

漢文自動訓読ツールUD-Kundokuの開発

安岡孝一*

1 はじめに

日本における漢文(古典中国語)の受容は、一つには訓読という形でおこなわれてきた。訓読は、言語処理という観点から見た場合、VO型の孤立語である古典中国語を、OV型の膠着語である日本語の書き下し文に変換する、という過程の一種だとみなせる。訓読を、返り点と送り仮名に分けたならば、VO型からOV型への変換を返り点が担い、孤立語から膠着語への変換を送り仮名が担っている、と考えることもできるだろう。

一方、コンピュータによる漢文の自動解析は、形態素解析→依存文法解析→直接構成鎖解析という手順によって、白文の統語構造を得ることができる(図1)、というのが、筆者の目論見^[1]である。入力された白文に対し、形態素解析によって、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。直接構成鎖解析によって、各文の統語構造を解析木の形で得る。

ただし、コンピュータによる返り点の自動付与に関しては、直接構成鎖解析をおこなう必要は無く、依存文法解析までで(ほぼ)十分である。卑近な言い方をすれば、訓読の返り点は、漢文の統語構造ではなく、係り受け関係によって(ほぼ)記述可能^[2]である。一方、送り仮名の自動付与に関しては、膠着語としての日本語を特徴づける格助詞・接続助詞・活用語尾を、各単語ないしは構成鎖の末尾に付与する、という問題に帰着できる。

これらのアイデアに基づき、漢文自動訓読ツールUD-Kundokuを開発した。入力された古典中国語テキストに対し、最初に依存文法解析をおこない、次に返り点の自動付与をおこなって語順を入れ替え、最後に送り仮名を自動付与する、というツールである。本稿では、このUD-Kundokuの設計と実装について述べる。

2 Universal Dependencies による漢文の依存文法解析

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、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^[3](以下「UD」)の古典中国語への適用を研究してきた。依存文法解析それ自体は、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

*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

^[1]安岡孝一: 漢文の形態素解析・依存文法解析・直接構成鎖解析, 東方學報, 第94冊(2019年12月), pp.330-322.

^[2]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.

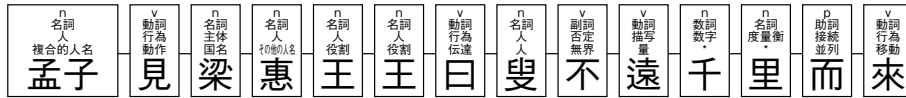
^[3]Joakim Nivre: Towards a Universal Grammar for Natural Language Processing, CICLing 2015: 16th International Conference on Intelligent Text Processing and Computational Linguistics (April 2015), pp.3-16.

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

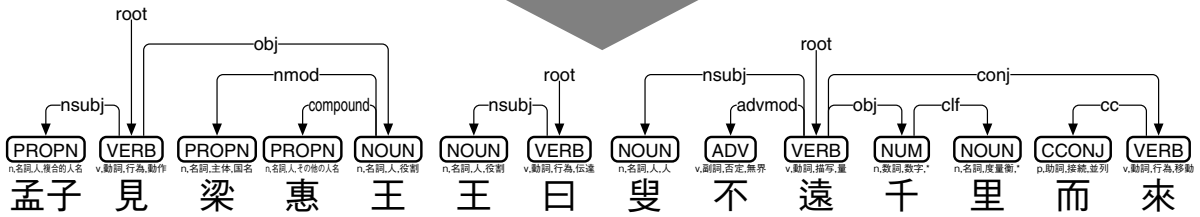
^[5]Igor A. Mel'čuk: Dependency Syntax: Theory and Practice, New York: State University of New York Press (1988).

