尼自利」の CoNLL-U・UD 案

```
# text = 因崇之曰久delta部加茂伊
1  因      因     VERB  v, 動詞, 行為, 動作          _  2  advmod  _  SpaceAfter=No
2  崇      崇     VERB  v, 動詞, 行為, 態度          _  0  root    _  SpaceAfter=No
3  之      之     PRON  n, 代名詞, 人称, 止格        _  2  obj     _  SpaceAfter=No
4  曰      曰     VERB  v, 動詞, 行為, 伝達          _  2  parataxis  _  SpaceAfter=No
5  久delta  kor   VERB  他動詞                       _  6  acl     _  Lang=ain|SpaceAfter=No
6  部      pe     NOUN  形式名詞                     _  7  nmod    _  Lang=ain|SpaceAfter=No
7  加茂伊  kamuy  NOUN  名詞                         _  4  obj     _  Lang=ain|SpaceAfter=No
```

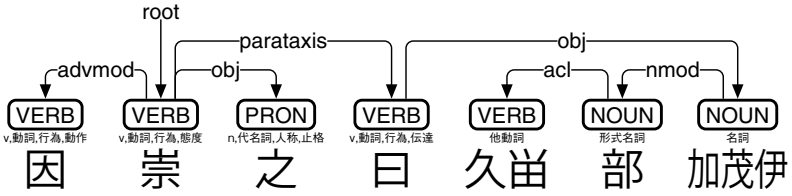

図 9: 『狄島夜話記』「因崇之曰久delta部加茂伊」の CoNLL-U・UD 案

[10] 日本書紀 上, 日本古典文學大系 67, 東京: 岩波書店 (1967 年 3 月).

[11] 日本書紀 下, 日本古典文學大系 68, 東京: 岩波書店 (1965 年 7 月).

[12] 日本霊異記, 新日本古典文学大系 30, 東京: 岩波書店 (1996 年 12 月).

[13] 御成敗式目ハンドブック, 東京: 吉川弘文館 (2024 年 3 月).

[14] 安岡孝一:『狄島夜話記』におけるアイヌ語について, 日本漢字学会第 8 回研究大会予稿集 (2025 年 12 月), pp.33-45.

[15] 釋天嶺: 松島夜話, 松島: 東礬, 京都: 柳枝軒多左衞門, 江戸: 小川彦九郎 (元文 3 年 2 月).

[16] 福山天蔭: 詠史日本樂府物語, 東京: 東白堂書房 (1938 年 6 月).

[17] 猪口篤志: 日本漢詩 上, 新釈漢文大系 45, 東京: 明治書院 (1972 年 8 月).

[18] 猪口篤志: 日本漢詩 下, 新釈漢文大系 46, 東京: 明治書院 (1972 年 9 月).

[19] 藤井茂利: 東アジア比較言語学研究 (二) —「者」の用法について —, 福岡大学総合研究所報, 第 162 号 [人文科学編 (第 99 号)] (1994 年 6 月), pp.7-22.

[20] 榎本福寿:『日本書紀』の「之」に関する調査研究報告, 京都語文, 第 9 号 (2002 年 10 月), pp.88-135.

[21] 草野清民: 國語の特有せる語法—總主, 帝國文學, 第 5 巻, 第 5 (1899 年 5 月), pp.16-29.

[22] 山崎直樹, 守岡知彦, 安岡孝一: 古典中国語形態素解析のための品詞体系再構築, 人文科学とコンピュータシンポジウム「じんもんこん 2012」論文集 (2012 年 11 月), pp.39-46.

[23] Mark Sebba, Shahrzad Mahootian and Carla Jonsson: Language Mixing and Code-Switching in Writing, New York: Routledge (2012).

[24] 近藤明日子: 近代文語 UniDic 短単位規程集, Ver.1.1, 立川: 国立国語研究所コーパス開発センター (2016 年 3 月).

[25] 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 (2019 年 3 月), pp.3-36.

[26] Lee SeungJae (李丞宰): Old Korean Writing on Wooden Tablets and its Implications for Old Japanese Writing, Seoul Journal of Korean Studies, Vol.27, No.2 (December 2014), pp.151-185.

[27] 安岡孝一, 安岡素子: ローマ字・カタカナ・キリル文字によるアイヌ語 Universal Dependencies の可能性, Evidence-based Linguistics Workshop 2023 発表論文集 (2023 年 9 月), pp.47-60.