情報処理学会論文誌 Vol.63 No.2 355–363 (Feb. 2022)

古典中国語（漢文）Universal Dependenciesとその応用

安岡 孝一[1,a)] ウィッテルン クリスティアン[1,b)] 守岡 知彦[1,c)] 池田 巧[1,d)] 山崎 直樹[2,e)]
二階堂 善弘[2,f)] 鈴木 慎吾[3,g)] 師 茂樹[4,h)] 藤田 一乗[1,i)]

受付日 2021年5月8日, 採録日 2021年9月9日

概要：Universal Dependencies に基づいて，『孟子』『論語』『禮記』『十八史略』の依存構造（係り受け）コーパスを製作した．さらに，このコーパスを用いて，古典中国語の文切り・形態素解析・係り受け解析を統合的に行う解析システムも開発した．Universal Dependencies は，書写言語における品詞・形態素属性・依存構造（係り受け情報）を，言語に依存せず記述する手法である．Universal Dependencies の係り受け記述は，いわゆる動詞中心主義であり，言語横断的であると同時に，古典中国語における動賓終構造の記述にも適している．ただし，Universal Dependencies におけるコピュラ文の記述方法は，古典中国語のコピュラ文との間で微妙に齟齬があり，結果として，補語が節であるようなコピュラ文（約 1.6%）に関しては，記述は行えるものの記法上の問題が残った．

キーワード：漢文コーパス，係り受け解析，形態素解析

Designing Universal Dependencies for Classical Chinese and Its Application

Koichi Yasuoka[1,a)] Christian Wittern[1,b)] Tomohiko Morioka[1,c)] Takumi Ikeda[1,d)]
Naoki Yamazaki[2,e)] Yoshihiro Nikaido[2,f)] Shingo Suzuki[3,g)] Shigeki Moro[4,h)]
Kazunori Fujita[1,i)]

Received: May 8, 2021, Accepted: September 9, 2021

Abstract: We have developed a Classical Chinese Corpus of the whole texts of “孟子”, “論語”, “禮記”, and “十八史略”, using Universal Dependencies (UD). Utilizing our Classical Chinese Corpus, we have also developed syntax analyzers, which perform sentence segmentation, word tokenization, part-of-speech tagging, and dependency parsing of Classical Chinese texts. UD is very suitable to annotate the predicate-object-final structure of Classical Chinese. But, particularly in a copula sentence whose predicate consists of a clause, the guideline of UD is insufficient for annotating Classical Chinese texts, thus we need several extensions of UD.

Keywords: classical Chinese corpus, dependency parsing, morphological analysis

1. はじめに

古典中国語（漢文）は，どのような言語構造を持っているのだろうか．それは，コンピュータによる解析が可能な構造なのだろうか．古典中国語は，単語と単語の間に区切りがなく，文と文の間にも区切りがない．これが，白文と呼ばれる古典中国語の書写形態であり，傍目には，漢字が

[1] 京都大学
Kyoto University, Kyoto 606–8501, Japan
[2] 関西大学
Kansai University, Suita, Osaka 564–8680, Japan
[3] 大阪大学
Osaka University, Minoh, Osaka 562–8558, Japan
[4] 花園大学
Hanazono University, Kyoto 606–8456, Japan
[a)] yasuoka@kanji.zinbun.kyoto-u.ac.jp
[b)] wittern@zinbun.kyoto-u.ac.jp
[c)] tomo@zinbun.kyoto-u.ac.jp
[d)] ikeda@zinbun.kyoto-u.ac.jp
[e)] ymzknk@kansai-u.ac.jp
[f)] nikaido@kansai-u.ac.jp
[g)] suzukish@lang.osaka-u.ac.jp
[h)] s-moro@hanazono.ac.jp
[i)] fujita.kazunori.7z@kyoto-u.ac.jp

連続的に並んでいるだけである．しかも，古典中国語の語彙や文法は，現代中国語とも大きく異なっており [1]，現代中国語の解析手法を，古典中国語にそのまま援用できない．

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」を結成して以来，我々は，古典中国語の文法解析の研究に邁進してきた．様々な手法に手を出したり，あるいは手を引いたりする中で，我々は，MeCab [2] を用いた形態素解析手法，すなわち単語切りと品詞付与を同時に行う手法が，古典中国語にも適用可能だと確信 [3] し，それに適した 4 階層の品詞体系を練り上げる [4] とともに，古典中国語の形態素解析システムを製作 [5] した．これにより，古典中国語の文法解析への足がかりができたことから，次なる解析手法をどうすべきか，我々は模索していた．

2016 年 4 月に新たな共同研究班「東アジア古典文献コーパスの実証研究」を立ち上げ，さらなる手法を試していく中で，我々は，Universal Dependencies（以下「UD」）に出会った．UD は，言語横断的な依存構造（係り受け）コーパスを開発するプロジェクト [6] である．品詞・形態素属性・係り受け情報を言語に依存せず記述することで，言語横断的な情報とすることを目指しており，カレル大学の LINDAT/CLARIN を中心に，世界中で開発が行われている．ただ，UD2.0（2017 年 3 月発表）の時点では，ヨーロッパの各言語（単語の間に区切りを持つ）が UD における開発の中心であり，単語の間に区切りを持たない書写言語は，現代中国語（繁体）UD と現代日本語 UD [*1] だけだった．古典中国語のような，単語の間にも文の間にも区切りを持たない書写言語は，まだ誰も試していなかった．

