

# 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

安岡 孝一<sup>1,a)</sup>

**概要:** 古典中国語(漢文)の解析手法として、Universal Dependenciesの拡張により、構成鎖(catena)解析をおこなう手法を提案する。言語横断的な依存構造記述である Universal Dependencies は、ニューラルネットを用いた言語解析ツール等に採用されており、言語をまたいだ係り受け解析に非常に有用である。しかし、Universal Dependencies は、句構造という考え方をかなり意識的に排除しており、古典中国語の文法構造に適用した場合、この点で、どうしてもいくつかの齟齬が生じてしまう。本稿では、これらの齟齬を出来る限り解決すべく、Universal Dependencies を直接構成鎖解析へと拡張する形で、句構造の導入を試みる。

**キーワード:** 漢文コーパス、構成鎖不可分性、依存文法解析、直接構成素解析、句構造解析

## 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「東アジア古典文献コーパスの実証研究」(班員: 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、李媛、白須裕之、藤田一乗)では、現在、古典中国語(漢文)の依存文法解析に精力を傾注しており、その道具立ての一つとして、Universal Dependencies (以下「UD」)[1]の古典中国語への適用を研究している。依存文法解析それ自体は、Tesnière [2]の構造的統語論に源を発し、Мельчук [3]の有向グラフ記述によって、一応の完成を見た手法である。その最大の特長は、言語横断的な記述が可能だという点にあり、Мельчукの手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が可能としているのである。

ただし、句構造を考慮しない、というUDの特長は、古典中国語の文法解析においては必ずしも利点ではなく、実際、いくつかの齟齬が起こっている。たとえば、「是民受之也」という文をUDで記述すると、見かけ上の二重主語の問題[4]が起こってしまう。この文は「民受之」をXとお

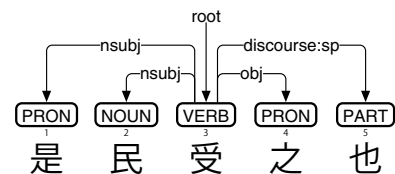


図1 「是民受之也」の古典中国語 UD

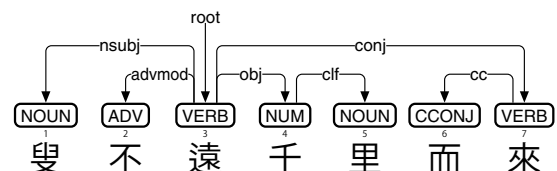


図2 「叟不遠千里而來」の古典中国語 UD

くと、「是X也」というコピュラ文だが、「是」がXの主語であることを表すリンク(nsubj)と、「民」が「受」の主語であることを表すリンクの見分けがつかず、「受」から2本のnsubjが出てるように見える(図1)のである。あるいは、「叟不遠千里而來」(図2)のような「不Y而Z」という形の構文においては、「不」がYだけに係っているのか、それとも「Y而Z」全体を否定しているのか、UDでは必ずしもハッキリしない[5]。これらの齟齬は、そもそもUDが句構造を考慮していないからであり、依存文法解析においては宿命とも言えるのだが、われわれとしては、これらを解決して、さらに先へと進みたい。

本稿では、古典中国語UDに句構造を導入するにあたり、Wells [6]の直接構成素解析(immediate constituent analysis)

<sup>1</sup> 京都大学 Kyoto University

<sup>a)</sup> yasuoka@kanji.zinbun.kyoto-u.ac.jp

表1 古典中国語 UD 依存構造タグ

|                           | Nominals                                                            | Clauses                                                   | Modifier Words                                        | Function Words                                    |
|---------------------------|---------------------------------------------------------------------|-----------------------------------------------------------|-------------------------------------------------------|---------------------------------------------------|
| <b>Core arguments</b>     | nsubj 主語<br>↪nsubj:pass [受動文]<br>obj 目的語<br>iobj 間接目的語              | csubj 節主語<br>↪csubj:pass [受動文]<br>ccomp 節目的語<br>xcomp 節補語 |                                                       |                                                   |
| <b>Non-core arguments</b> | obl 斜格補語<br>vocative 呼称語<br>expl 形式語<br>dislocated 外置語              | advcl 連用修飾節                                               | advmod 連用修飾語<br>discourse 談話要素<br>↪discourse:sp [文助詞] | aux 動詞補助成分<br>cop 繫辞 (copula)<br>mark 標識 (marker) |
| <b>Nominal dependents</b> | nmod 体言による連体修飾語<br>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<br>clf 類別詞<br>case 格表示                    |
| <b>Coordination</b>       | <b>MWE</b>                                                          | <b>Loose</b>                                              | <b>Special</b>                                        | <b>Other</b>                                      |
| conj 接続<br>cc 接続詞         | compound 複合 (endocentric)<br>flat 並列 (exocentric)<br>↪flat:vv [動詞類] | list 細目<br>parataxis 隣接表現                                 | orphan 親なし                                            | punct 句読点<br>root 親                               |

を拡張する形で、UD を変換する手法を提案する。Wells の直接構成素解析は、その後に Chomsky [7] によって歪められてしまったものの、元々は、単語境界における構成素の離れやすさ・離れにくさを扱う手法である。これに対し本稿では、Osborne [8] の構成鎖 (catena) を導入することにより、構成鎖の離れにくさ (不可分性) を全順序関係として定式化する。さらに、その順序付けアルゴリズムによって、古典中国語 UD を構成鎖の解析木へと変換する。

なお、本稿のアイデアの核となる部分は、守岡 [9] の UD 階層化と、Lee [10] の Constraint-based Maximum Entropy Parsing に、その着想を得ている。また、本稿で示すアルゴリズムは、その本質的な部分において、田中 [11] が UD Japanese-KTC で用いた変換手法の「逆回し」である。それぞれに感謝の意を述べたい。

## 2. 古典中国語 UD の直接構成鎖解析への拡張

$n$  個の単語  $w_i$  から成る古典中国語の文  $S=w_1w_2...w_n$  に対し、 $S$  の UD を 5 つ組  $[i, w_i, p_i, h_i, r_i]$  で与える。ただし

$p_i$ :  $w_i$  の品詞 (PROPN・NOUN・PRON・NUM・VERB・ADP・ADV・AUX・PART・INTJ・PROPN・SCONJ・CCONJ・PUNCT・SYM のいずれか [4])

