古典中国語（漢文）の依存文法解析と直接構成素解析

安岡　孝一（やすおか　こういち）

❋ はじめに

言語処理という側面から見ると、古典中国語（漢文）の白文というのは、かなりやっかいなシロモノである。単語と単語の間に区切りがない。文と文の間にも区切りがない。漢字がズラズラと、切れ目なく並ぶだけである。こんなもの、どうやって読めばいいのか。

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」では、現在、古典中国語における文法解析の自動化に全力で取り組んでいる。漢文の白文に対し、形態素解析・依存文法解析・直接構成素解析を順におこなうことで、白文の統語構造が解析可能となる、というのが、われわれの見通しである。形態素解析 [1] によって、単語切りをおこなうと同時に、各単語の品詞を得る。依存文法解析 [2] によって、単語と単語の間の係り受け関係を解析すると同時に、文の切れ目を得る。直接構成素解析 [3] によって、各文の統語構造を解析木の形で得る。例として「孟子見梁惠王王曰叟不遠千里而來」という白文に対し、形態素解析・依存文法解析・直接構成素解析を順におこなう際の流れ（イメージ図）を、図 1 に示す。

❋ 漢文の形態素解析

古典中国語の形態素解析 [4] において、われわれは、MeCab という汎用の形態素解析ソフトウェアを用いることにした。MeCab は、もともとは日本語の形態素解析用だった [1] が、言語、辞書、コーパスに依存しない汎用的な設計がなされており、辞書とコーパスを準備すればいかなる言語にも対応できる、というのが売りである。

MeCab の辞書には「品詞」（複数の階層が可能）が必要なことから、われわれは、日本語と漢文をつなぐ「構造」の一種である訓読に着目し、返り点を「品詞」に反映させることを考えた。すなわち、訓読における返り点が、漢文の動賓構造を表しているとみなし、動詞類に「v」という「品詞」を、賓語に「n」という「品詞」を、その他の語に「p」という「品詞」を、それぞれ、MeCab 漢文辞書の「第 1 階層の品詞」として定めた。次に「第 2 階層の品詞」を「名詞」「代名詞」「数詞」「動詞」「前置詞」「副詞」「助動詞」「助詞」「感嘆詞」の 9 種類とした。従来の漢文文法などで見られた「形容詞」を廃止して、「動詞」と統合している [5] のが特徴である。さらに「第 3 階層の品詞」として 44 種類の意味素性を、「第 4 階層の品詞」として 88 種類の小素性を定義し、形態素解析の結果として得られる各単語を、意味の面からも捉えやすいよう工夫した。

MeCab を用いた形態素解析において、その中心となるアイデアは、CRF（Conditional Random Fields）である。われわれの漢文形態素解析に即して言えば、解析したい白文を MeCab 漢文辞書に基いて、可能性のある全ての単語（4 階層の品詞を含む）の組み合わせの列に変換する。例として「孟