幾多の検討 [8] と作業 [9] を経て，我々は，四書（孟子・論語・大學・中庸）の依存構造コーパス（11176 文，55026 語，56768 字）を製作 [10] し，UD2.4（2019 年 5 月発表）に参画することで，UD における開発の一翼を担うことにした．開発は現在も継続しており，本稿執筆時点の UD2.8（2021 年 5 月発表）は，『孟子』『論語』『禮記』『十八史略』の依存構造コーパス（55514 文，269002 語，280769 字）を収録している．本稿では，古典中国語 UD の開発における，我々の研究成果を俯瞰する．さらにその応用として，古典中国語の文切り・形態素解析（単語切りと品詞付与）・係り受け解析を統合的に行う解析システムについて述べる．

2. 古典中国語 UD の開発

2.1 UD の基本構造

UD は，書写言語における品詞・形態素属性・依存構造（係り受け情報）を，言語に依存せ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，すべての構文構造を単語間のリンクで記述するのが特徴である [11]．

表 1 CoNLL-U のタブ区切りフィールド
Table 1 CoNLL-U format of UD.

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．
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．
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．係り受け元がない場合は，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．言語固有の拡張も可．
9. DEPS：複数の係り受け元を持つ場合，すべての HEAD:DEPREL ペア．
10. MISC：その他のアノテーション．

〈https://universaldependencies.org/format.html〉[*3] より抄訳

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 UTF-8 テキストが定められている．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，単語間の係り受け[*2]に関するフィールドである．

我々の古典中国語 UD では，これらのフィールドのうち，ID（2.2 節），FORM（2.2 節），LEMMA（2.2 節），UPOS（2.3 節），XPOS（2.3 節），FEATS（2.3 節），HEAD（2.4 節），DEPREL（2.4 節），MISC（2.2 節）の 9 つを用いる．DEPS フィールドは使用しない．

2.2 単語切り

我々は，MeCab による形態素解析手法の成果 [5] を，古典中国語 UD における単語切りにそのまま用いることにした．たとえば「孟子見梁惠王」という文は，形態素解析によって「孟子」「見」「梁」「惠」「王」という 5 つの単語に分割し，それぞれ 1〜5 の ID を振ったうえで，各単語を FORM に入れる（**図 1**）．LEMMA は，対応する FORM の「康熙字典体」とする．これらに加え，MISC には各単語のグロ

[*1] UD2.0 時点の現代日本語 UD は，google 提供の Wikipedia 係り受けデータを変換したものであり，再整備 [7] 後の現代日本語 UD とは異なる．

[*2] UD における係り受け情報は，基本的に，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，その単語に入る有向枝に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，すべ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ありうるが，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．

[*3] アクセス日 2021 年 5 月 5 日．以下，すべての URL で同様．

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
1	孟子	孟子	PROPN	n, 名詞, 人, 複合的人名	NameType=Prs	2	nsubj	_	Gloss=Mencius|SpaceAfter=No
2	見	見	VERB	v, 動詞, 行為, 動作	_	0	root	_	Gloss=see|SpaceAfter=No
3	梁	梁	PROPN	n, 名詞, 主体, 国名	Case=Loc|NameType=Nat	5	nmod	_	Gloss=[country-name]|SpaceAfter=No
4	惠	惠	PROPN	n, 名詞, 人, その他の人名	NameType=Prs	5	compound	_	Gloss=Hui|SpaceAfter=No
5	王	王	NOUN	n, 名詞, 人, 役割	_	2	obj	_	Gloss=king|SpaceAfter=No

図 1 「孟子見梁惠王」の CoNLL-U データ
Fig. 1 CoNLL-U of “孟子見梁惠王”.

表 2 XPOS から UPOS への変換
Table 2 Conversion of XPOS to UPOS.

XPOS	UPOS	特例	
n,名詞	NOUN	n, 名詞, 人, 姓氏	PROPN
		n, 名詞, 人, 名	
		n, 名詞, 人, その他の人名	
		n, 名詞, 人, 複合的人名	
		n, 名詞, 主体, 国名	
		n, 名詞, 固定物, 地名	
n,代名詞	PRON		
n,数詞	NUM		
v,動詞	VERB		
v,前置詞	ADP		
v,副詞	ADV		
v,助動詞	AUX		
p,助詞	PART	p, 助詞, 接続, 属格	SCONJ
		p, 助詞, 接続, 並列	CCONJ
p,感嘆詞	INTJ		
p,接尾辞	PART		
s,記号	SYM		

表 3 XPOS から FEATS への変換
Table 3 Conversion of XPOS to FEATS.