$h_i$ :  $w_i$  を終点とするリンクの始点番号

$r_i$ :  $w_i$  を終点とするリンクのタグ (表 1)

とする。なお、 $r_i=\text{root}$  であるような  $w_i$  は、 $S$  中にただ 1 つとし、 $r_i=\text{root} \Leftrightarrow h_i=0$  とする。また、古典中国語の UD 有向グラフには、リンクのループも交差もない (projective) ものとする。さらに、UD 上での単語  $w_i$  の深さ  $d_i$  と、構成鎖  $C_i$  を、以下のように定義する。

$d_i$ : root から  $w_i$  に至る有向リンク数 (root を含む)

$C_i$ :  $w_i$  を始点とする有向部分木の全単語 ( $w_i$  を含む)

図 1 の UD における各値を表 2 に、図 2 の UD における各

表2 図 1 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$        | $d_i$ | $C_i$ |
|-----|-------|-------|-------|--------------|-------|-------|
| 1   | 是     | PRON  | 3     | nsubj        | 2     | 是     |
| 2   | 民     | NOUN  | 3     | nsubj        | 2     | 民     |
| 3   | 受     | VERB  | 0     | root         | 1     | 是民受之也 |
| 4   | 之     | PRON  | 3     | obj          | 2     | 之     |
| 5   | 也     | PART  | 3     | discourse:sp | 2     | 也     |

表3 図 2 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$  | $d_i$ | $C_i$   |
|-----|-------|-------|-------|--------|-------|---------|
| 1   | 叟     | NOUN  | 3     | nsubj  | 2     | 叟       |
| 2   | 不     | ADV   | 3     | advmod | 2     | 不       |
| 3   | 遠     | VERB  | 0     | root   | 1     | 叟不遠千里而來 |
| 4   | 千     | NUM   | 3     | obj    | 2     | 千里      |
| 5   | 里     | NOUN  | 4     | clf    | 3     | 里       |
| 6   | 而     | CCONJ | 7     | cc     | 3     | 而       |
| 7   | 來     | VERB  | 3     | conj   | 2     | 而來      |

値を表 3 に示す。なお、projective な UD における  $C_i$  は、文  $S$  中においていずれも連続な単語列となり (付録参照)、この場合において  $C_i$  は構成素とみなしうる。

集合  $F = \{C_1, C_2, ..., C_n\}$  上での全順序関係 (等価を含む) が、以下の 4 条件を満たす時、この全順序関係を構成鎖不可分性 (catena inseparability) と呼ぶ。

- a)  $d_j < d_k \Rightarrow C_j < C_k$
- b)  $d_j = d_k$  かつ  $h_j < h_k \Rightarrow C_j < C_k$
- c)  $j < k < h_j = h_k \Rightarrow C_j < C_k$  もしくは  $C_j = C_k$
- d)  $h_j = h_k < j < k \Rightarrow C_j > C_k$  もしくは  $C_j = C_k$

条件 a) は、UD 有向グラフを木構造として見た際に、root からみて「浅い」構成鎖ほど離れやすくすることで、いわゆる幅優先の走査順を保障する条件である。条件 c) と d) は、ある単語から複数のリンクが出ている場合に、リンクが「深い」構成鎖をやや離れやすくする条件である。条件 b) は、これら以外の場合に、文頭に近い構成鎖から離れや

すくする条件である.

文  $S$  と集合  $F = \{C_1, C_2, \dots, C_n\}$  に対して, ある構成鎖不可分性が与えられたならば, その昇順に  $S$  を分割していく作業が, Wells [6] の直接構成素解析にあたる. たとえば表 3 の「叟不遠千里而來」に対し,  $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$  という構成鎖不可分性が(仮に)与えられたならば, この文の直接構成素解析は, Wells の記法で以下のように書ける.

$C_3 \downarrow$  (叟不遠千里而來)  
 叟 不 遠 千 里 而 來  
 $C_1=C_7 \downarrow$  (叟) (而來)  
 叟 | 不 遠 千 里 | 而 來  
 $C_2 \downarrow$  (不)  
 叟 | 不 || 遠 千 里 | 而 來  
 $C_4 \downarrow$  (千里)  
 叟 | 不 || 遠 ||| 千 里 | 而 來  
 $C_5 \downarrow$  (里)  
 叟 | 不 || 遠 ||| 千 |||| 里 | 而 來  
 $C_6 \downarrow$  (而)  
 叟 | 不 || 遠 ||| 千 |||| 里 | 而 || 來

$C_i$  による分割の「残余」を  $\overline{C}_i$  と記すことにすると, 上の  $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$  における残余は, 以下のとおりとなる ( $\phi$  は長さ 0 の単語列).

|                                                 |                             |
|-------------------------------------------------|-----------------------------|
| $\overline{C}_3 = \phi$                         | $\overline{C}_4 = \text{遠}$ |
| $\overline{C}_1 = \overline{C}_7 = \text{不遠千里}$ | $\overline{C}_5 = \text{千}$ |
| $\overline{C}_2 = \text{遠千里}$                   | $\overline{C}_6 = \text{來}$ |

文  $S$  と集合  $F = \{C_1, C_2, \dots, C_n\}$  に対して, ある構成鎖不可分性が与えられた際に, 残余の集合 ( $\phi$  を除く) を  $G = \{\overline{C}_1, \overline{C}_2, \dots, \overline{C}_n\} - \{\phi\}$  とおく. この時,  $F \cup G$  上での (単語列の真部分集合の) 半順序関係  $\supset$  を, 直接構成鎖解析木 (以下, IC 解析木) と呼ぶ. 例として, 表 3 の「叟不遠千里而來」に対し,  $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$  が与えられた場合の IC 解析木を図 3 に示す.

