東方學報 京都第 94 冊（2019）：330-322 頁

漢文の形態素解析・依存文法解析・直接構成鎖解析

安 岡 孝 一

筆者が班長を務める京都大學人文科學研究所共同研究班「東アジア古典文獻コーパスの實證研究」では，現在，古典中國語（漢文）における文法解析の自動化に全力で取り組んでいる。漢文の白文に對し，形態素解析・依存文法解析・直接構成鎖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，單語切りをおこなうと同時に，各單語の品詞を得る。依存文法解析[2]によって，單語と單語の間の係り受け關係を解析すると同時に，文の切れ目を得る。直接構成鎖解析[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」という白文に對し，形態素解析・依存文法解析・直接構成鎖解析を順におこなう際の流れ（イメージ圖）を，圖 1 に示す。

なお，本研究は，科學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。

漢文の形態素解析

漢文の形態素解析において，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ている。MeCab は，もともとは日本語の形態素解析用だった[1]が，言語，辭書，コーパスに依存しない汎用的な設計がなされており，辭書とコーパスを準備すればいかなる言語にも對應できる。

MeCab の辭書には「品詞」（複數の階層が可能）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」の一種である訓讀に着目し，返り點を「品詞」に反映させることを考えた。すなわち，訓讀における返り點が，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 漢文辭書の「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」「數詞」「動詞」「前置詞」「副詞」「助動詞」