XPOS	FEATS
n, 名詞, 人, 姓氏	NameType=Sur
n, 名詞, 人, 名	NameType=Giv
n, 名詞, 人, その他の人名	NameType=Prs
n, 名詞, 人, 複合的人名	NameType=Prs
n, 名詞, 主体, 国名	Case=Loc|NameType=Nat
n, 名詞, 固定物, 地名	Case=Loc|NameType=Geo
n, 名詞, 固定物, 地形	Case=Loc
n, 名詞, 固定物, 建造物	Case=Loc
n, 名詞, 固定物, 関係	Case=Loc
n, 名詞, 制度, 場	Case=Loc
n, 名詞, 時	Case=Tem
n, 名詞, 度量衡	NounType=Clf
n, 代名詞, 人称	PronType=Prs
n, 代名詞, 指示	PronType=Dem
n, 代名詞, 疑問	PronType=Int
n, 数詞, 干支	NumType=Ord
v, 動詞, 行為, 分類	Degree=Equ
v, 動詞, 描写	Degree=Pos
v, 副詞, 時相, 過去	AdvType=Tim|Tense=Past
v, 副詞, 時相, 現在	AdvType=Tim|Tense=Pres
v, 副詞, 時相, 将来	AdvType=Tim|Tense=Fut
v, 副詞, 時相, 完了	AdvType=Tim|Aspect=Perf
v, 副詞, 時相	AdvType=Tim
v, 副詞, 疑問, 原因	AdvType=Cau
v, 副詞, 程度, 極度	AdvType=Deg|Degree=Sup
v, 副詞, 程度, やや高度	AdvType=Deg|Degree=Cmp
v, 副詞, 程度, 軽度	AdvType=Deg|Degree=Pos
v, 副詞, 否定	Polarity=Neg
v, 助動詞, 受動	Voice=Pass
v, 助動詞, 可能	Mood=Pot
v, 助動詞, 必要	Mood=Nec
v, 助動詞, 願望	Mood=Des

ス（近似的な逐語英訳）を，「Gloss=各単語のグロス」の形で入れる．なお，MISC における「SpaceAfter=No」は，単語の直後に空白がないことを示している．

2.3 品詞付与

古典中国語形態素解析で用いた 4 階層の品詞体系 [5] は，ほぼそのまま[*4]古典中国語 UD の XPOS に引き継いでいる（図 1）．さらに，4 階層品詞から**表 2** に基づいて，UPOS を自動で導出している．UD は 17 種類の UPOS を規定しているが，古典中国語 UD では表 2 の 14 種類を使用しており，残る 3 種類（ADJ・PUNCT・X）[*5]は使用していない．

また，4 階層品詞から**表 3** に基づいて，FEATS を自動で導出している．ただし「v, 動詞, 存在, 存在」のうち，否定の動詞（「無」など）には Polarity=Neg を，コピュラ動詞（「為」など）には VerbType=Cop を，それぞれ FEATS に入れている．また「n, 代名詞, 人称」のうち，1 人称代名詞には Person=1 を，2 人称代名詞には Person=2 を，3 人称代名詞には Person=3 を，再帰代名詞には Reflex=Yes を，それぞれ FEATS に追加している．

なお，LINDAT/CLARIN との協議により，UD2.5（2019 年 11 月発表）以降，一部の VERB については，以下の変更が行われた．DEPREL が cop の場合には，UPOS を強制的に AUX とする．DEPREL が advmod の場合には，UPOS を強制的に ADV とし，FEATS に VerbForm=Conv を追加する．DEPREL が amod の場合には，FEATS に VerbForm=Part を追加する．これらの措置に，我々は渋々

[*4] 「p, 接尾辞」と「s, 記号, 一般」を追加した．
[*5] ADJ は adjective（形容詞）を意味するが，我々の 4 階層品詞体系は形容詞を廃止している [4]．PUNCT は punctuation（句読点）を意味するが，我々の古典中国語 UD は白文を基にしており，句読点は現れない．X は other（その他の品詞）を意味するが，我々は使用していない．

表 4 古典中国語 UD における係り受けタグ（DEPREL）
Table 4 DEPREL for Classical Chinese UD.

	Nominals	Clauses	Modifier Words	Function Words
Core arguments	nsubj 主語 ↪nsubj:pass [受動文] 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 ↪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

⟨https://universaldependencies.org/u/dep⟩ を拡張

従ったが，必ずしも納得したわけではない [12]．

2.4 係り受け

古典中国語 UD では，**表 4** に示す 38 種類の係り受けタグを，DEPREL に用いている．係り受けタグのうち 32 種類は，もともと UD で規定されているものであり，残り 6 種類（nsubj:pass，obl:tmod，obl:lmod，discourse:sp，compound:redup，flat:vv）は，その拡張形である [10]．root はリンク元を持たない（HEAD を 0 とする）が，それ以外のリンクは，リンク元の単語とリンク先の単語を 1 つずつ有する．リンクはループしない．さらに，古典中国語 UD では，リンクどうしが交差しない（planar），root をまたぐリンクが存在しない（projective），という制限も設けた．