次に、この IC 解析木を、古典中国語 UD と融合することを考えてみよう．端的には、UD 有向グラフの 5 つ組  $[i, w_i, p_i, h_i, r_i]$  のうち、 $w_i$  を、 $C_i$  と  $\overline{C}_i$  ( $\phi$  を除く) に拡張する．それに伴い  $h_i$  を拡張して、 $F \cup G$  上での半順序関係  $\preceq$  に合致させる． $r_i$  は、 $C_i$  を終点とするリンクへ移動し、 $\overline{C}_i$  を終点とするリンクにはタグを付与しない．

このような方法で、図2の「叟不遠千里而來」に対し、表3において図3の条件でIC融合UDを作成(表4)したところ、図4が得られた。「不」が「遠千里」に係っていて、「而來」には係っていないことが、一目瞭然である。同様に、図1の「是民受之也」に対し、表2において $C_3 < C_5 < C_1 < C_2 < C_4$ という構成鎖不可分性でIC融合UDを作成したところ、図5が得られた。「是」が「民受之」の主語(nsubj)であり、「民」が「受之」の主語であることが見て取れる。また、等価を含まない構成鎖不可分性は2分木のIC融合UDを構

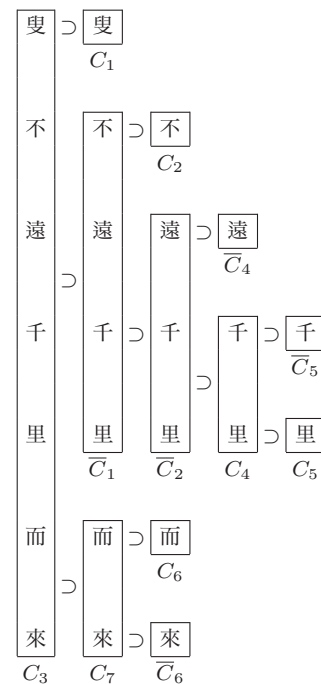


図3 表3の $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$ によるIC解析木

表4 表3の  $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$  による拡張

| $i$ | $w_i$                | $p_i$ | $C_i / \bar{C}_i$ | $h_i$       | $r_i$  |
|-----|----------------------|-------|-------------------|-------------|--------|
| 1   | 叟<br>( $C_1$ )       | NOUN  | 叟                 | $C_3$       | nsubj  |
|     |                      |       | 不遠千里              |             | -      |
| 2   | 不<br>( $C_2$ )       | ADV   | 不                 | $\bar{C}_1$ | advmod |
|     |                      |       | 遠千里               |             | -      |
| 3   | 遠<br>( $\bar{C}_4$ ) | VERB  | 叟不遠千里而來           | 0           | root   |
|     |                      |       | $\phi$            |             | -      |
| 4   | 千<br>( $\bar{C}_5$ ) | NUM   | 千里                | $\bar{C}_2$ | obj    |
|     |                      |       | 遠                 |             | -      |
| 5   | 里<br>( $C_5$ )       | NOUN  | 里                 | $C_4$       | clf    |
|     |                      |       | 千                 |             | -      |
| 6   | 而<br>( $C_6$ )       | CCONJ | 而                 | $C_7$       | cc     |
|     |                      |       | 來                 |             | -      |
| 7   | 來<br>( $\bar{C}_6$ ) | VERB  | 而來                | $C_3$       | conj   |
|     |                      |       | ( $\bar{C}_1$ )   |             | -      |

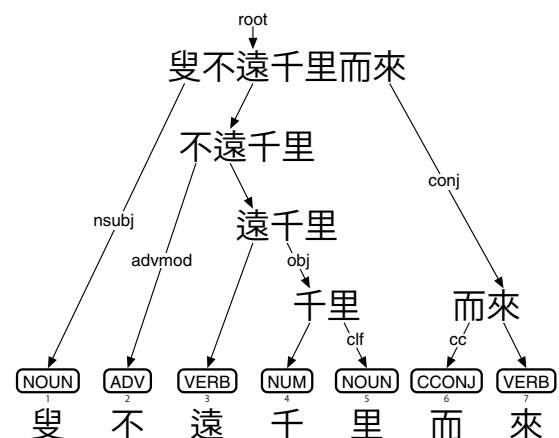


図4 表4にもとづくIC融合UD

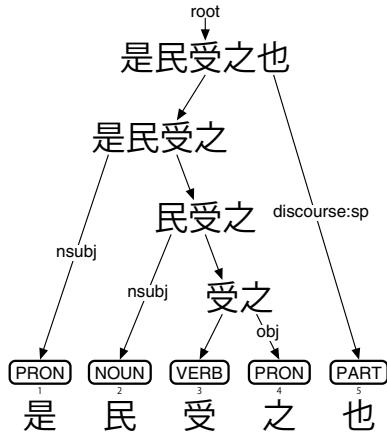


図5 表2の  $C_3 < C_5 < C_1 < C_2 < C_4$  による IC 融合 UD

成し、等価を含む構成鎖不可分性は3分木以上となることも、それぞれ図5・図4から理解できる。

ちなみに、IC 融合 UD から元の UD への逆変換は、以下の手順でおこなうことができる。

- (1) IC 融合 UD において、タグの無いリンク ( $\bar{C}_i$  へのリンク) を、双方向リンクに置き換える。
- (2) 構成鎖のうち、単語  $w_i$  以外のノードを除去する。その際に、各単語への到達性を維持する形で、リンクを縮退する。

図5から図1へ、図4から図2へ、それぞれ逆変換できることを確認されたい。

### 3. 構成鎖不可分性の順序付けアルゴリズム

前章の議論を要約すると、projective な UD に対して、適切な構成鎖不可分性を与えることができれば、直接構成要素解析をおこなったのと同様の結果が得られる、ということである。では、適切な構成鎖不可分性を与える順序付けアルゴリズムは、実際に書くことができるのか。構成鎖不可分性の4条件にしたがって、 $C_j$  と  $C_k$  の順序付けアルゴリズム Inseparability を、python 風に書いてみよう(図6)。

実のところ、図6が判定しているのは条件a)とb)のみで、条件c)は InseparabilityLL に、条件d)は InseparabilityRR に、その他 ( $j < h_j = h_k < k$ ) は InseparabilityLR に、それぞれ押しつけている。というのも、InseparabilityLL・RR・LR は、対象言語の語順の自由さにより、異なる可能性が高い。

図7では、古典中国語 UD に限定した上で、とりあえず  $r_j$  と  $r_k$  のみに着目して、仮に InseparabilityLL・RR・LR を書いてみた。仮のアルゴリズムではあるものの、表2に対しては  $C_3 < C_5 < C_1 < C_2 < C_4$  という構成鎖不可分性が、表3に対しては  $C_3 < C_1 = C_7 < C_2 < C_4 < C_5 < C_6$  という構成鎖不可分性が、それぞれ図6の Inseparability で得られる。