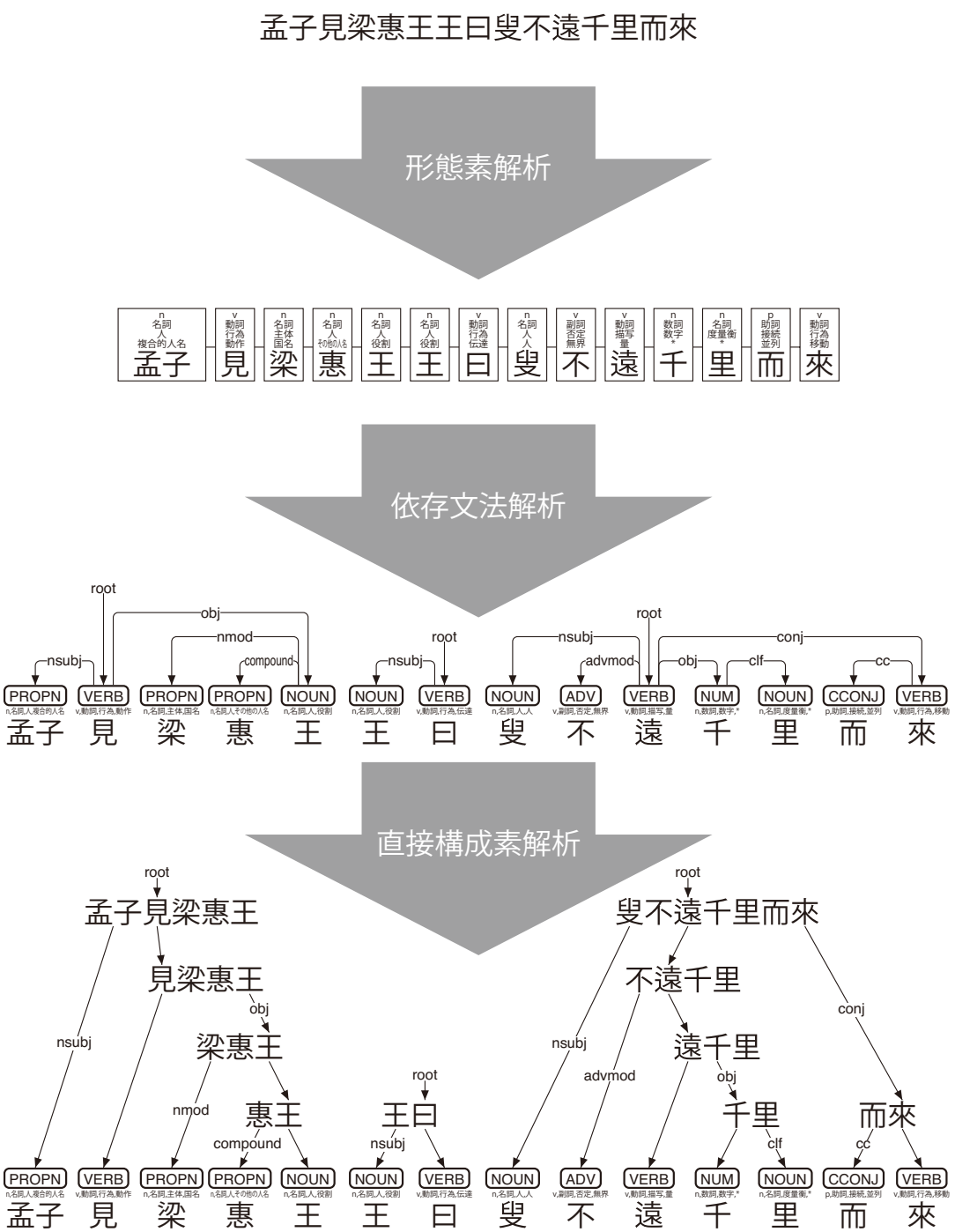

図 1: 漢文の形態素解析・依存文法解析・直接構成素解析

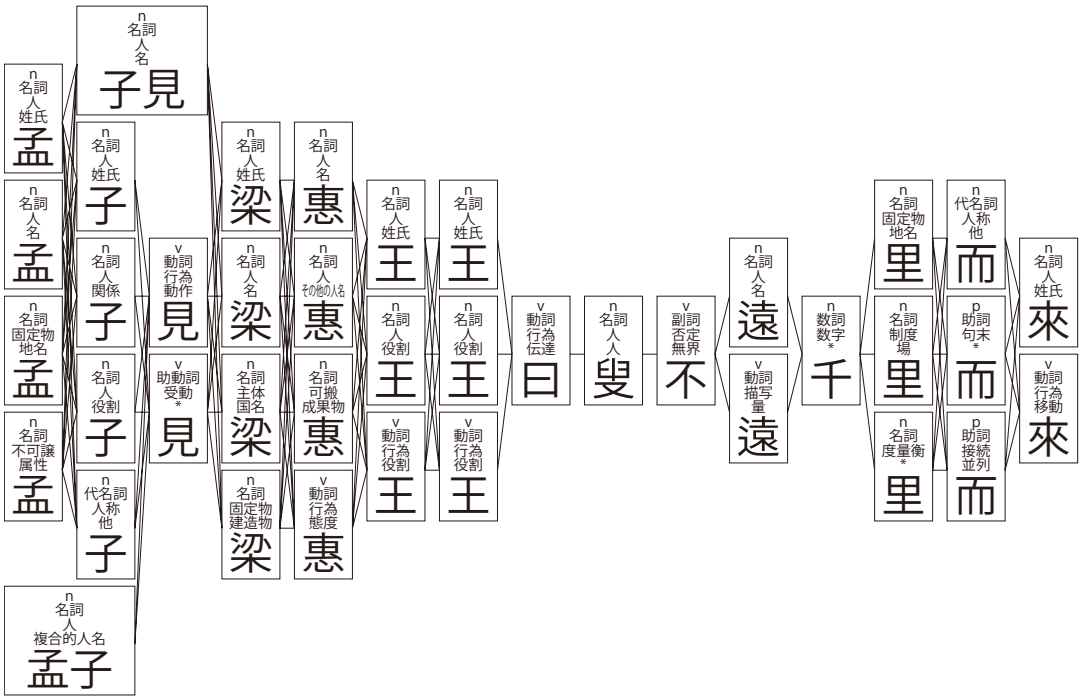

図 2: CRF を用いた漢文の形態素解析

子見梁惠王王曰叟不遠千里而來」という白文に対する CRF（の一部）を、図 2 に示す。「孟子」には「n, 名詞 , 人 , 複合的人名」という「品詞」が与えられており、その一方で、「孟」には 4 種類の「品詞」が、「子」には 4 種類の「品詞」が与えられうることから、これらの組み合わせが全て列挙されている。さらに「子見」に「n, 名詞 , 人 , 名」という「品詞」が与えられていて、「見」は「v, 動詞 , 行為 , 動作」と「v, 助動詞 , 受動 ,*」の可能性がある。これらの組み合わせに対し、MeCab は、各単語の出現確率と、隣り合う単語どうしの共起確率から、全ての組み合わせの中で最も確率が高くなるような単語列を抽出する。そのようにして、図 1 の形態素解析結果が得られるわけである。

❖ 漢文の依存文法解析

古典中国語の依存文法解析 [6] において、われわれは、Universal Dependencies[7]（以下「UD」）という、言語横断的な依存構造記述を用いることにした。UD は、品詞・形態素属性・依存構造情報を、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できるよう、全ての構文構造を「単語」間の依存関係で記述するのが特徴である。

UD には 17 種類の品詞が準備されているが、これらのうち、われわれの古典中国語 UD では PROPN・NOUN・PRON・NUM・VERB・ADP・ADV・AUX・PART・SCONJ・CCONJ・INTJ の 12 種類を使用している [8]。残りの 5 種類（ADJ・DET・PUNCT・SYM・X）は、使用していない。12 種類の UD 品詞へは、漢文形態素解析の結果として得られる 4 階層の品詞を、自動変換している。「単語」間の係り受け関係に対しては、UD 依存構造の「単語」間リンクを用いて表現し、各リンクに表 1 の UD 依存構造タグ 34 種類を付与している [8]。タグのうち 30 種類は、もともと UD で規定されているものであり、4 種類（nsubj:pass・csubj:pass・discourse:sp・flat:vv）は、その派生形である。root はリンク元を持たないが、他のタグによるリンクは、リンク元の「単語」とリンク先の「単語」を 1 つずつ有する。たとえば、漢文の動賓構

	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 → csubj:pass [受動文] ccomp 節目的語 xcomp 節補語		