孟子見梁惠王王曰叟不遠千里而來

形態素解析



依存文法解析



直接構成鎖解析

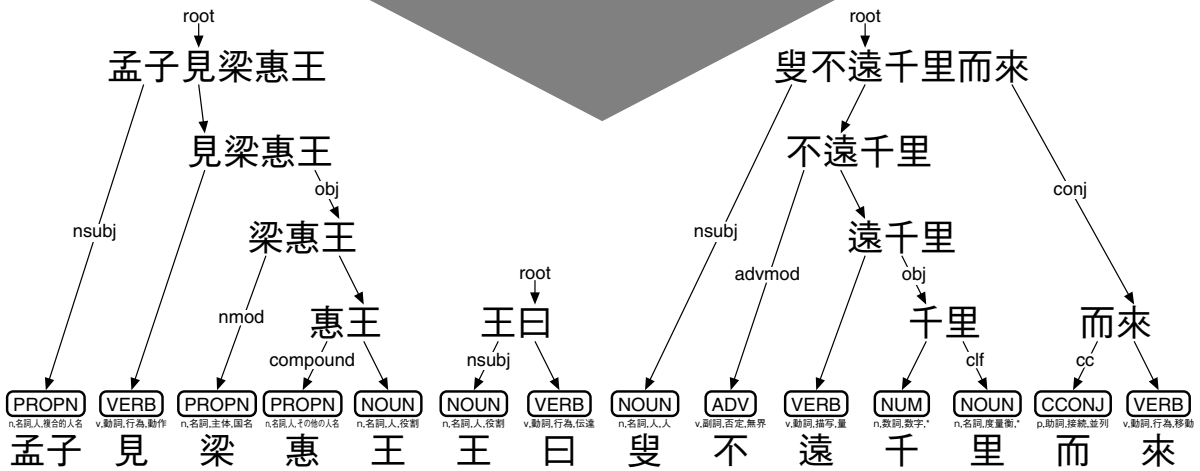


図 1: 白文「孟子見梁惠王王曰叟不遠千里而來」に対する統語構造解析の流れ

表 1: 古典中国語に対する UD 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 ccomp 節目的語 xcomp 節補語		
Non-core arguments	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 繫辞 (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 [重疊] flat 並列 (exocentric) ↔flat:vv [動詞類]	list 細目 parataxis 隣接表現	orphan 親なし	root 親

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を可能としているのである。

UDにおける係り受け関係の記述は、文中の単語をノードとする有向グラフにおいて、単語間の依存関係をリンクで表現する。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必ず1本とする。また、リンクはループしない。リンクには、それぞれUD依存構造タグを付与する。古典中国語UDにおいては、表1に示す38種類^[6]のタグを使用している。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、6種類(nsubj: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。

白文に対する依存文法解析は、その前段階として、単語切りという処理を必要とする。白文では、単語と単語の間に区切りがないことから、単語というものを処理単位とする依存文法解析に際し、まず、白文を単語に区切る処理が必要となるのである。この処理をわれわれは、漢文の形態素解析^[7]という形で実現し、白文の単語切りをおこなうと同時に、各単語に対して4階層の品詞を得ている(図1)。また、この際に、UD向け品詞(PROPN・NOUN・PRON・NUM・VERB・ADP・ADV・AUX・PART・SCONJ・CCONJ・INTJ・SYM)も同時に得ている。

^[6]Koichi Yasuoka: Universal Dependencies Treebank of the Four Books in Classical Chinese, DADH2019: 10th International Conference of Digital Archives and Digital Humanities (December 2019), pp.20-28.

^[7]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.