ただし、現時点での図6・7の順序付けアルゴリズムは、もちろん完璧ではないし万能でもない。たとえば、「孔子不得中道而與之」(図8・表5)という「不Y而Z」形の文

```
def Inseparability(j, k) :
    if d_j < d_k :
        return 「C_j < C_k」
    if d_j > d_k :
        return 「C_k < C_j」
    if h_j < h_k :
        return 「C_j < C_k」
    if h_j > h_k :
        return 「C_k < C_j」
    if j < k < h_j = h_k :
        return InseparabilityLL(j, k)
    if k < j < h_j = h_k :
        return InseparabilityLL(k, j)
    if h_j = h_k < j < k :
        return InseparabilityRR(j, k)
    if h_j = h_k < k < j :
        return InseparabilityRR(k, j)
    if j < k :
        return InseparabilityLR(j, k)
    if k < j :
        return InseparabilityLR(k, j)
    return 「C_j = C_k」
```

図6 順序付けアルゴリズム Inseparability(j, k)

```
def InseparabilityLL(j, k) :
    if r_j = compound and r_k = compound :
        if ∃ i ただし h_i = h_j かつ j < i < k :
            return (InseparabilityLL(j, i) かつ InseparabilityLL(i, k))
        return 「C_j = C_k」
    return 「C_j < C_k」

def InseparabilityRR(j, k) :
    for x in [conj, flat, list, parataxis, discourse:sp] :
        if r_j = x and r_k = x :
            if ∃ i ただし h_i = h_j かつ j < i < k :
                return (InseparabilityRR(j, i) かつ InseparabilityRR(i, k))
            return 「C_j = C_k」
    return 「C_k < C_j」

def InseparabilityLR(j, k) :
    for x in [case, mark, parataxis, discourse:sp, punct] :
        if r_k = x :
            return 「C_k < C_j」
    if r_j = nsubj and r_k = conj :
        if ∃ i ただし r_i = nsubj かつ j < i < h_i = h_j :
            return 「C_j < C_k」
        return 「C_j = C_k」
    return 「C_j < C_k」
```

図7 古典中国語 UD 向け (仮) InseparabilityLL・RR・LR

を見てみよう。この文においては、「不」が「Y而Z」全体に係っていることから、 $C_3 < C_1 < C_2 < C_7 < C_5 < C_4 < C_6 < C_8$  という構成鎖不可分性が適切だと考えられる(図9)。しかしながら、現時点の図6・7の順序付けアルゴリズムは、このような構成鎖不可分性を出力できない。代わりに得られるのは、 $C_3 < C_1 = C_7 < C_2 < C_5 < C_4 < C_6 < C_8$  という「不」が「得中道」にだけ係った構成鎖不可分性である。

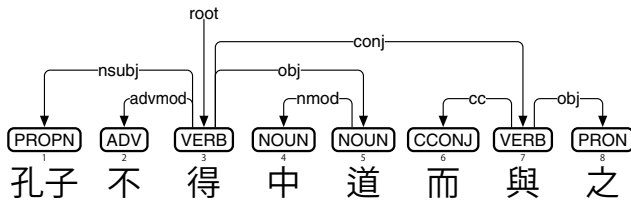


図8 「孔子不得中道而與之」の古典中国語 UD

表5 図8のUDにおける  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$  | $d_i$ | $C_i$     |
|-----|-------|-------|-------|--------|-------|-----------|
| 1   | 孔子    | PROPN | 3     | nsubj  | 2     | 孔子        |
| 2   | 不     | ADV   | 3     | advmod | 2     | 不         |
| 3   | 得     | VERB  | 0     | root   | 1     | 孔子不得中道而與之 |
| 4   | 中     | NOUN  | 5     | nmod   | 3     | 中         |
| 5   | 道     | NOUN  | 3     | obj    | 2     | 中道        |
| 6   | 而     | CCONJ | 7     | cc     | 3     | 而         |
| 7   | 與     | VERB  | 3     | conj   | 2     | 而與之       |
| 8   | 之     | PRON  | 7     | obj    | 3     | 之         |

これは、図7のアルゴリズムにおいて、 $r_j$  と  $r_k$  のみに着目しているための限界であり、図9と図4の間の問題を解決するためには、他の要素にも着目する必要があるということである。しかしながら、他の要素にも着目した場合、図7のアルゴリズムを解析的に記述するのは、かなり膨大な作業が予想される。むしろ、機械学習などの手法によって、InseparabilityLL・LR・RRを構成するやり方に、挑戦すべきだろう。

#### 4. おわりに

古典中国語 UD に構成鎖不可分性を導入し、構成鎖による解析を試みた。これにより、古典中国語 UD に句構造を導入する目途が立ったといえる。さらに、構成鎖不可分性

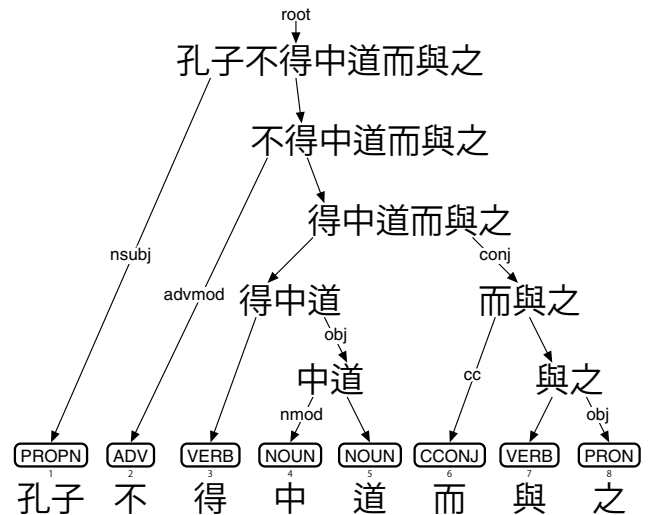


図9 表5の  $C_3 < C_1 < C_2 < C_7 < C_5 < C_4 < C_6 < C_8$  によるIC融合UD

を自動生成するための順序付けアルゴリズムを試作し、その有効性を検証した。ただし、アルゴリズム全体を解析的に書くのは困難が伴う、という点が課題として残されている。また、本研究の副産物として、UDの公式フォーマットである CoNLL-U の拡張をおこない、コメント行に構成鎖不可分性 (catena inseparability) を、そのまま記載する方法 (図10) を考案した。