Non-core arguments	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 → 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	compound 複合（endocentric） flat 並列（exocentric） → flat:vv [動詞類]	list 細目 parataxis 隣接表現		root 親

表 1: 古典中国語に対する UD 依存構造タグ

造は、動詞をリンク元、賓語をリンク先、とするobj というリンクで表現する。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、図 1 の依存文法解析結果にも示すとおり、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。また、リンクはループしない。さらに、われわれの古典中国語 UD では、リンクどうしが交差しない、root をまたぐリンクも存在しない、という制限も設けている。

依存文法解析のための手法は、これまでに数多く提案されているが、われわれの古典中国語 UD のように、複数の root を持ち（dependency forest）、UD 依存構造のリンクどうしが交差せず（planar）、root をまたぐリンクがない（projective）、という条件においては、arc-planar[9] という（非決定性）アルゴリズムが、有効だと考えられる。arc-planar は、単語列の先頭から末尾に向かって「垣根」（stack-buffer boundary）を移動していく、というイメージで処理をおこなう。「垣根」がおこなう遷移は、Shift・Reduce・Left-Arc・Right-Arc の 4 種類である。

- Shift「垣根」を右に 1 単語分、移動する。
- Reduce「垣根」のすぐ左の単語を除去して、解析結果へ移す。
- Left-Arc「垣根」のすぐ右の単語から、すぐ左の単語へリンクを繋ぐ。
- Right-Arc「垣根」のすぐ左の単語から、すぐ右の単語へリンクを繋ぐ。

単語が全て Reduce されて、「垣根」がポツンと取り残された時点で、arc-planar は終了である。arc-planar による「孟子見梁惠王王曰叟不遠千里而來」の依存文法解析の様子を、図 3 に示す。あとは、リンクが入っていない「見」「曰」「遠」に root を刺すことで、図 1 の依存文法解析結果が得られるわけである。

ただし、arc-planar における「垣根」の遷移は、実際には非決定的[10] である。図 3 では解析過程を一本道で示したが、現実には、各局面において複数の可能性が、枝分かれとして存在する。これら複数の可能性については、それぞれの遷移を選択した場合を、確率的に並行して解析することになる。