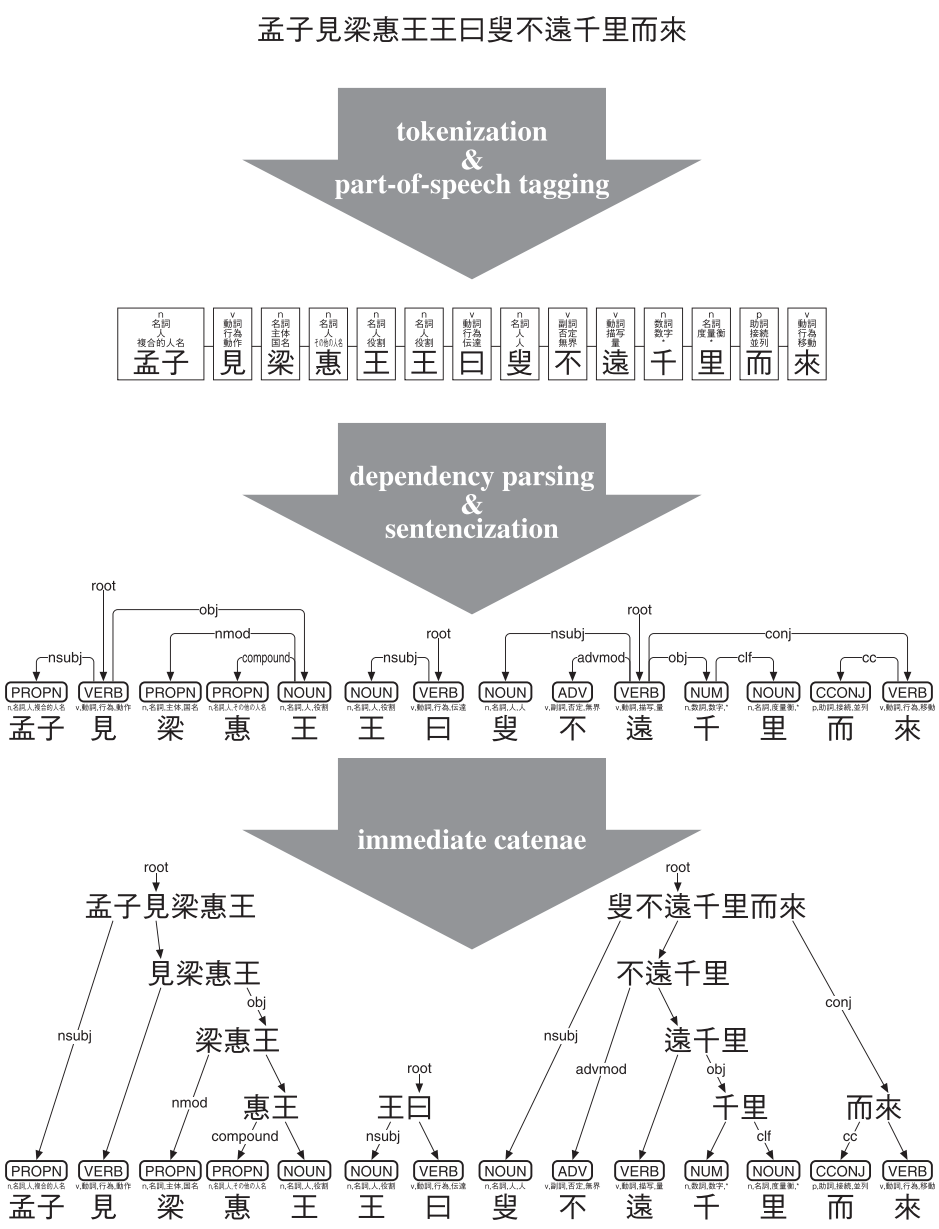

圖 1　漢文の形態素解析・依存文法解析・直接構成鎖解析

「助詞」「感嘆詞」「接尾辭」の 10 種類とした。従來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[4]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單語を，意味の面からも捉えやすいよう工夫した[5]。

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

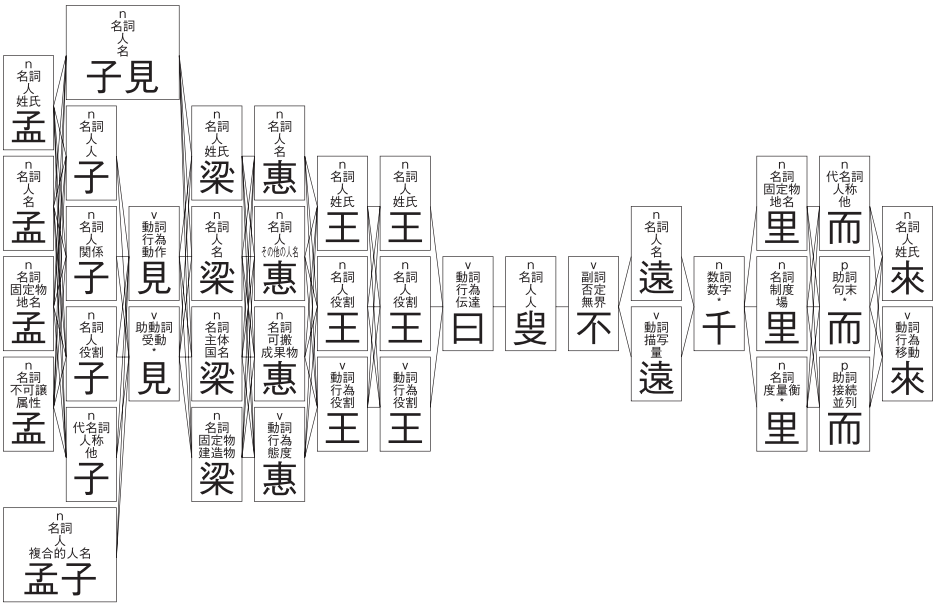

圖 2　CRF を用いた漢文の形態素解析

Random Fields）である。われわれの漢文形態素解析に卽して言えば，解析したい白文を MeCab 漢文辭書に基いて，可能性のある全ての單語（4 階層の品詞を含む）の組み合わせの列に變換する。例として「孟子見梁惠王王曰叟不遠千里而來」という白文に對する CRF（の一部）を，圖 2 に示す。「孟子」には「n，名詞，人，複合的人名」という「品詞」が與えられており，その一方で，「孟」には 4 種類の「品詞」が，「子」には 4 種類の「品詞」が與えられうることから，これらの組み合わせが全て列擧されている。さらに「子見」に「n，名詞，人，名」という「品詞」が與えられていて，「見」は「v，動詞，行爲，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對し，MeCab は，各單語の出現確率と，隣り合う單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單語列を抽出する。