本稿の手法は、リンクに交差がない古典中国語 UD において有効であり、他の言語の UD においても、リンクに交差がない場合は同様に有効だと考えられる。一方、リンクに交差がある UD に対しては、解析木にも交差が起こってしまうことを、付録に示す。

なお、本研究は、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。

|                                           |    |    |       |              |                       |   |              |   |                                      |
|-------------------------------------------|----|----|-------|--------------|-----------------------|---|--------------|---|--------------------------------------|
| # text = 是民受之也                            |    |    |       |              |                       |   |              |   |                                      |
| # catena_inseparability = 3<5<1<2<4       |    |    |       |              |                       |   |              |   |                                      |
| 1                                         | 是  | 是  | PRON  | n,代名詞,指示,*   | PronType=Dem          | 3 | nsubj        | - | Gloss=this SpaceAfter=No             |
| 2                                         | 民  | 民  | NOUN  | n,名詞,人,人     | -                     | 3 | nsubj        | - | Gloss=people SpaceAfter=No           |
| 3                                         | 受  | 受  | VERB  | v,動詞,行為,得失   | -                     | 0 | root         | - | Gloss=receive SpaceAfter=No          |
| 4                                         | 之  | 之  | PRON  | n,代名詞,人称,止格  | Person=3 PronType=Prs | 3 | obj          | - | Gloss=[3PRON] SpaceAfter=No          |
| 5                                         | 也  | 也  | PART  | p,助詞,句末,*    | -                     | 3 | discourse:sp | - | Gloss=[final-particle] SpaceAfter=No |
| # text = 叟不遠千里而來                          |    |    |       |              |                       |   |              |   |                                      |
| # catena_inseparability = 3<1=7<2<4<5<6   |    |    |       |              |                       |   |              |   |                                      |
| 1                                         | 叟  | 叟  | NOUN  | n,名詞,人,人     | -                     | 3 | nsubj        | - | Gloss=old-gentleman SpaceAfter=No    |
| 2                                         | 不  | 不  | ADV   | v,副詞,否定,無界   | Polarity=Neg          | 3 | advmod       | - | Gloss=not SpaceAfter=No              |
| 3                                         | 遠  | 遠  | VERB  | v,動詞,描写,量    | Degree=Pos            | 0 | root         | - | Gloss=distant SpaceAfter=No          |
| 4                                         | 千  | 千  | NUM   | n,数詞,数字,*    | -                     | 3 | obj          | - | Gloss=thousand SpaceAfter=No         |
| 5                                         | 里  | 里  | NOUN  | n,名詞,度量衡,*   | NounType=Class        | 4 | clf          | - | Gloss=[distance-unit] SpaceAfter=No  |
| 6                                         | 而  | 而  | CCONJ | p,助詞,接続,並列   | -                     | 7 | cc           | - | Gloss=and SpaceAfter=No              |
| 7                                         | 來  | 來  | VERB  | v,動詞,行為,移動   | -                     | 3 | conj         | - | Gloss=come SpaceAfter=No             |
| # text = 孔子不得中道而與之                        |    |    |       |              |                       |   |              |   |                                      |
| # catena_inseparability = 3<1<2<7<5<4<6<8 |    |    |       |              |                       |   |              |   |                                      |
| 1                                         | 孔子 | 孔子 | PROPN | n,名詞,人,複合的人名 | NameType=Prs          | 3 | nsubj        | - | Gloss=Confucius SpaceAfter=No        |
| 2                                         | 不  | 不  | ADV   | v,副詞,否定,無界   | Polarity=Neg          | 3 | advmod       | - | Gloss=not SpaceAfter=No              |
| 3                                         | 得  | 得  | VERB  | v,動詞,行為,得失   | -                     | 0 | root         | - | Gloss=get SpaceAfter=No              |
| 4                                         | 中  | 中  | NOUN  | n,名詞,固定物,關係  | Case=Loc              | 5 | nmod         | - | Gloss=centre SpaceAfter=No           |
| 5                                         | 道  | 道  | NOUN  | n,名詞,制度,儀礼   | -                     | 3 | obj          | - | Gloss=doctrine SpaceAfter=No         |
| 6                                         | 而  | 而  | CCONJ | p,助詞,接続,並列   | -                     | 7 | cc           | - | Gloss=and SpaceAfter=No              |
| 7                                         | 與  | 與  | VERB  | v,動詞,行為,交流   | -                     | 3 | conj         | - | Gloss=participate SpaceAfter=No      |
| 8                                         | 之  | 之  | PRON  | n,代名詞,人称,止格  | Person=3 PronType=Prs | 7 | obj          | - | Gloss=[3PRON] SpaceAfter=No          |

図10 CoNLL-U フォーマットの catena inseparability 拡張

## 付録 他言語 UD への応用

本稿の手法に関し、他の言語の UD への応用について、以下では考察してみよう。図 11 に示したのは、Chomsky [7] の「Colorless green ideas sleep furiously」という英文を、StanfordNLP [12] の en\_ewt モデル 0.1.0 で依存文法解析した結果の UD である。

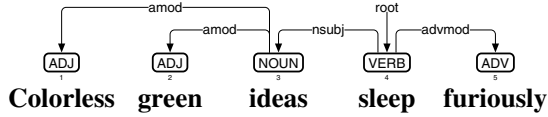


図 11 「Colorless green ideas sleep furiously」の英語 UD

表 6 図 11 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$     | $p_i$ | $h_i$ | $r_i$  | $d_i$ | $C_i$                                 |
|-----|-----------|-------|-------|--------|-------|---------------------------------------|
| 1   | Colorless | ADJ   | 3     | amod   | 3     | Colorless                             |
| 2   | green     | ADJ   | 3     | amod   | 3     | green                                 |
| 3   | ideas     | NOUN  | 4     | nsubj  | 2     | Colorless green ideas                 |
| 4   | sleep     | VERB  | 0     | root   | 1     | Colorless green ideas sleep furiously |
| 5   | furiously | ADV   | 4     | advmod | 2     | furiously                             |

