『日本書紀』Universal Dependenciesへの挑戦

安岡孝一・ウィッテルン クリスティアン・池田巧・藤田一乘 (京都大学)
守岡知彦 (国文学研究資料館)
山崎直樹・二階堂善弘 (関西大学)
鈴木慎吾 (大阪大学)
師茂樹 (花園大学)

概要: われわれの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」では，古典中国語 (漢文) の周辺言語に対し，Universal Dependencies (UD) の適用に挑戦している．その対象言語の一つが日本漢文であり，手始めとして『日本書紀』に着手した.『日本書紀』は，基本的に古典中国語で書かれており，古典中国語 UD で記述可能だと考えていたからだ．しかし，読みが甘かった.『日本書紀』は手ごわすぎる．古典中国語 UD に押し込めるには，無理がある．本稿では『日本書紀』におけるいくつかの例を示し，それらに関する解決法を模索する．

キーワード: 日本漢文，形態素解析，文法解析

Towards Universal Dependencies of Nihon Shoki

Koichi Yasuoka / Christian Wittern / Takumi Ikeda / Kazunori Fujita (Kyoto University)
Tomohiko Morioka (National Institute of Japanese Literature)
Naoki Yamazaki / Yoshihiro Nikaido (Kansai University)
Shingo Suzuki (Osaka University)
Shigeki Moro (Hanazono University)

Abstract: Our research group “Applied Study of Classical Chinese Corpora” (Institute for Research in Humanities, Kyoto University) is investigating to apply Universal Dependencies (UD) of Classical Chinese to other languages including Old Japanese. We chose Nihon Shoki (日本書紀) for our new UD corpus, since 日本書紀 looked to be written in Classical Chinese. We have realized that we were too courageous and that we thought it too easy. In this paper we mention our efforts and odds to develop 日本書紀 UD.

Keywords: Japanized Classical Chinese, Morphological Analysis, Dependency-Parsing

1 はじめに

書写言語としての上代日本語は，漢字で書かれている．日本固有の文字 (ひらがな・カタカナ) が創られる以前に，百済・隋・唐で用いられていた漢字が飛鳥・奈良に流入し，漢字を用いて上代日本語が記述されたためである．また，漢字の流入は同時に，漢文 (古典中国語) の流入でもあった．結果として，漢字による上代日本語の書写は，漢字音を用いて上代日本語を直接記述する記法と，上代日本語を漢文に翻訳して間接記述する記法が，両方，用いられることとなった．前者の代表を『万葉集』とするなら，後者の代表は『日本書紀』である．

われわれは，2008 年 4 月に京都大学人文科学研究所共同研究班「東アジア古典文献コーパスの研究」を組織し，続く 2013 年 4 月に共同研究班「東アジア古典文献コーパスの応用研究」を組織して，古典中国語に対する形態素解析の研究をおこなってきた [1]. さらに 2016 年 4 月には共同研究班「東アジア古典文献コーパスの実証研究」を組織し，続く 2020 年 4 月に共同研究班「古典中国語のコーパスの研究」を組織して，古典中国語に Universal Dependencies (UD) を適用する形で，形態素解析と係り受け解析を並行しておこなう手法を開発した [2].

2023 年 4 月に共同研究班「古典中国語コーパスの応用研究」を組織するにあたり，われわれは，これまで開発してきた古典中国語 UD を周辺言語に応用することを考え，対象言語の一つとして，上代日本語の

解析に着手することにした．その際『古事記』や『万葉集』は後回しにして，まずは『日本書紀』をターゲットとして選んだ.『日本書紀』の方が，古典中国語に近いと考えたからである．

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) 23H03690『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開発，およびそれらを用いたコーパス作成作業は，文部科学省『AI 等の活用を推進する研究データエコシステム構築事業』の支援を受けている．

2　古典中国語 UD 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 [3]．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 [4] に源を発し，Мельчук の有向グラフ記述 [5] 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．これにより，言語横断的な文法構造記述を可能としている．