抽出した單語列に對し，Universal Dependencies 向け品詞（表 1）とグロス（近似的な逐語英譯）の付與を竝行しておこない，形態素解析を完成する。グロスは基本的に[6]に依っているが，一人稱・二人稱・三人稱は［1PRON］・［2PRON］・［3PRON］で，受動の助動詞は［PASS］で，完了の助詞は［PFV］で，疑問の助詞は［Q］で表した。例として「孟子見梁惠王王曰叟不遠千里而來」に對する形態素解析の結果を，圖 3 に示す。

表 1　MeCab 向け漢文品詞體系と Universal Dependencies 向け品詞

大品詞，品詞	UD 向け品詞	特　例	
n, 名　詞	NOUN	n, 名詞，人，姓氏 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，主體，國名 n, 名詞，固定物，地名	PROPN
n, 代名詞	PRON		
n, 數　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，接續，屬格 p, 助詞，接續，竝列	SCONJ CCONJ
p, 感嘆詞	INTJ		
p, 接尾辭	PART		
s, 記　號	PUNCT		
s, 文　字	SYM		

圖 3 「孟子見梁惠王王曰叟不遠千里而來」の形態素解析結果

漢文の依存文法解析

漢文の依存文法解析[7]において，われわれは，Universal Dependencies[8]（以下「UD」）という，言語横斷的な依存構造記述を用いている。UD は，品詞・形態素屬性・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關係を記述できるよう，全ての構文構造を「單語」間の依存關係で記述するのが特徵である。

「單語」間の係り受け關係に對しては，UD 依存構造の「單語」間リンクを用いて表現し，各リンクに表 2 の UD 依存構造タグ 39 種類を付與している。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，7 種類（nsubj : pass・csubj : pass・obl : tmod・obl : lmod・discourse : sp・compound : redup・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「單語」とリンク先の「單語」を 1 つずつ有する。たとえば，漢文の動賓構造は，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。リンクの本數は「單語」の個數に等しく，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「單語」から出るリンクは複數

表2　古典中國語（漢文）Universal Dependencies の依存構造タグ

	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 dependents	obl 斜格補語 ↪obl : tmod［時］ ↪obl : lmod［場所］ vocative 呼稱語 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 繫辭（copula） mark 標識（marker）
Nominal dependents	nmod 體言による連體修飾語 nummod 數量による修飾語	acl 連體修飾節	amod 用言による連體修飾語	det 決定詞 clf 類別詞 case 格表示
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj 接續 cc 接續詞	fixed 固着 compound 複合（endocentric） ↪compound : redup［重疊］ flat 竝列（exocentric） ↪flat : vv［動詞類］	list 細目 parataxis 隣接表現	orphan 親なし	root 親

ありうるが，各「單語」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，われわれの古典中國語（漢文）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，これまでに數多く提案されているが，われわれの古典中國語（漢文）UD のように，複數の 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），という條件においては，arc-planar[9]という（非決定性）アルゴリズムが，有效だと考えられる。arc-planar は，單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，というイメージで處理をおこなう。「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單語分，移動する。
・**Reduce**　「垣根」のすぐ左の單語を除去して，解析結果へ移す。
・**Left-Arc**　「垣根」のすぐ右の單語から，すぐ左の單語へリンクを繫ぐ。
・**Right-Arc**　「垣根」のすぐ左の單語から，すぐ右の單語へリンクを繫ぐ。

單語が全て **Reduce** されて，「垣根」がポツンと取り殘された時點で，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」の依存文法解析の樣子を，圖4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