図 11 の UD 有向グラフにおいて、5 つ組  $[i, w_i, p_i, h_i, r_i]$  と深さ  $d_i$  と構成鎖  $C_i$  は、表 6 のようになっている。この  $C_i$  に図 6・7 の順序付けアルゴリズムを(そのまま)適用すると、現状では  $C_4 < C_3 < C_5 < C_1 < C_2$  という構成鎖不可分性が得られる。この構成鎖不可分性にしがうと、 $\bar{C}_i$  は

$$\begin{aligned}\bar{C}_4 &= \phi \\ \bar{C}_3 &= \text{sleep furiously} & \bar{C}_1 &= \text{green ideas} \\ \bar{C}_5 &= \text{sleep} & \bar{C}_2 &= \text{ideas}\end{aligned}$$

となり、IC 融合 UD を作成すると、表 7・図 12 のようになる。筆者の見る限り、この「Colorless green ideas sleep furiously」という英文に対し、本稿の手法は正しく(機械的に)直接構成素解析をおこなうことができており、図 12 の IC 融合 UD は Chomsky [7] の句構造による文法木と同型である。

表 7 表 6 の  $C_4 < C_3 < C_5 < C_1 < C_2$  による拡張

| $i$ | $w_i$                    | $p_i$ | $C_i / \bar{C}_i$                        | $h_i$       | $r_i$  |
|-----|--------------------------|-------|------------------------------------------|-------------|--------|
| 1   | Colorless<br>( $C_1$ )   | ADJ   | Colorless                                | $C_3$       | amod   |
|     | green                    |       | green ideas                              |             | -      |
| 2   | green<br>( $C_2$ )       | ADJ   | green                                    | $\bar{C}_1$ | amod   |
|     | ideas                    |       | ideas                                    |             | -      |
| 3   | ideas<br>( $\bar{C}_2$ ) | NOUN  | Colorless green ideas                    | $C_4$       | nsubj  |
|     | sleep furiously          |       | sleep furiously                          |             | -      |
| 4   | sleep<br>( $\bar{C}_5$ ) | VERB  | Colorless green ideas<br>sleep furiously | 0           | root   |
|     |                          |       | $\phi$                                   | -           | -      |
| 5   | furiously<br>( $C_5$ )   | ADV   | furiously                                | $\bar{C}_3$ | advmod |
|     |                          |       | sleep                                    |             | -      |

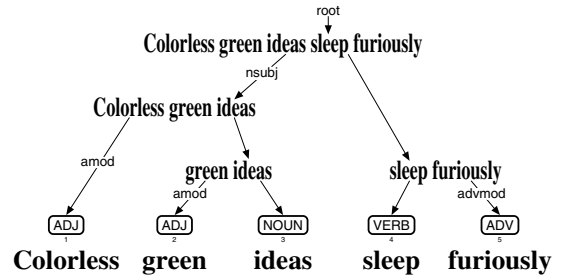


図 12 表 7 にもとづく IC 融合 UD

ただし、英語 UD は、このようなスジのいい例ばかりではない。われわれの古典中国語 UD は、リンクの交差を許していないが、英語 UD においては、リンクの交差が起こる(non-projective) 場合があるのだ。図 13 は、「I rode a horse here which had no name」という英文を、同様に StanfordNLP で依存文法解析した結果の UD である。「here」へのリンクと「had」へのリンクが、交差しているのが見て取れる。

この場合、本稿の手法にしたがって  $C_i$  を導出すると、表 8 に示すとおり  $C_4$  が「a horse which had no name」となってしまう。文中において不連続である。すなわち、non-projective な UD においては、 $C_i$  が不連続な単語列になりうる。一般的な文法解析において、不連続な単語列による要素を許すかどうかについては、そもそも Wells [6] においても議論があるが、ここでは不連続な単語列も構成鎖として扱うことにしよう。

不連続な単語列を許した上で、表 8 の  $C_i$  に図 6・7 の順序付けアルゴリズムを(そのまま)適用すると、現状では  $C_2 < C_1 < C_5 < C_4 < C_3 < C_7 < C_6 < C_9 < C_8$  という構成鎖不可分性が得られる。これによって、図 14 の IC 融合 UD が得られるが、解析木にも交差が発生してしまう。このような交差のある解析木を許すかどうかは、解析対象の性質に

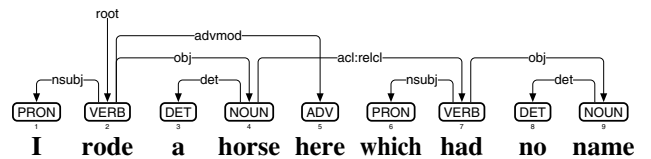


図 13 リンクに交差がある英語 UD の例

表 8 図 13 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$     | $d_i$ | $C_i$                                 |
|-----|-------|-------|-------|-----------|-------|---------------------------------------|
| 1   | I     | PRON  | 2     | nsubj     | 2     | I                                     |
| 2   | rode  | VERB  | 0     | root      | 1     | I rode a horse here which had no name |
| 3   | a     | DET   | 4     | det       | 3     | a                                     |
| 4   | horse | NOUN  | 2     | obj       | 2     | a horse which had no name             |
| 5   | here  | ADV   | 2     | advmod    | 2     | here                                  |
| 6   | which | PRON  | 7     | nsubj     | 4     | which                                 |
| 7   | had   | VERB  | 4     | acl:relcl | 3     | which had no name                     |
| 8   | no    | DET   | 9     | det       | 5     | no                                    |
| 9   | name  | NOUN  | 7     | obj       | 4     | no name                               |

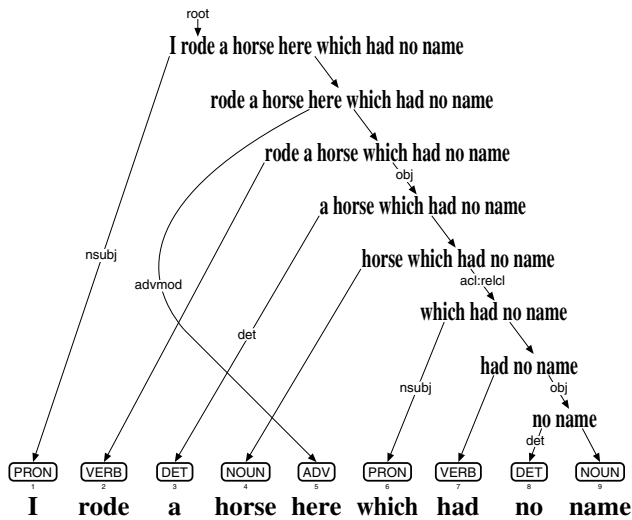


図 14 表 8 の IC 融合 UD 例 ( $C_2 < C_1 < C_5 < C_4 < C_3 < C_7 < C_6 < C_9 < C_8$ )