UD の係り受けリンクは，メルチュク依存文法 [13] の後裔であり，いわば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）は，体言の修飾語だと見なす [6]．我々の古典中国語 UD も，これに従っている．たとえば，図 1 の「孟子見梁惠王」においては，動詞「見」を root として，そこから主語「孟子」を nsubj，目的語「王」を obj とするリンクを記述している．さらに「王」からは，「惠」を compound，「梁」を nmod とするリンクを記述しており，全体として**図 2** に示す係り受けツリーを構成している．古典中国語の動賓終構造を記述する際，UD の動詞中心主義は，非常に適切だといえる [14]．

一方，コピュラ文の UD は，補語をリンク元として，主語へとリンクする．古典中国語 UD も同様である．たとえば

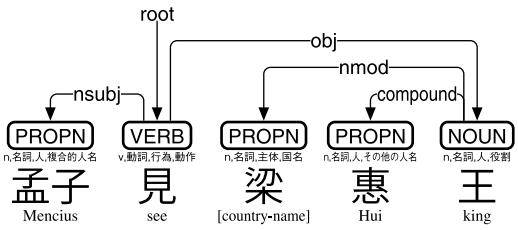


図 2 「孟子見梁惠王」（図 1 参照）の係り受けツリー
Fig. 2 Dependency tree from CoNLL-U of Fig. 1.

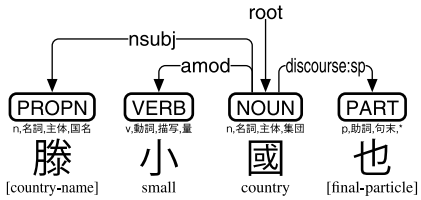


図 3 「滕小國也」の係り受けツリー
Fig. 3 Dependency tree of “滕小國也”.

「滕小國也」というコピュラ文では，補語「國」を root として，そこから主語「滕」を nsubj でリンクしている（**図 3**）．なお，句末の「也」は cop ではなく，discourse:sp でリンク[*6]している．単純なコピュラ文に関しては，古典中国語においても，UD の記述ルールは適切に機能している．

しかし，コピュラ文に関する UD の記述ルールは，補語が節であるような複文を，うまく記述できない．古典中国語 UD における例として，「是民受之也」を見てみよう．この文は「民受之」を P とおくと，「是 P 也」というコピュ

[*6] 現代中国語 UD に倣って [15]，discourse:sp は文助詞（sentence particle）を意味しており，動賓終構造（predicate-object-final structure）の終（final）を記述するのに用いる．

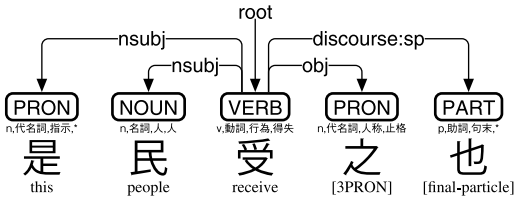


図 4 「是民受之也」の係り受けツリー
Fig. 4 Dependency tree of “是民受之也”.

ラ文である．「民受之」については，動詞「受」を root として，そこから主語「民」を nsubj，目的語「之」を obj とすればよい．また，「是 P 也」については，P から「是」へは nsubj，P から「也」へは discourse:sp となる．しかし，これを図示すると，妙なことになる（**図 4**）．P を代表するのは「受」であることから，「受」から「是」へ nsubj，「受」から「也」へ discourse:sp となり，結果として「受」から 2 本の nsubj が出てしまう．これでは，二重主語文と見分けがつかない．

つまり，補語が節であるようなコピュラ文を，UD は記述しきれておらず，これが UD の限界の 1 つなのである．なお，『孟子』『論語』『禮記』『十八史略』の依存構造コーパス 55514 文のうち，補語が節であるようなコピュラ文は 907 文（約 1.6%）あり，これら 907 文には，UD における記述上の問題[*7]が残されている．

2.5 開発環境とツール群

古典中国語 UD の開発を円滑に行うべく，我々は，独自の GitLab サーバ[*8]を立ち上げている．GitLab サーバには，古典中国語 UD の全データを収録するとともに，ビジュアル・エディタ[*9]を JavaScript と SVG（Scalable Vector Graphics）で実装した．これにより，Web ブラウザ上でマウスなどを用いて，古典中国語 UD データを編集可能である．新規データの採録においては，Kanripo [19] の原テキストを基に，MeCab と UDPipe [20] を組み合わせた UD-Kanbun [*10]によって，初期データを作成している．

なお，古典中国語 UD の開発は，科学研究費補助金基盤研究（B）17H01835『古典漢文形態素コーパスにもとづく動詞の作用域の自動抽出』の研究助成を受けている．

3. 古典中国語 UD に基づく解析システム

UD 依存構造コーパスの目的の 1 つは，UD を用いた文法解析システムの開発を促す [6] ことにある．UD を用いた機械学習などの手段により，書写言語の文切り・単語切り・品詞付与・係り受け解析を行う，というタスクを設定 [23] することで，解析システムが世界中で開発されている．