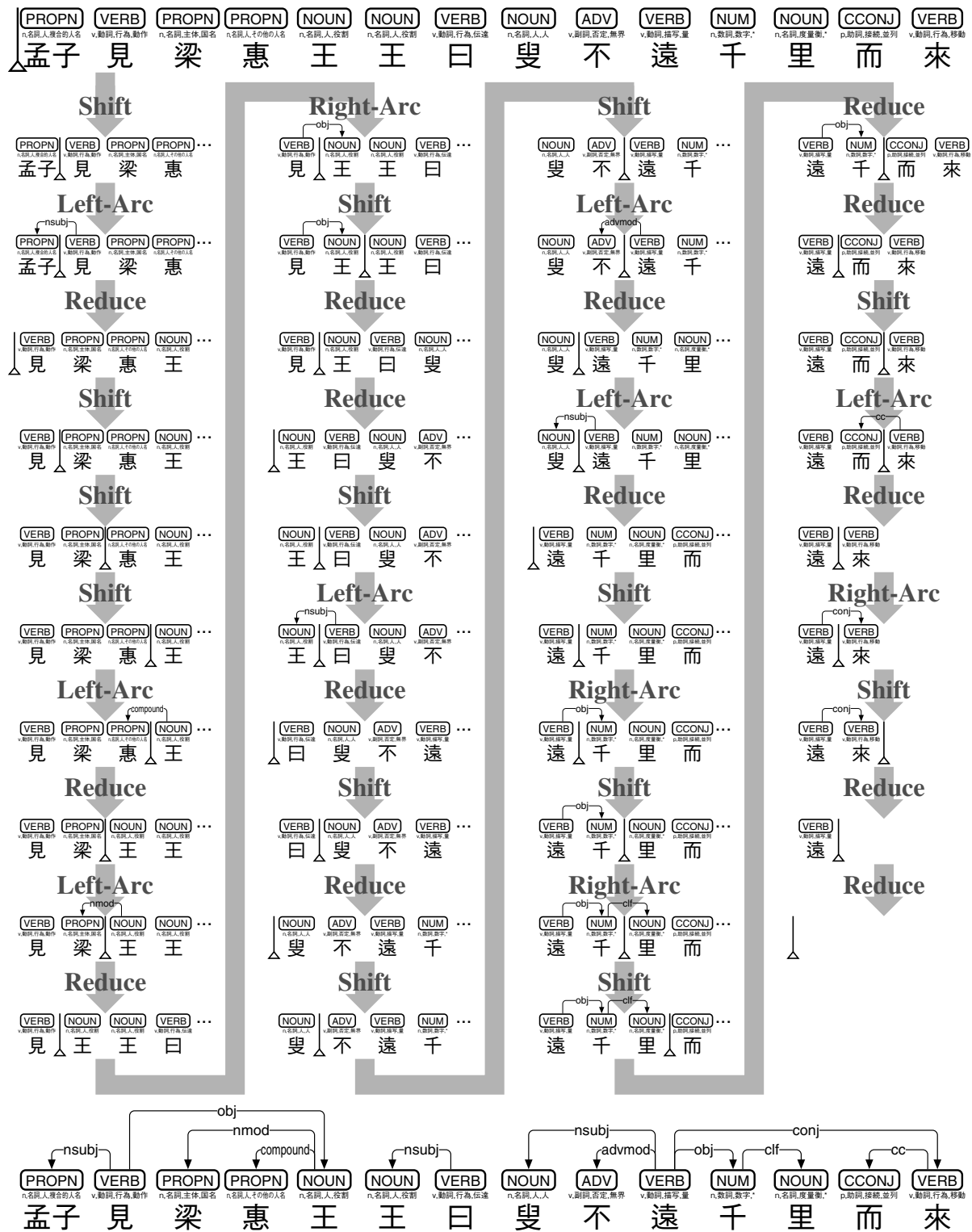


図 2: arc-planar による漢文の依存文法解析

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^[8] 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 2 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、実際には非決定的である。図 2 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。

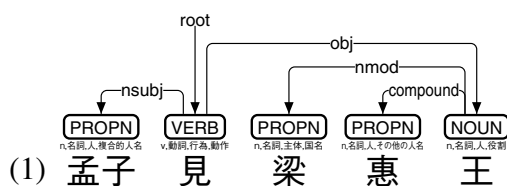
3 漢文の依存文法解析にもとづく返り点の自動付与

漢文の依存文法解析の結果から、いくつかのルールにもとづいて、訓読の返り点を導出したい。どのようなルールが必要となるのか、以下では「孟子定本」に対する依存文法解析結果をもとに、右向き UD リンク、左向き UD リンク、特別扱いすべき個別の文字、返り点の付け替え、の 4 つの視点から、返り点のルールを考えてみよう。なお、以下の文例では、返り元と返り先を視覚化すべく、レ点の代わりに一二点を用いている。各文例に対する富山房『漢文体系』での返り点は、図 3 を参照されたい。

3.1 右向き UD リンクに対する返り点のルール

右向き UD リンクに対する返り点のルールは、以下のようなものが考えられる。

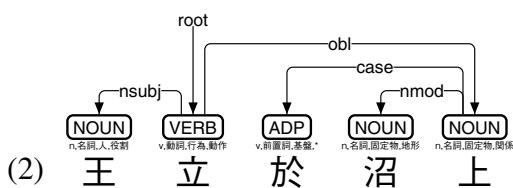
ルール① 右向き obj リンクは、リンク先からリンク元へ返り点を打つ (文例 1)。



孟子見梁惠王

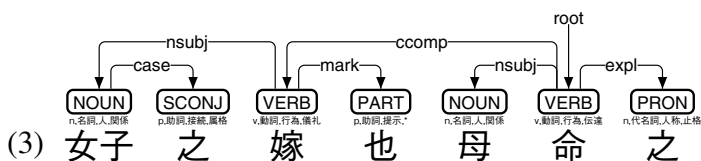
[8] Carlos Gómez-Rodríguez, Joakim Nivre: A Transition-Based Parser for 2-Planar Dependency Structures, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (July 2010), pp.1492-1501.