もよる [13] が、議論の余地があるだろう。

日本語 UD においても、やや稀ではあるものの、リンクに交差がある文例 (図 15) が存在する。交差がない文例 (図 16) と比較検討してみよう。

図 15 の UD に対し、本稿の手法にしたがって  $C_i$  を導出すると、表 9 に示すとおり  $C_5$  が「これが正しい」となってしまう、文中において不連続である。一方、図 16 の UD に対し、本稿の手法にしたがって  $C_i$  を導出すると、やはり  $C_5$  が「これが正しい」となる (表 10) が、こちらは文中において連続している。すなわち、不連続な単語列となる  $C_i$  が現れるのは、UD のリンクに交差がある場合に限定されるといえる。

不連続な単語列を許した上で、表 9 の  $C_i$  に図 6・7 の順序付けアルゴリズムを (そのまま) 適用すると、現状では  $C_7 < C_3 < C_5 < C_4 < C_6 < C_1 < C_2$  という構成鎖不可分性が得られる。これによって、図 17 の IC 融合 UD が得られるが、やはり解析木にも交差が発生してしまう。一方、表 10 の  $C_i$  に図 6・7 の順序付けアルゴリズムを (そのまま) 適用

表 9 図 15 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$ | $d_i$ | $C_i$            |
|-----|-------|-------|-------|-------|-------|------------------|
| 1   | これ    | PRON  | 5     | nsubj | 3     | これが <sup>s</sup> |
| 2   | が     | ADP   | 1     | case  | 4     | が <sup>s</sup>   |
| 3   | 僕     | PRON  | 7     | nsubj | 2     | 僕は               |
| 4   | は     | ADP   | 3     | case  | 3     | は                |
| 5   | 正しい   | ADJ   | 7     | ccomp | 2     | これが正しいと          |
| 6   | と     | ADP   | 5     | mark  | 3     | と                |
| 7   | 思う    | VERB  | 0     | root  | 1     | これが僕は正しいと思う      |

表 10 図 16 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$ | $d_i$ | $C_i$            |
|-----|-------|-------|-------|-------|-------|------------------|
| 1   | 僕     | PRON  | 7     | nsubj | 2     | 僕は               |
| 2   | は     | ADP   | 1     | case  | 3     | は                |
| 3   | これ    | PRON  | 5     | nsubj | 3     | これが <sup>s</sup> |
| 4   | が     | ADP   | 3     | case  | 4     | が <sup>s</sup>   |
| 5   | 正しい   | ADJ   | 7     | ccomp | 2     | これが正しいと          |
| 6   | と     | ADP   | 5     | mark  | 3     | と                |
| 7   | 思う    | VERB  | 0     | root  | 1     | 僕はこれが正しいと思う      |

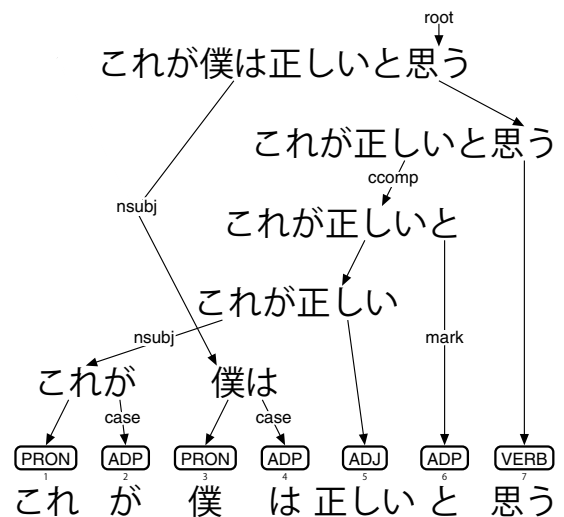


図 17 表 9 の IC 融合 UD 例 ( $C_7 < C_3 < C_5 < C_4 < C_6 < C_1 < C_2$ )

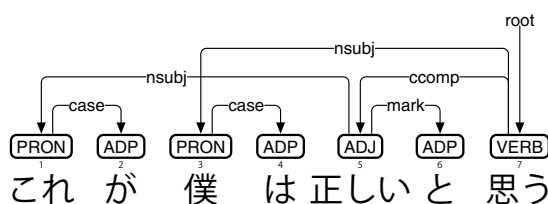


図 15 リンクに交差がある日本語 UD の例

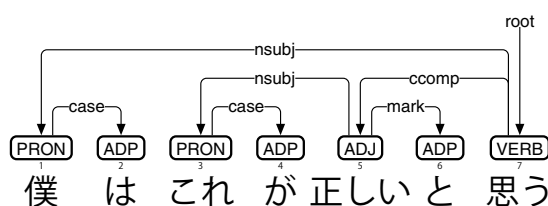


図 16 「僕はこれが正しいと思う」の日本語 UD

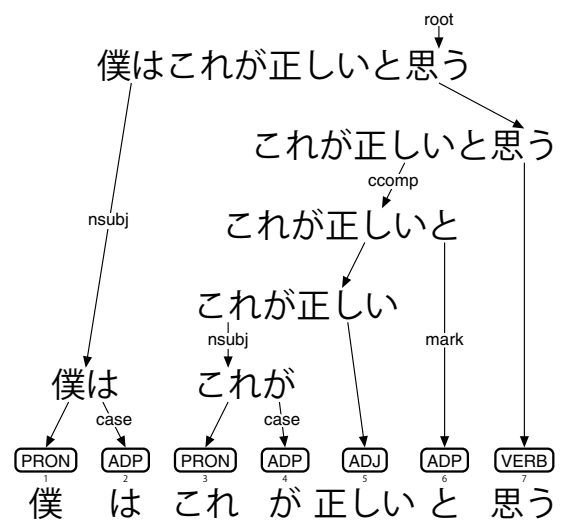


図 18 表 10 の IC 融合 UD 例 ( $C_7 < C_1 < C_5 < C_2 < C_6 < C_3 < C_4$ )