UD 係り受け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，表 1 に示す 10 個のタブ区切りフィールドで構成される．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の可能性があるが，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 依存文法の後裔にあたり，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 [6] が，Мельчук とは異なっている．また，コピュラ文においては動詞中心主義を採らず，補語をリンク元として，主語や繋辞へとリンクする．

UD を古典中国語へ適用 [2] するに際して，われわれは係り受けタグ (DEPREL) を拡張した (図 3 のうち nsubj:pass，nsubj:outer，csubj:outer，obl:tmod，obl:lmod，discourse:sp，compound:redup，flat:vv)．特に discourse:sp は文助詞 (sentence particle) を表しており [7]，古典中国語における動賓終構造 (predicate-object-final structure) の終を，記述するために用いている．例として「未聞好學者也」の CoNLL-U と，deplacy [8] による可視化を図 1 に示す.「聞」から「者」へ obj が,「聞」から「也」へ discourse:sp が繋がれており,「聞-者-也」という動賓終構造を，UD で記述できている．

3　古典中国語 UD による『日本書紀』

『日本書紀』[9, 10] の各文に対し，SuPar-Kanbun [2] を用いて仮の古典中国語 UD を作成し，古典中国語 UD エディターで適切に編集した上で,『日本書紀』UD コーパスとして公開する．これが,『日本書紀』UD コーパス作成における，われわれの当初プランだった．

しかし，実際に作業を始めてみると，この「適切に編集」というのが『日本書紀』では難しい.『日本書紀』には，古典中国語としては不適切な表現，いわゆる倭習 [11] があり，これを古典中国語 UD として「適切に編集」してしまうと,『日本書紀』そのものの持つ依存構造から乖離してしまう．つまり，古典中国語 UD としては不適切であっても，それはそれとして残す必要があるのだ．以下に，作業上の問題となっている点を，かいつまんで示す．

3.1　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 [12] があ

表 1: CoNLL-U の各フィールド

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 (表 2)．
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．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．

表 2: UD 品詞タグ (UPOS)

Open class words	Closed class words	Other
ADJ 形容詞	ADP 側置詞	PUNCT 句読点
ADV 副詞	AUX 助動詞	SYM 記号
INTJ 感嘆詞	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 接辞	
	PRON 代名詞	
	SCONJ 従属接続詞	

表 3: 古典中国語 UD における係り受けタグ (DEPREL)

	Nominals	Clauses	Modifier Words	Function Words
Core arguments	nsubj 主語 ↪nsubj:pass [受動文] ↪nsubj:outer [外置] obj 目的語 iobj 間接目的語	csubj 節主語 ↪csubj:outer [外置] ccomp 節目的語 xcomp 節補語		
Non-core dependents	obl 斜格補語 ↪obl:tmod [時] ↪obl:lmod [場所] 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 ↪discourse:sp [文助詞]	aux 動詞補助成分 cop 繋辞 (copula) mark 標識 (marker)
Nominal dependents	nmod 体言による連体修飾語 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 compound 複合 (endocentric) ↪compound:redup [重畳] flat 並列 (exocentric) ↪flat:vv [動詞類]	list 細目 parataxis 隣接表現	orphan 親なし	root 親

```
# text = 未聞好學者也
1  未  未  ADV   v, 副詞, 否定, 有界  Polarity=Neg  2  advmod       _  SpaceAfter=No
2  聞  聞  VERB  v, 動詞, 行為, 伝達  _             0  root         _  SpaceAfter=No
3  好  好  VERB  v, 動詞, 行為, 態度  _             5  acl          _  SpaceAfter=No
4  學  學  NOUN  n, 名詞, 行為, ∗     _             3  obj          _  SpaceAfter=No
5  者  者  PART  p, 助詞, 提示, ∗     _             2  obj          _  SpaceAfter=No
6  也  也  PART  p, 助詞, 句末, ∗     _             2  discourse:sp  .  SpaceAfter=No
```