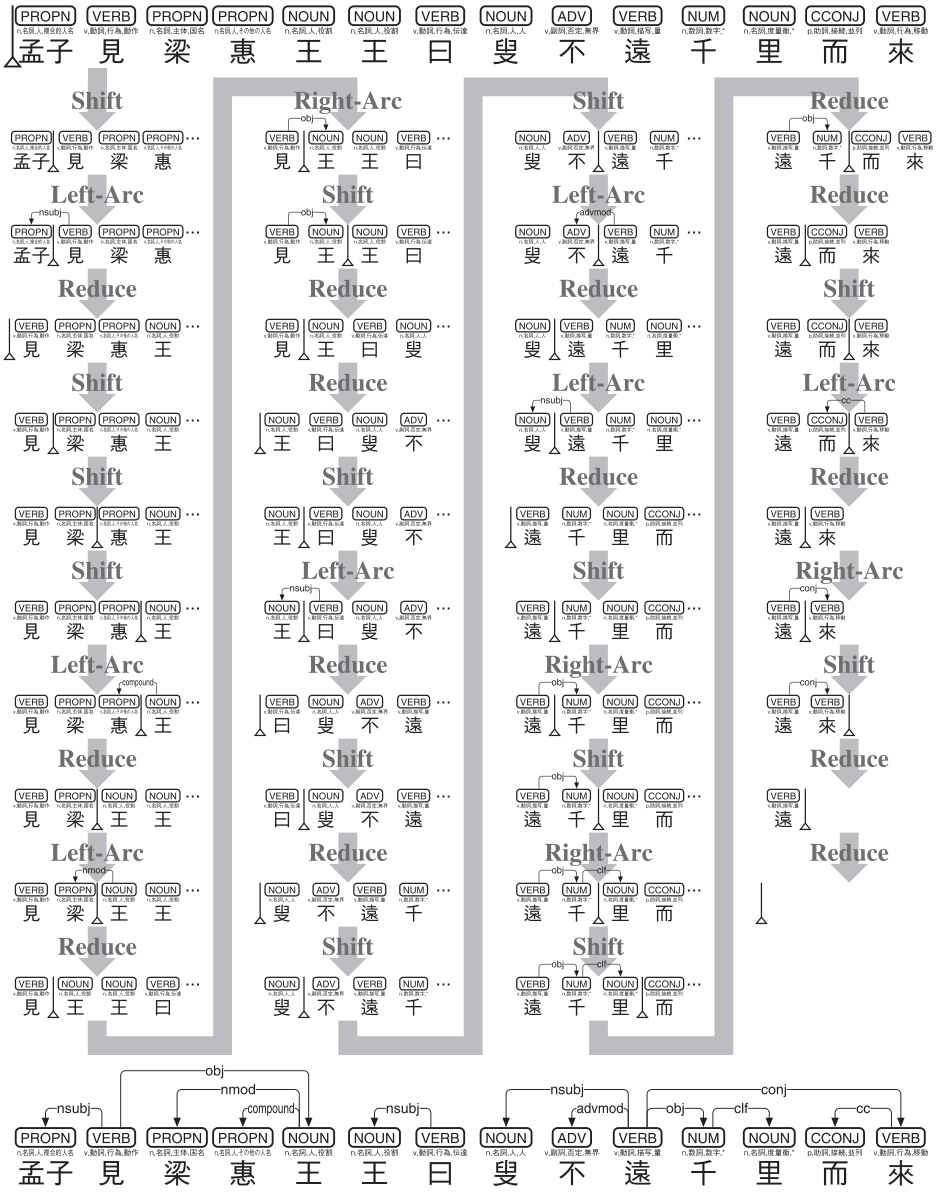

圖 4　arc-planar による漢文の依存文法解析

で，圖 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，實際には非決定的 * である。圖 4 では

---

* 「垣根」が單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，↗

解析過程を一本道で示したが，現實には，各局面において複數の可能性が，枝分かれとして存在する。これら複數の可能性については，それぞれの遷移を選擇した場合を，確率的に竝行して解析することになる。われわれが解析に用いている UDPipe[10]においても，ほぼ，そのような形で實裝がおこなわれている。