カレル大学の UDPipe や，スタンフォード大学の Stanza [24] は，60 以上の言語に対して UD を機械学習しており，文切り・単語切り・品詞付与・係り受け解析を，統合的にサポートしている．古典中国語も，UDPipe や Stanza がサポートする言語に含まれている[*11]が，総合的な解析精度では，我々の UD-Kanbun が優って[*12]いた．ところが，オレゴン大学から Trankit [26] が発表された結果，UD-Kanbun の優位は揺らぎ始めた．

Trankit は，XLM-RoBERTa [27]（base）を事前学習モデルとして，文切り・単語切り・品詞付与・係り受け解析の学習に UD2.5 を用いている．UD2.5 には我々の古典中国語 UD が含まれているが，XLM-RoBERTa の 100 言語に古典中国語は含まれていない．そうすると Trankit においては，現代中国語を含む事前学習が，古典中国語の解析にも効いている可能性が考えられる．ならば，事前学習に XLM-RoBERTa ではなく，古典中国語に特化した言語モデル[*13]を使えば，どうなるだろう．

北京理工大学の GuwenBERT [28] [*14]を拡張する形で，我々は，古典中国語 RoBERTa モデル roberta-classical-chinese-base-char を製作し，Transformers [31] の言語モデルとして公開[*15]した．roberta-classical-chinese-base-char は，古典中国語のマスキング（穴埋め）問題を解く言語モデルで，繁体字や簡体字だけでなく，日本の常用漢字にも対応している．この roberta-classical-chinese-base-char を事前学習モデルとして，我々の古典中国語 UD で機械学習（fine-tuning）することを考えた．ただ，Transformers の RoBERTa モデルにおける漢字の扱いは，基本的に 1 文字 =1 単語であり，古典中国語 UD とは単語長が異なっている．このような場合，通常は，長い単語を短く分けることで，単語長の問題を回避する．漢字であれば，複数字の単語を 1 文字ずつに分割して RoBERTa モデルを使うのが，一般的な手法である．

[*7] この問題を解決すべく，我々は，構成鎖（catena）[16] を用いた UD 拡張手法を提案している [17]．しかしながら，我々の拡張提案は，他の各言語 UD からの支持が得られず，UD の公式拡張 ⟨https://universaldependencies.org/u/overview/enhanced-syntax.html⟩ には取り込まれていない．

[*8] ⟨https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun⟩

[*9] deplacy [18] の `deplacy.serve` モジュールとしてパッケージ化を行い，⟨https://pypi.org/project/deplacy⟩ で公開した．

[*10] UD-Kanbun は，単語切り（FORM・LEMMA・MISC の生成）と品詞付与（XPOS・UPOS・FEATS の生成）を MeCab 0.996 で，文切り（ID の生成）と係り受け解析（HEAD・DEPREL の生成）を UDPipe 1.2.0 で，それぞれ行っている [21]．なお，文切りを『古詩文斷句』[22] WebAPI に「外注」する機能も実装しているが，本稿では使用していない．

[*11] UDPipe 1.2.0 は UD2.5 を，Stanza 1.2 は UD2.7（2020 年 11 月発表）を機械学習に用いており，この中に，我々の古典中国語 UD も含まれている．

[*12] UD-Kanbun は，MeCab と UDPipe と StanfordNLP（Stanza の前身）[25] を比較して，それらの「いいとこ取り」で設計しており，優っているのは当然である [21]．

[*13] 予備実験として，GuwenBERT [28][*14]を係り受け解析に用いたところ，UD-Kanbun より解析精度の向上が見られた．そこで，このシステムを GuwenCOMBO と名づけて公開した [29]．

[*14] ⟨https://huggingface.co/ethanyt/guwenbert-base⟩

[*15] ⟨https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-base-char⟩

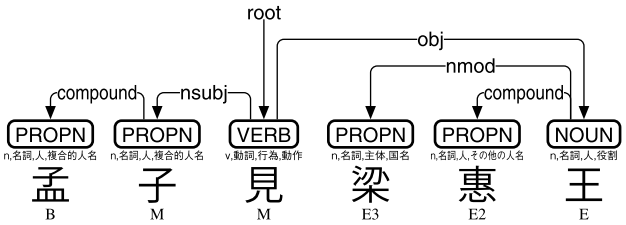


図 5 「孟子見梁惠王」を 1 文字単位に分割
Fig. 5 Character-tokenization of “孟子見梁惠王”.