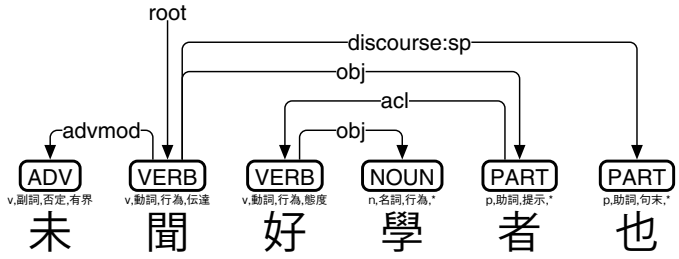

図 1: 古典中国語 CoNLL-U と deplacy による可視化

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 2)．というのも，同様の「之」に「于時火火出見尊乃歌之」(図 3) などの例があるが，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，UD 依存構造としては問題が無さそうだからである．まあ，作業の都合上と言ってもいい．ただし，動詞の直後の「之」であっても,「此海陸不相通之縁也」のような接続助詞の「之」については，SCONJ とみなし mark で繋ぐことにした (図 4).

3.2 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 5) 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 6) ことにした．主語の「者」と条件節の「者」[13] 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．

3.3 謙譲語の「罷」

「遣罷歸之」の「罷」は，上代日本語における謙譲語「まかる」[14] である．古典中国語の「罷」とは異なる意味だが，いずれも動詞なので，とりあえず VERB としておいた (図 7).

3.4 固有名詞と尊称

図 5 の「脚摩乳」「手摩乳」は，夫婦であり「摩乳」を共有していることから，たとえば「脚」「摩乳」と「手」「摩乳」である可能性が考えられる．考えられるが，それは上代日本語としての可能性であって，古典中国語ではないことから，われわれは「脚摩乳」と「手摩乳」を，それぞれ 1 語とした．

ただし,「彦」「姫」「尊」「命」「主」などの尊称については，図 3 の「火火出見」「尊」のように，別の語として分離することにした．

3.5 和歌

いわゆる和歌は，どう考えても古典中国語 UD では書けない．漢字音を用いて上代日本語を直接記述しているのだから，あるいは近代日本語 UD [15] で書けるかもしれない．とりあえず,「飫企都鄧利軻茂豆句志磨爾和我謂禰志伊茂播和素邏珥譽能據鄧馭劉母」に対応する近代日本語 UD を，UniDic 品詞を XPOS