漢文の直接構成鎖解析

直接構成鎖解析（immediate catena analysis）[3] は，構成素（constituent）[11] による古典的な文法解析を，構成鎖（catena）[12] を用いて依存文法へと擴張する手法である。われわれは，漢文の依存文法記述に UD を用いていることから，直接構成鎖解析においても，UD を擴張する形で解析手法の開發をおこなった。

構成鎖は，依存文法の文法木における連結な部分グラフである。たとえば，圖 5 左上の UD「孟子見梁惠王」は，以下に示す 17 種類の構成鎖を含んでいる。

「孟子見梁惠王」「孟子見梁王」「孟子見惠王」「見梁惠王」「孟子見王」
「見梁王」「見惠王」「梁惠王」「孟子見」「見王」「梁王」「惠王」「孟子」
「見」「梁」「惠」「王」

直接構成鎖解析における基本コンセプトは，リンクの除去である。最初に root リンクの除去をおこない，そこから文法木を幅優先でリンクの除去をおこない，最終的には全てのリンクを除去する。圖 5 左を見てみよう。最初に root リンクを除去し，極大構成鎖「孟子見梁惠王」を得る。次に nsubj リンクを除去すると，「孟子見梁惠王」は，「孟子」と「見梁惠王」の 2 つの構成鎖に分割される。次に obj リンクを除去すると，「見梁惠王」は，「見」と「梁惠王」の 2 つの構成鎖に分割される。次に nmod リンクを除去すると，「梁惠王」は，「梁」「惠王」の 2 つの構成鎖に分割される。最後に compound リンクを除去すると，「惠王」は，「惠」と「王」の 2 つの構成鎖に分割される。この分割の樣子を，再度，構文木として組み上げたのが圖 5 右であり，これが「孟子見梁惠王」に對する直接構成鎖解析の構文木である。