❖ 漢文の直接構成素解析

依存文法解析によって得られる UD 依存構造に対し、われわれは、階層化 UD 依存構造[11] を導入した。階層化 UD 依存構造は UD 依存構造の変形であり、リンク先として辿っていくことができる単語を全て、リンク元に集約する形で階層化をおこなう（図 4）。ただし、UD 依存構造のリンクどうしが交差しない、root をまたぐリンクがな

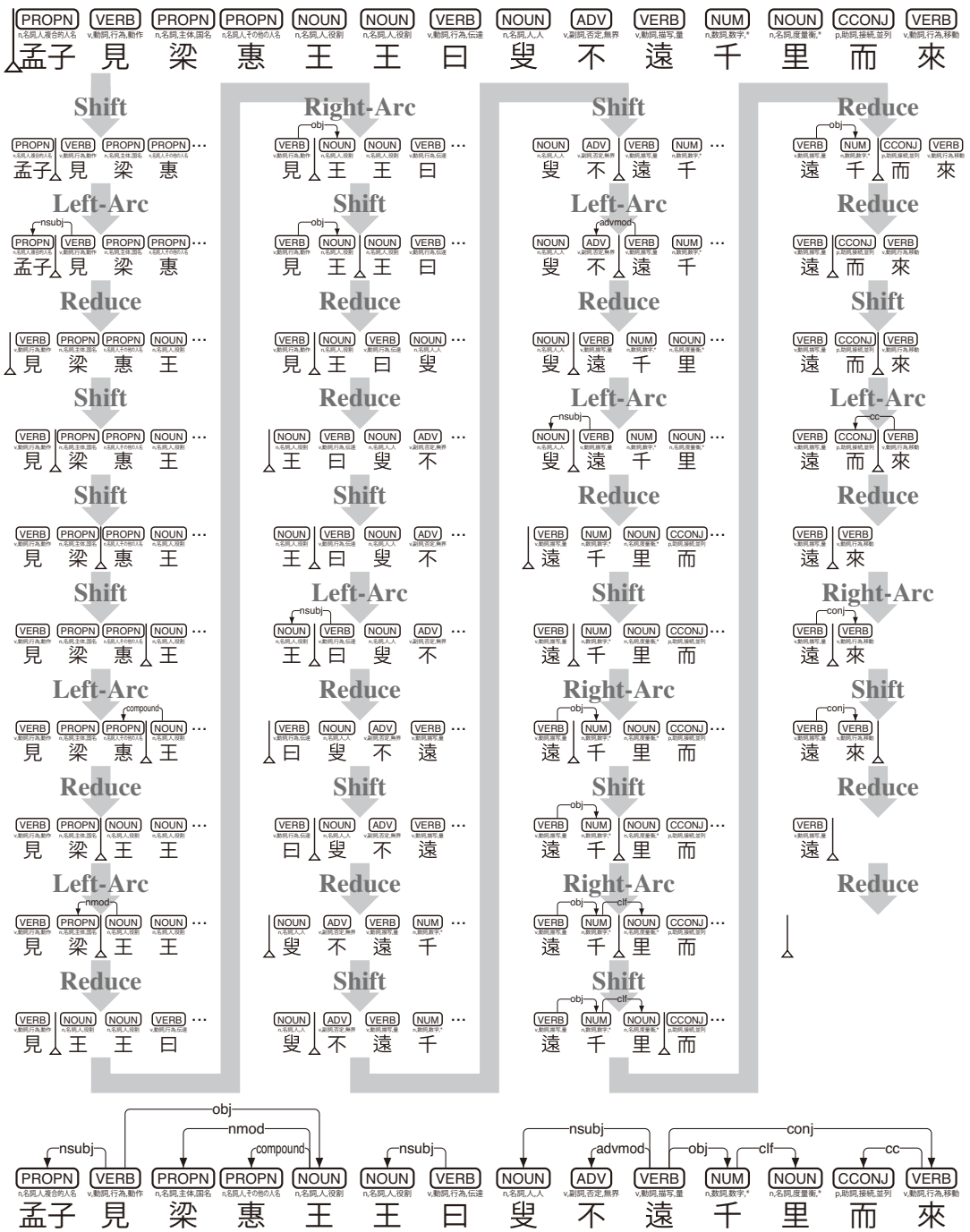

図 3: arc-planar による漢文の依存文法解析

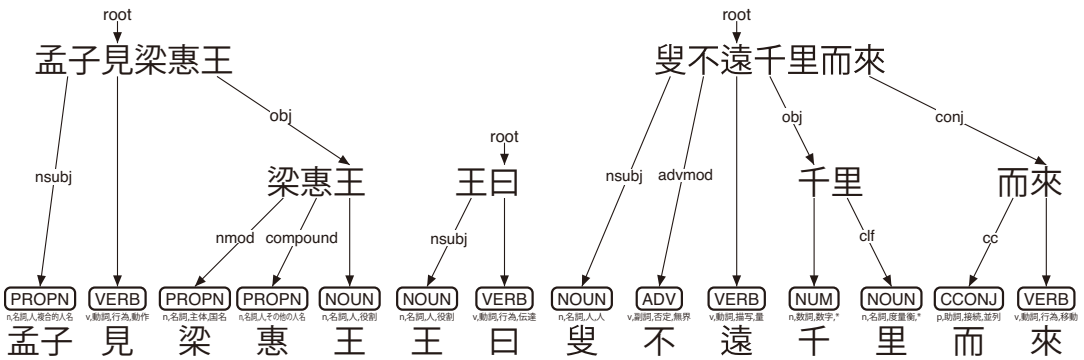

図 4: 階層化 UD 依存構造

い、というのが、階層化 UD 依存構造への変形可能条件である。

階層化 UD 依存構造は、直接構成素解析の前段階とみなすことができる。図 4 の階層化 UD 依存構造であれば、これに「見梁惠王」「惠王」「不遠千里」「遠千里」の 4 つの中間ノードを加えることで、直接構成素解析の結果解析木（図 1）を導出できる。では、どういう形で中間ノードを加えるのが適切なのか。少し議論してみよう。

図 4 の「孟子見梁惠王」というノードからは、nsubj と obj のリンクが両方 [12] 出ている。一方、図 1 解析木の「孟子見梁惠王」からは、nsubj は出ているが、obj は出ていない。これを nsubj ＜ obj という半順序で表すことにする。同様に、図 4 の「梁惠王」というノードからは、nmod と compound のリンクが両方 [12] 出ている。一方、図 1 解析木の「梁惠王」からは、nmod は出ているが、compound は出ていない。これを nmod ＜ compound という半順序で表すことにする。

これらの半順序を階層化 UD 依存構造に適用し、半順序を満たすような中間ノードを加えることで、直接構成素解析が可能となる。たとえば、図 4 の「叟不遠千里而來」からは、nsubj と advmod と obj と conj のリンクが出ている [12] が、ここに nsubj ＜ obj および nsubj ＜ advmod という条件があれば、obj と advmod のリンクを nsubj より後で出すために、「不遠千里」という中間ノードを加えるのが適切、ということになる。さらに、advmod ＜ obj が追加されて、nsubj ＜ advmod ＜ obj という条件ならば、「不遠千里」と「遠千里」という 2 つの中間ノードを加えるのが適切、ということになる。

ただし、これらの半順序は、多数の例文の間では時に矛盾を孕む。したがって、これらの半順序は確率として扱うべきであり、直接構成素解析における中間ノードの追加は、それぞれの可能性を、確率的に並行して解析することになるだろう。

☘ おわりに