すると、現状では  $C_7 < C_1 < C_5 < C_2 < C_6 < C_3 < C_4$  という構成鎖不可分性が得られる。これによって、図 18 の IC 融合 UD が得られるが、こちらの解析木に交差は無い。

ここで、図 17 と図 18 の IC 融合 UD を比較してみると、語順を除いて同型とみなせる。いずれも妥当な解析結果だと考えられるが、図 17 の交差の問題は残る。

ドイツ語 UD においては、また別の問題が起こりうる。図 19 に示したのは、「Er sieht sehr gut aus」という独文を、StanfordNLP の de\_gsd モデル 0.1.0 で依存文法解析した結果の UD である。この文の「sieht」と「aus」は、動詞「aussehen」の 3 人称単数現在形「aussieht」が分離したものだと解され、compound:prt という特殊なリンクで繋がれている。

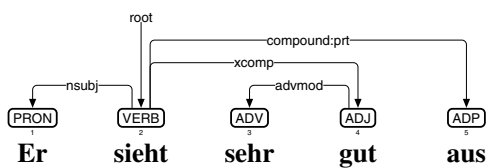


図 19 「Er sieht sehr gut aus」のドイツ語 UD

表 11 図 19 の UD における  $[i, w_i, p_i, h_i, r_i]$  および  $d_i, C_i$

| $i$ | $w_i$ | $p_i$ | $h_i$ | $r_i$        | $d_i$ | $C_i$                 |
|-----|-------|-------|-------|--------------|-------|-----------------------|
| 1   | Er    | PRON  | 2     | nsubj        | 2     | Er                    |
| 2   | sieht | VERB  | 0     | root         | 1     | Er sieht sehr gut aus |
| 3   | sehr  | ADV   | 4     | advmod       | 3     | sehr                  |
| 4   | gut   | ADJ   | 2     | xcomp        | 2     | sehr gut              |
| 5   | aus   | ADP   | 2     | compound:prt | 2     | aus                   |

本稿の手法にしたがって  $C_i$  を導出すると、表 11 のようになる。ここで「aussieht」を考慮すると、 $C_2 < C_1 < C_4 < C_5 < C_3$  という構成鎖不可分性が、分離動詞を表現する点では妥当だと考えられる (図 20)。だが、この構成鎖不可分性の  $C_4 < C_5$  は、条件 d) に違反しており、本稿の手法では導出できない。

表 11 を構成鎖不可分性の 4 条件に適合させるならば、たとえば  $C_2 < C_1 < C_4 = C_5 < C_3$  が考えられる。しかしながら、図 21 の IC 融合 UD が「Er sieht sehr gut aus」の解析結果として妥当かどうかは、非常に疑問が残る。ドイツ語

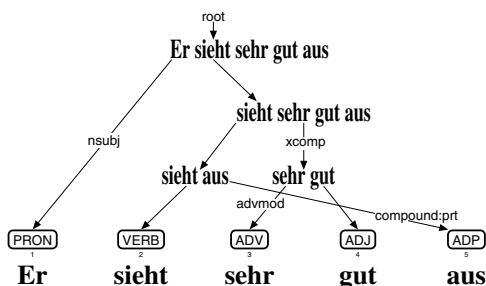


図 20 表 11 の IC 融合 UD 例 ( $C_2 < C_1 < C_4 < C_5 < C_3$ )

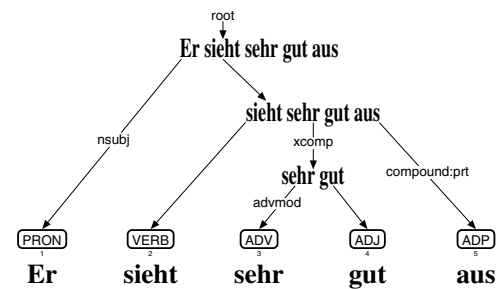


図 21 表 11 の IC 融合 UD 例 ( $C_2 < C_1 < C_4 = C_5 < C_3$ )

UD に対しては、図 20 の IC 融合 UD が導出できるよう、構成鎖不可分性の 4 条件を緩和すべきかもしれない。

本稿の手法を、英語 UD・日本語 UD・ドイツ語 UD に応用する場合について、ざっと概観した。古典中国語 UD とは異なる問題があり、どうやら言語ごとに手法を変えざるを得ない、というところまでは判明したものの、具体的な変更点には立ち入ることができなかった。これらについては、またいずれ、稿を改めて議論したい。

## 参考文献

- [1] 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.
- [2] Lucien Tesnière: Éléments de Syntaxe Structurale, Paris: C. Klincksieck (1959).
- [3] Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).
- [4] 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 (2018 年 10 月).
- [5] 安岡孝一: 漢文の依存文法解析と返り点の関係について, 日本漢字学会第 1 回研究大会予稿集 (2018 年 12 月), pp.33-48.
- [6] Rulon S. Wells: Immediate Constituents, Language, Vol.23, No.2 (April-June 1947), pp.81-117.
- [7] Noam Chomsky: Syntactic Structures, Hague: Mouton (1957).
- [8] Timothy Osborne, Michael Putnam, Thomas Groß: Catenae: Introducing a Novel Unit of Syntactic Analysis, Syntax, Vol.15, No.4 (December 2012), pp.354-396.
- [9] 守岡知彦: 古典中国語 UD コーパスの IPFS を用いた表現の試み, 情報処理学会研究報告, Vol.2018-CH-118 (2018 年 8 月), No.6, pp.1-7.
- [10] Young-Suk Lee, Zhiguo Wang: Language Independent Dependency to Constituent Tree Conversion, Proceedings of COLING 2016: the 26th International Conference on Computational Linguistics (December 2016), pp.421-428.
- [11] 田中貴秋: UD Japanese-KTC: 京大コーパス句構造版からの Universal Dependencies 化, 第 1 回 Universal Dependencies 公開研究会 (2018 年 6 月).
- [12] Peng Qi, Timothy Dozat, Yuhao Zhang, Christopher D. Manning: Universal Dependency Parsing from Scratch, Proceedings of the CoNLL 2018 Shared Task (October 2018), pp.160-170.
- [13] Marco Kuhlmann: Mildly Non-Projective Dependency Grammar, Computational Linguistics, Vol.39, No.2 (June 2013), pp.355-387.