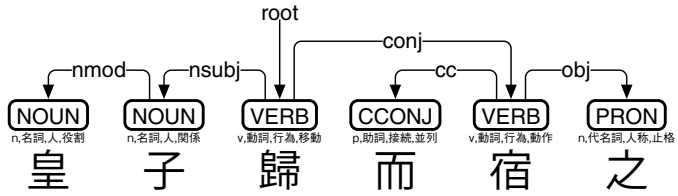

図 2: 巻二十六「皇子歸而宿之」の UD 依存構造案

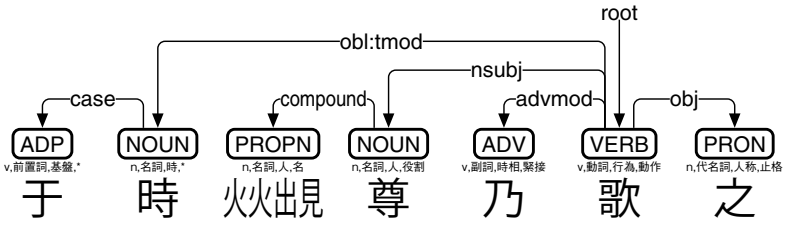

図 3: 巻二「于時火火出見尊乃歌之」の UD 依存構造案

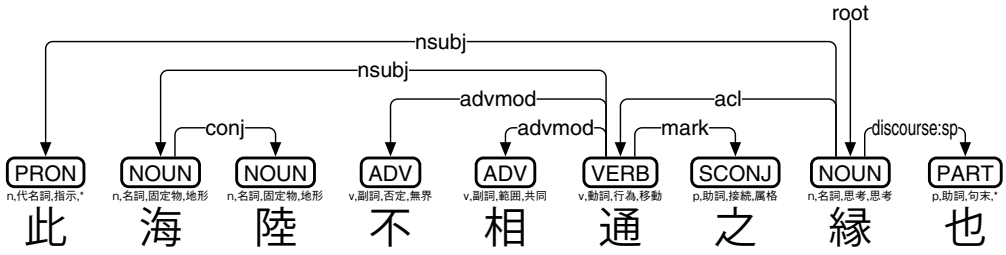

図 4: 巻二「此海陸不相通之緣也」の UD 依存構造案

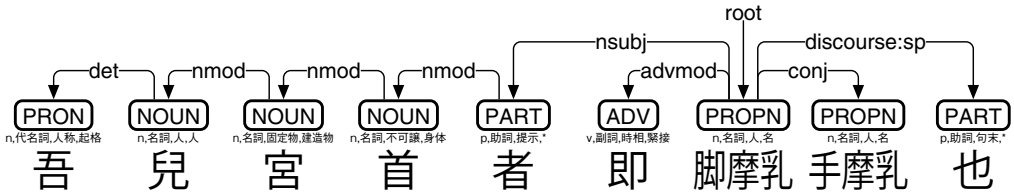

図 5: 巻一「吾兒宮首者即脚摩乳手摩乳也」の UD 依存構造案

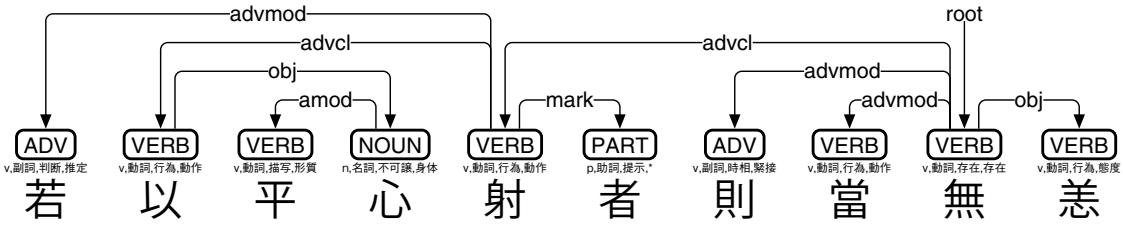

図 6: 巻二「若以平心射者則當無恙」の UD 依存構造案

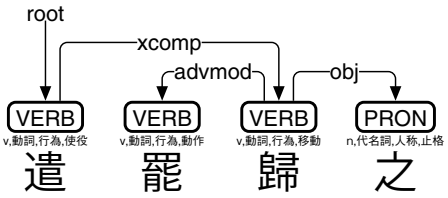

図 7: 巻十七「遣罷歸之」の UD 依存構造案

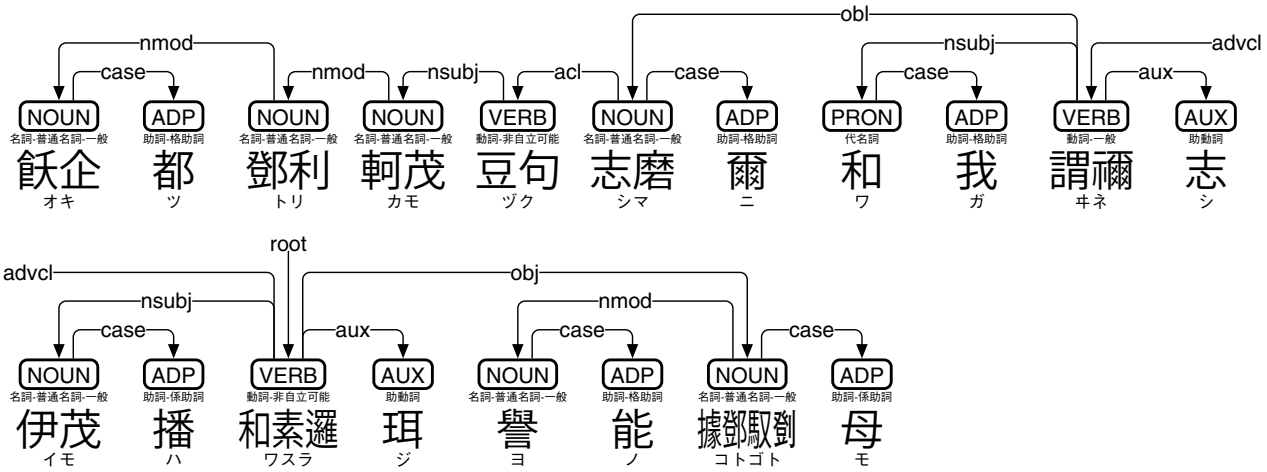

図 8: 巻二「飫企都鄧利軻茂豆句志磨爾和我謂禰志伊茂播和素邏珥譽能據鄧馭劉母」の UD 依存構造案

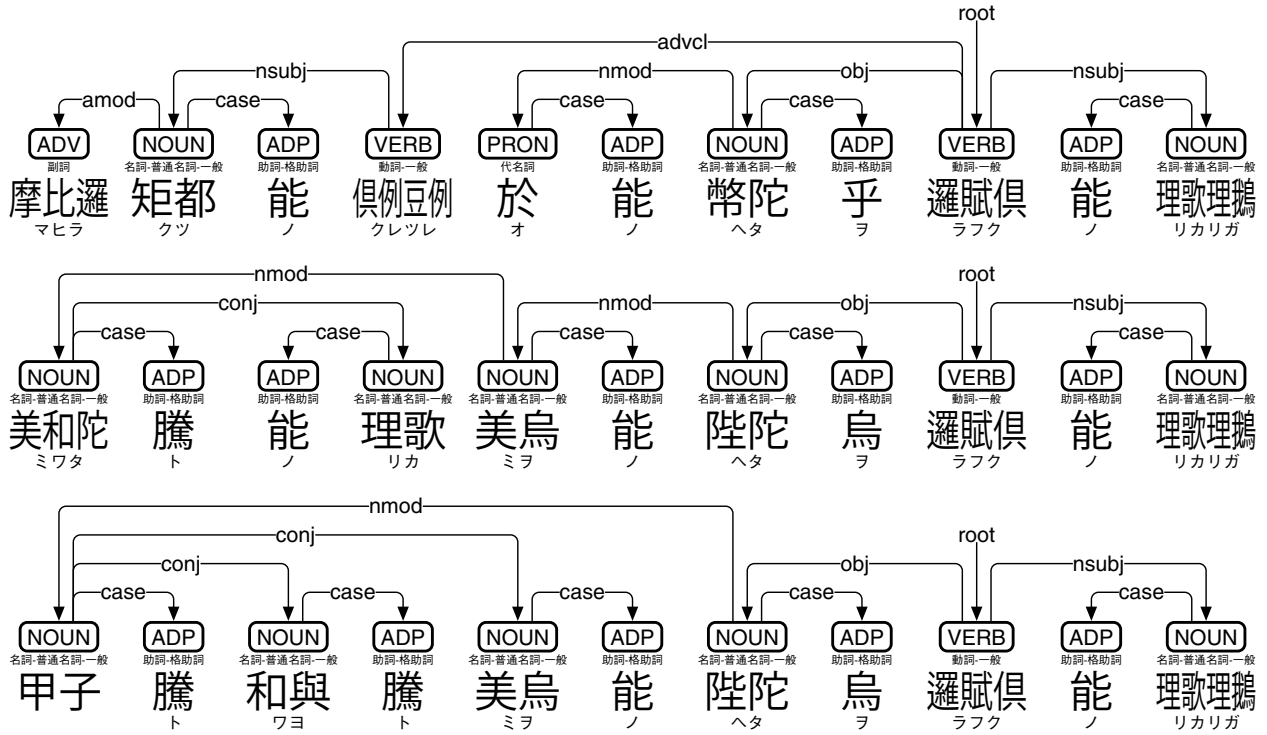

図 9: 巻二十六「摩比邏矩都能倶例豆例於能幣陀乎邏賦倶能理歌理鵝美和陀騰能理歌美烏能陛陀烏邏賦倶能理歌理鵝甲子騰和與騰美烏能陛陀烏邏賦倶能理歌理鵝」の UD 依存構造案

に入れつつ手作業で作ってみた (図 8)．しかし，これが UD として正しいのかどうか，われわれとしては確かめようがない．また，たとえこの UD が正しいとしても，和歌を全て手作業で処理するわけにもいかず，読み下し文などから，いったん仮の近代日本語 UD を自動生成した上で，それを適切に編集する仕掛けが必要だと思われる．

3.6 童謡