それなら，古典中国語 UD を最初から 1 文字単位に分割しておいて，それを機械学習した方が，全体としての解析精度は上がる[*16]はずだ．「孟子見梁惠王」であれば，「孟」と「子」の両方に「孟子」の品詞を継承し，それらを compound でつなげば[*17]よい（**図 5**）．この場合，文切り→形態素解析（単語切り＋品詞付与）→係り受け解析，という旧来の解析順序ではなく，文字切り→品詞付与→文切り→係り受け解析→単語組み上げ，という解析順序にするのが，roberta-classical-chinese-base-char には適している．そう，我々は考えた．

このアイデアを基に，我々は，古典中国語解析システム SuPar-Kanbun を製作・公開[*18]した．SuPar-Kanbun は，品詞付与（XPOS・UPOS・FEATS をまとめて生成）を，Transformers 4.0.1（以降）の系列ラベリングで実装しており，古典中国語 UD（1 文字単位版）と roberta-classical-chinese-base-char で学習している．文切りも，B・M・E3・E2・E および S をラベルとする [33] 系列ラベリング[*19]で，同様に実装している．係り受け解析は，SuPar[*20]の Biaffine [35] 実装を用いており，古典中国語 UD（1 文字単位版）と roberta-classical-chinese-base-char で学習している．文字切りと単語組み上げ（同一の XPOS が compound でつながっている場合，それらの文字を 1 語と見なす）は，我々の独自実装である．

大学入学共通テストの令和 3 年度本試験『国語』（2021 年 1 月 16 日実施）第 4 問の問題文を，手作業で UD 化し [29]，そこから白文を抽出[*21]して，文切り・形態素解析・係り受け解析の総合評価を行った．評価指標は，LAS（Labeled Attachment Score）/ MLAS（Morphology-aware Labeled Attachment Score）/ BLEX（Bi-LEXical dependency score）[23] である．**表 5** に評価結果を示す．【問題文 I】（欧陽脩『欧陽文忠公集』より）においては，SuPar-Kanbun が比較的良い精度を示している．【問題文 II】（『韓非子』より）においては，Trankit と SuPar-Kanbun が，それぞれ高い精度を示しているようである．

表 5 共通テスト『国語』第 4 問ベンチマーク（LAS/MLAS/BLEX）
Table 5 LAS/MLAS/BLEX on the Common Test for University Admissions 2021.

	【問題文 I】	【問題文 II】
UDPipe 1.2.0	45.66/35.60/41.88	62.12/50.00/57.41
Stanza 1.2	47.00/41.67/42.71	69.70/60.00/63.64
Trankit 1.0.1	57.53/50.00/54.17	89.39/78.18/83.64
UD-Kanbun 3.1.5	53.21/46.56/48.68	69.70/64.81/66.67
GuwenCOMBO 1.3.8	56.88/50.79/52.91	74.24/67.89/69.72
SuPar-Kanbun 1.1.5	68.81/62.77/67.02	83.33/80.00/80.00

なお，古典中国語 UD にもとづく解析システムの研究（開発ならびに評価）は，科学研究費補助金基盤研究（B）20H04481『古典漢文依存文法コーパスに基づく係り受け構造の自動抽出』の研究助成を受けている．

4. おわりに

我々が製作した『孟子』『論語』『禮記』『十八史略』古典中国語 UD について，その概要を述べた．また，我々の古典中国語 UD を基にした文切り・形態素解析・係り受け解析システムに対し，解析精度の評価を行った．結果として，6 つのシステムのうち，我々の SuPar-Kanbun の解析精度が最も高かった．ただし，これは，あくまで本稿執筆時点でのことである．いずれは，古典中国語を得意とする新たなシステムが現れて，我々を追い抜いていくに違いない．それが研究の広がりというものなのだし，我々としては，それがうれしい．

本稿で示した古典中国語 UD の手法は，日本の変体漢文などにも適用可能だろうと，我々は考えている．ただし，古典中国語の手法を直接に適用するのは無理で，近代日本語における手法 [36] を中古日本語へと拡張し，中古日本語と古典中国語の手法を融合して変体漢文に適用するのが，迂遠だが確実な方法だと思われる．まだまだ道程は長い．今後の我々の研究の進展に期待されたい．

*16 現代日本語の形態素解析に同様のアイデアを用いた例 [32] があり，我々は，これにヒントを得ている．
*17 UD の仕様としては，compound より goeswith の方が適切 [30] だが，リンクが逆方向となってしまう．
*18 ⟨https://pypi.org/project/suparkanbun⟩
*19 ラベリング例を，図 5 の最下段に示す．なお，Transformers の文切りトークン [CLS]・[SEP] は使用せず，複数の文を直接くっつけた形で fine-tuning を行っている．
*20 初期の SuPar 1.0.0 [34] は漢字の扱いに難があったため，修正版を SuPar 1.0.1a1 として，プレリリースしてもらった．
*21 【問題文 I】の白文は「吾有千里馬毛骨何蕭森疾馳如奔風白日無留陰徐駆当大道歩驟中五音馬雖有四足遅速在吾心六轡応吾手調和如瑟琴東西与南北高下山与林惟意所欲適九州可周尋至哉人与馬両楽不相侵伯楽識其外徒知価千金王良得其性此術固已深良馬須善馭吾言可為箴」，【問題文 II】の白文は「凡御之所貴馬体安于車人心調于馬而後可以進速致遠今君後則欲逮臣先則恐逮于臣夫誘道争遠非先則後也而先後心在于臣尚何以調於馬此君之所以後也」である．