ルール② 右向きの obli リンクは、リンク先からリンク元へ返り点を打つ (文例 2)。



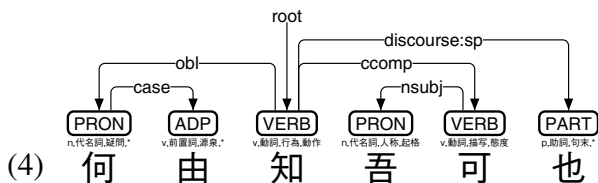
王立於沼上

ルール③ 右向きの expl リンクは、リンク先からリンク元へ返り点を打つ (文例 3)。



女子之嫁也母命之

ルール④ 右向きの ccomp および xcomp リンクは、リンク先からリンク元へ返り点を打つ (文例 4)。

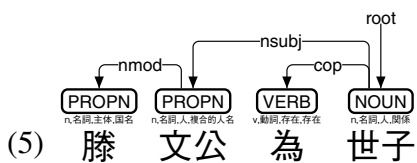


何由知吾可也

3.2 左向きの UD リンクに対する返り点のルール

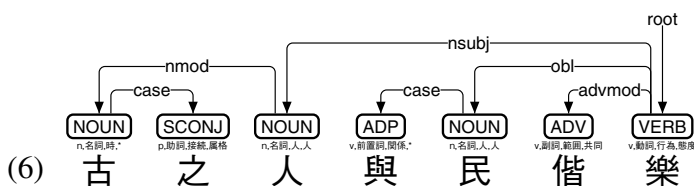
左向きの UD リンクに対する返り点のルールは、以下のようなものが考えられる。

ルール⑤ 左向きの cop リンクは、リンク元からリンク先へ返り点を打つ (文例 5)。



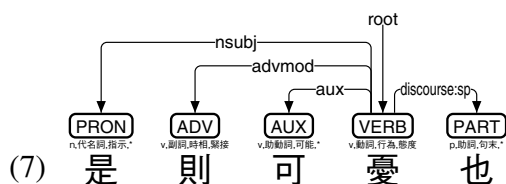
滕文公為世子

ルール⑥ 左向きの case および mark リンクは、リンク元からリンク先へ返り点を打つ (文例 6)。ただし、リンク先の形態素解析結果が「v, 前置詞, 基盤」の場合は、返り点を打たない (文例 2)。



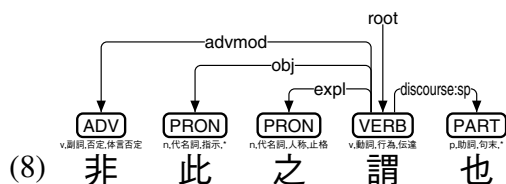
古之人與民偕樂

ルール⑦ 左向きの **aux** リンクは、リンク元からリンク先へ返り点を打つ(文例7)。



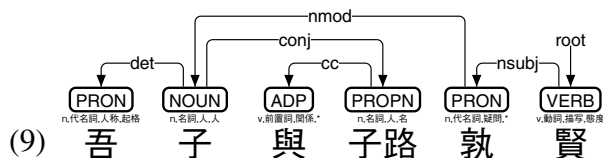
是則可憂也

ルール⑧ 左向きの **advmod** リンクは、リンク先の形態素解析結果が「v, 副詞, 否定」「v, 副詞, 判断, 逆接」「v, 副詞, 時相, 将来」の場合、あるいはリンク先が「難」「易」の場合に限って、リンク元からリンク先へ返り点を打つ(文例8)。



非此之謂也

ルール⑨ 左向きの **cc** リンクは、リンク先の形態素解析結果が「v, 前置詞, 関係」の場合に限って、リンク元からリンク先へ返り点を打つ(文例9)。

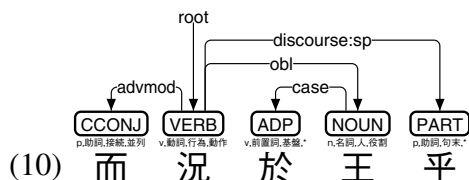


吾子與子路孰賢

3.3 個別の文字に対するルール

個別の文字に対する返り点のルールは、以下のようなものが考えられる。以下のルールの中には、富山房『漢文大系』の「孟子定本」における安井衡の「返りグセ」のようなものも含まれている気がするが、ルール化できそうなものは、できるだけルール化することにする。