さらに難しいのは童謡である．中でも手ごわいのが,「摩比邏矩都能倶例豆例於能幣陀乎邏賦倶能理歌理鵝美和陀騰能理歌美烏能陛陀烏邏賦倶能理歌理鵝甲子騰和與騰美烏能陛陀烏邏賦倶能理歌理鵝」である.「まひらくつのくれつれをのへたをらふくのりかりがみわたとのりかみをのへたをらふくのりかりが甲子とわよとみをのへたをらふくのりかりが」らしい [10] のだが，さっぱり読めない．単語の切れ目すらわからず，形態素解析も係り受け解析も不可能である.「へたをらふくのりかりが」を「田辺を雁々の喰らふ」と読むなど，文字の順序を入れ替える [16] しか無さそうだが，それは UD としては禁じ手であり，かなり悩ましい (図 9).

4 おわりに

UD で『日本書紀』を記述するにあたり，古典中国語 UD の適用を考えた．動詞の直後の「之」，条件節の「者」，謙譲語の「罷」，固有名詞と尊称，に関しては，それぞれ微妙に難しい点はあるものの，何とか古典中国語 UD の範囲で記述可能だと考えられる．しかしながら，和歌や童謡はどうにもならない．古典中国語 UD では歯が立たない．

これを敷衍すると，漢字音を用いて上代日本語を直接書写した「文」は，古典中国語 UD では記述できそうにない，ということを意味している．代わりとしては近代日本語 UD が有力そうだが，どの程度うまくいくのか，現時点では全く見当がつかない．何とか良い記述手法を見つけたいが，まだまだ道程は長い．今後のわれわれの研究の進展に期待されたい．

参考文献

[1] 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: 古典中国語（漢文）の形態素解析とその応用, 情報処理学会論文誌, Vol.59, No.2 (2018 年 2 月), pp.323-331.

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

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

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

[6] 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.

[7] Herman Leung, Rafaël Poiret, Tak-sum Wong, Xinying Chen, Kim Gerdes and John Lee: Developing Universal Dependencies for Mandarin Chinese, 12th Workshop on Asian Language Resources (December 2016), pp.20-29.

[8] 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.

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

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

[11] 森博達: 日本書紀の謎を解く, 中公新書 1502, 東京: 中央公論新社 (1999 年 10 月).

[12] 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.

[13] 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.

[14] 中島京子: 中古語「まかる」の一考察, 語文研究, 第 22 号 (1966 年 10 月), pp.14-25.

[15] 安岡孝一: 形態素解析部の付け替えによる近代日本語 (旧字旧仮名) の係り受け解析, 情報処理学会研究報告, Vol.2020-CH-124『人文科学とコンピュータ』, No.3 (2020 年 9 月), pp.1-8.

[16] 神田秀夫: 斉明紀童謡溯考, 國語と國文學, 第 440 号 (1960 年 11 月), pp.12-27.