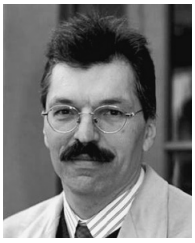
参考文献

[1] Huang, L., Peng, Y., Wang, H. and Wu, Z.: Statistical Part-of-Speech Tagging for Classical Chinese, *Proc. TSD 2002: 5th International Conference on Text, Speech, and Dialogue*, pp.115–122 (2002).
[2] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conference on Empirical Methods*

in Natural Language Processing, pp.230–237 (2004).
[3] 守岡知彦：MeCab を用いた古典中国語形態素解析器の改良，情報処理学会研究報告，Vol.2009-CH-84, No.3, pp.1–5 (2009).
[4] 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，pp.39–46 (2012).
[5] 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, pp.323–331 (2018).
[6] Nivre, J.: Towards a Universal Grammar for Natural Language Processing, *CICLing 2015*: *16th International Conference on Intelligent Text Processing and Computational Linguistics*, pp.3–16 (2015).
[7] 松田 寛，若狭 絢，山下華代，大村 舞，浅原正幸：UD Japanese GSD の再整備と固有表現情報付与，言語処理学会第 26 回年次大会発表論文集，pp.133–136 (2020).
[8] 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 Universal Dependencies への挑戦，情報処理学会研究報告，Vol.2018-CH-116, No.20, pp.1–8 (2018).
[9] 安岡孝一：古典中国語 Universal Dependencies で読む『孟子』，センター研究年報 2018 別冊，京都大学人文科学研究所附属東アジア人文情報学研究センター (2019).
[10] Yasuoka, K.: Universal Dependencies Treebank of the Four Books in Classical Chinese, *DADH2019*: *10th International Conference of Digital Archives and Digital Humanities*, pp.20–28 (2019).
[11] de Marneffe, M.-C., Manning, C.D., Nivre, J. and Zeman, D.: Universal Dependencies, *Computational Linguistics*, Vol.47, No.2, pp.255–308 (2021).
[12] 安岡孝一：漢日英 Universal Dependencies 平行コーパスとその差異，人文科学とコンピュータシンポジウム「じんもんこん 2019」論文集，pp.43–50 (2019).
[13] Mel'čuk, I.A.: *Dependency Syntax: Theory and Practice*, State University of New York Press, New York (1988).
[14] 安岡孝一：Universal Dependencies にもとづく古典中国語（漢文）の依存文法解析，センター研究年報 2018，京都大学人文科学研究所附属東アジア人文情報学研究センター (2018).
[15] Leung, H., Poiret, R., Wong, T., Chen, X., Gerdes, K. and Lee, J.: Developing Universal Dependencies for Mandarin Chinese, *12th Workshop on Asian Language Resources*, pp.20–29 (2016).
[16] Osborne, T., Putnam, M. and Groß, T.: Catenae: Introducing a Novel Unit of Syntactic Analysis, *Syntax*, Vol.15, No.4, pp.354–396 (2012).
[17] 安岡孝一：Universal Dependencies の拡張にもとづく古典中国語（漢文）の直接構成鎖解析の試み，情報処理学会研究報告，Vol.2019-CH-120, No.1, pp.1–8 (2019).
[18] 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，pp.95–100 (2020).
[19] ウィッテルン・クリスティアン：漢籍リポジトリ，センター研究年報 2015，京都大学人文科学研究所附属東アジア人文情報学研究センター (2016).
[20] Straka, M. and Straková, J.: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, *Proc. CoNLL 2017 Shared Task*, pp.88–99 (2017).
[21] 安岡孝一：四書を学んだ MeCab + UDPipe はセンター試験の漢文を読めるのか，東洋学へのコンピュータ利用，第 30 回研究セミナー，pp.3–110 (2019).
[22] 胡 韌奮，李 紳，諸 雨辰：基于深層語言模型的古漢語知識表示及自動断句研究，中文信息学報，Vol.35, No.4, pp.8–15 (2021).
[23] Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J. and Petrov, S.: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, *Proc. CoNLL 2018 Shared Task*, pp.1–21 (2018).
[24] Qi, P., Zhang, Y., Zhang, Y., Bolton, J. and Manning, C.D.: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.101–108 (2020).
[25] Qi, P., Dozat, T., Zhang, Y. and Manning, C.D.: Universal Dependency Parsing from Scratch, *Proc. CoNLL 2018 Shared Task*, pp.160–170 (2018).
[26] Van Nguyen, M., Lai, V.D., Veyseh, A.P.B. and Nguyen, T.H.: Trankit: A Light-Weight Transformer-based Toolkit for Multilingual Natural Language Processing, *EACL 2021*: *16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp.80–90 (2021).
[27] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.8440–8451 (2020).
[28] 閻 覃，遅 澤聞：基于継続訓練的古漢語語言模型，第 19 届中国計算語言学大会 "古聯杯" 古籍文献命名実体識別 (2020).
[29] 安岡孝一：Transformers の BERT は共通テスト『国語』を係り受け解析する夢を見るか，東洋学へのコンピュータ利用，第 33 回研究セミナー，pp.3–34 (2021).
[30] 安岡孝一：世界の Universal Dependencies と係り受け解析ツール群，第 3 回 Universal Dependencies 公開研究会 (2021).
[31] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing, *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp.38–45 (2020).
[32] Tolmachev, A., Kawahara, D. and Kurohashi, S.: Shrinking Japanese Morphological Analyzers With Neural Networks and Semi-supervised Learning, *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol.1, pp.2744–2755 (2019).
[33] 王 博立，史 暁東，蘇 勁松：一種基于循環神経網絡的古文断句方法，北京大学学報（自然科学版），Vol.53, No.2, pp.255–261 (2017).
[34] Zhang, Y., Li, Z. and Zhang, M.: Efficient Second-Order TreeCRF for Neural Dependency Parsing, *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp.3295–3305 (2020).
[35] Dozat, T. and Manning, C.D.: Deep Biaffine Attention for Neural Dependency Parsing, *5th International Conference on Learning Representations*, *C25* (2017).
[36] 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, No.3, pp.1–8 (2020).