同樣に，「王曰」「叟不遠千里而來」に對する直接構成鎖解析の構文木を，圖 1 下に示す。ただし，これらの直接構成鎖解析を適切におこなうためには，適切な順序でリンク

↘ **Right-Arc** の直後を **Shift** に，それぞれ決め打ちすることは（局面に應じて）可能である。

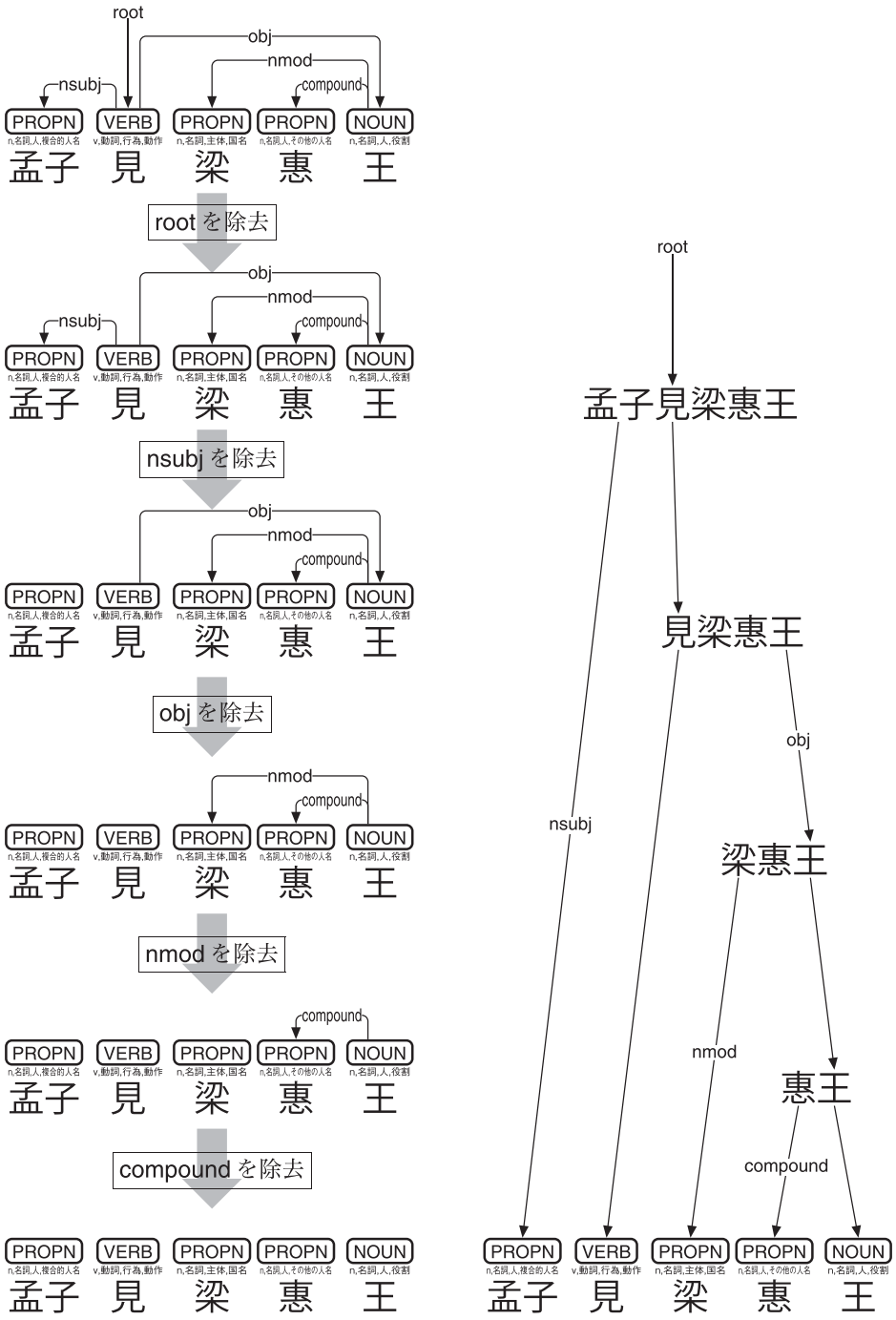

圖 5 「孟子見梁惠王」の直接構成鎖解析

の除去をおこなう必要がある上に，複數のリンクの同時除去も考慮に入れなければならない。しかし，そのような除去順序を導出するようなアルゴリズムを，現時點のわれわれは開發しきれていない。われわれの今後の研究の進展に期待されたい。

參 考 文 獻

［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］安岡孝一：Universal Dependencies の擴張にもとづく古典中國語（漢文）の直接構成鎖解析の試み，情報處理學會研究報告，Vol. 2019–CH–120（2019 年 5 月），No. 1, pp. 1–8.

［4］山崎直樹，守岡知彥，安岡孝一：古典中國語形態素解析のための品詞體系再構築，人文科學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.

［5］安岡孝一，ウィッテルン クリスティアン，守岡知彥，池田巧，山崎直樹，二階堂善弘，鈴木愼吾，師茂樹：古典中國語（漢文）の形態素解析とその應用，情報處理學會論文誌，Vol. 59, No. 2（2018 年 2 月），pp. 323–331.

［6］Edwin G. Pulleyblank : Lexicon of Reconstructed Pronunciation in Early Middle Chinese, Late Middle Chinese, and Early Mandarin, Vancouver : UBC Press（1991）.

［7］安岡孝一，ウィッテルン クリスティアン，守岡知彥，池田巧，山崎直樹，二階堂善弘，鈴木愼吾，師茂樹：古典中國語 Universal Dependencies への挑戰，情報處理學會研究報告，Vol. 2018–CH–116（2018 年 1 月），No. 20, pp. 1–8.

［8］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.

［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］Milan Straka and Jana Straková : Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe, Proceedings of the CoNLL 2017 Shared Task（August 2017），pp. 88–99.

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

［12］Timothy Osborne, Michael Putnam, Thomas Groß : Catenae : Introducing a Novel Unit of Syntactic Analysis, Syntax, Vol. 15, No. 4（December 2012），pp. 354–396.