古典中国語の形態素解析・依存文法解析・直接構成素解析に関して、ざっと、その概要を述べた。これらの解析を、できる限り正確におこなうためには、人手（もしくは半自動）で作成した漢文コーパスが必要となる。形態素解析における共起確率や、依存文法解析における遷移確率や、直接構成素解析における半順序確率は、大量の漢文コーパスから導出するしか、現実的な方法がないからだ。

われわれが構築中の漢文コーパスは

https://corpus.kanji.zinbun.kyoto-u.ac.jp/gitlab/Kanbun

において、逐次公開中である。ぜひ一度、触れてみてほしい。

注

[1] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing（July 2004）, pp.230-237.

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

[3] Rulon S. Wells: Immediate Constituents, Language, Vol.23, No.2（April-June 1947）, pp.81-117.

[4] 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.

[5] 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.

[6] 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（2018 年 1 月）, No.20, pp.1-8.

[7] 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.

[8] 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）.

[9] Carlos Gómez-Rodríguez, Joakim Nivre: A Transition-Based Parser for 2-Planar Dependency Structures, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics（July 2010）, pp.1492-1501.

[10]「垣根」が単語列の先頭にある局面では Shift を、末尾にある局面では Reduce をおこなうしかなく、これらの場合だけは決定的である。また、Left-Arc の直後を Reduce に、Right-Arc の直後を Shift に、それぞれ決め打ちすることは（局面に応じて）可能である。

[11] 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.

[12] head を表すリンク [11]（図 4 ではタグなしリンクで表現）も出ているが、このリンクは中間ノードの付加に用いることとし、半順序の比較には用いない。