安岡 孝一 （正会員）

1965年生．1990年京都大学大学院修士課程修了．京都大学博士（工学）．1990年京都大学大型計算機センター助手．1997年同助教授．2000年京都大学人文科学研究所附属漢字情報研究センター助教授．2009年同所附属東アジア人文情報学研究センター准教授．2015年同教授．人文科学と情報科学の橋渡しをすべく，人文情報学の研究に従事．日本漢字学会理事．電子情報通信学会，電気学会各会員．



ウィッテルン クリスティアン

1962年生．1991年ハンブルク大学修士（漢学），1998年ゲッティンゲン大学博士（哲学）．1998年中華佛學研究所（台北）副教授，中華電子佛典協會の研究・開発担当．2001年京都大学人文科学研究所附属漢字情報研究センター助教授．2009年同所附属東アジア人文情報学研究センター准教授．2012年同教授．文献学的な手法によって漢籍のデジタル・テキストのあるべきすがたを探る．日本デジタル・ヒューマニティーズ学会，国際仏教学学会各会員．

守岡 知彦 （正会員）

1969年生．1999年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了，博士（情報科学）．1999年電子技術総合研究所COE特別研究員．2000年京都大学人文科学研究所附属漢字情報研究センター助手．2009年同所附属東アジア人文情報学研究センター助教．漢字文献を中心とした人文情報学の研究に従事．

池田 巧

1962年生．1990年東京大学修士（中国語学），1993年東京大学大学院博士課程単位取得．山梨県立女子短期大学専任講師，立教大学助教授を経て，1999年京都大学人文科学研究所助教授，2013年同教授．専門は漢藏語方言史研究で，語彙の体系および文構造の記述分析を行っている．

山崎 直樹

1962年生．早稲田大学大学院文学研究科博士前期課程修了．同博士後期課程退学．早稲田大学助手，広島大学専任講師，大阪外国語大学助教授を経て，関西大学外国語学部教授．専門は言語構造の可視化とインストラクション設計．

二階堂 善弘 （正会員）

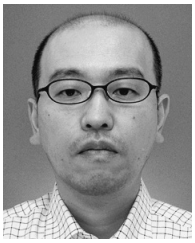
1962年生．1985年東洋大学文学部卒業，1997年早稲田大学大学院文学研究科博士課程退学，博士（文学）・博士（文化交渉学）．1997年東北大学大学院国際文化研究科助手，1998年茨城大学人文学部助教授，2004年関西大学文学部助教授，2005年に同教授．大型電算機プログラマの経験あり．専門は中国の民間信仰であるが，人文情報学においても『電脳中国学入門』（好文出版）等の著作がある．日本道教学会会員等．

鈴木 慎吾

1973年生．2007年大阪外国語大学博士（言語文化学）．2008年京都産業大学外国語学部助教．2011年大阪大学世界言語研究センター講師．2012年同言語文化研究科講師．2021年同准教授．中国語音の歴史的変遷に関する研究のかたわら中国語広東方言の教育に従事．日本中国語学会，日本中国学会，中国語教育学会各会員．

師 茂樹 （正会員）

1972年生．1995年早稲田大学第一文学部卒業，2001年東洋大学大学院文学研究科博士後期課程退学，博士（文化交渉学・関西大学）．現在，花園大学教授．仏教学（唯識思想・仏教論理学）とともに，文字符号化，漢字文献情報処理，文化遺産の3DCG復元等の研究にも取り組む．



藤田 一乘

1972 年生．1992 年龍谷大学文学部卒業．1999 年佛教大学大学院文学研究科修士課程修了．2004 年佛教大学大学院文学研究科博士課程満期退学．博士（文学），現在，佛教大学非常勤講師等．専門は近現代中国文学（胡適，周作人），近代中国の国語の成立，世界語の研究に従事．日本中国学会，東方学会各会員．