ルール⑩ ルール①～④で打った返り点において、返り先が「況」の場合、返り点を削除する。ただし、返り点の返り元から、左向きの **case** あるいは **mark** リンクが出ている場合は、返り点を削除する代わりに、返り先をそのリンク先に移動する。

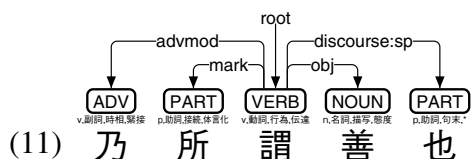


而況於王乎

文例10では、ルール②にもとづいて「王」から「況」へ返り点を打つべきところ、返り先を「於」へ移動している。このルール⑩は、動詞である「況」を、「況んや」と訓読するためのものである。

ルール⑪ ルール①～④で打った返り点において、返り先が「謂」であり、かつ、その「謂」から「所」へ左向きの mark リンクが出ている場合、「謂」を返り先とする返り点を削除する。

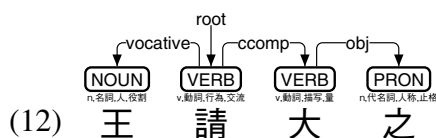
乃所謂善也



文例 11 では、「善」から「謂」への返り点(ルール①)を削除している。このルール⑪は、動詞である「謂」に対し、「所謂」を「いはゆる」と訓読するためのものである。

ルール⑫ ルール①あるいは④で打った返り点において、返り先が「請」であり、かつ、その「請」から vocative リンクが出ている場合、返り点を削除する。

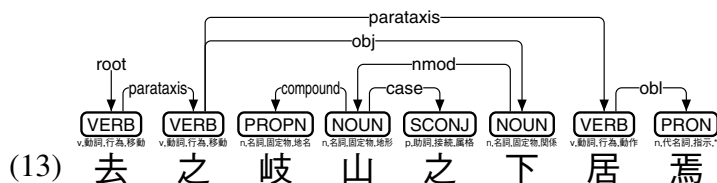
王請大之



文例 12 では、「大」から「請」への返り点(ルール④)を削除している。このルール⑫は、呼びかけを意味する「請」において、訓読を命令調にするためのものである。

ルール⑬ ルール②で打った返り点において、返り元が「焉」であり、かつ、返り先の形態素解析結果が「v, 動詞, 描写」以外の場合、返り点を削除する。

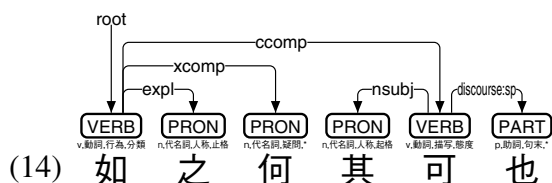
去之岐山之下居焉



文例 13 では、「焉」から「居」への返り点を削除している。このルール⑬は、「於之」を意味する代名詞「焉」に対し、比較を意味する場合を除いて、訓読しないためのものである。

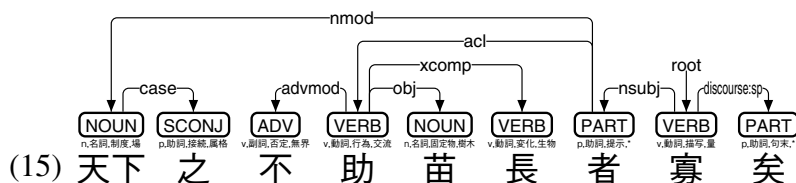
ルール⑭ ルール④で打った返り点において、返り先が「如」であり、かつ、その「如」から obj リンクもしくは expl リンクが出ている場合、返り点を削除する。

如之何其可也



文例 14 では、「何」から「如」への返り点と、「可」から「如」への返り点を、いずれも削除している。

ルール⑮ ルール④で打った返り点において、返り先が「助」であり、かつ、xcompリンクによる場合、返り点を削除する。加えて、その「助」が他のルールによる返り点の返り元である場合、返り元を、削除した返り点の返り元へ移動する。

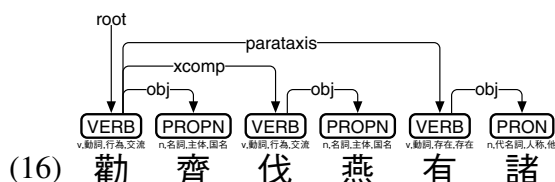


(15) 天下之不助苗長者寡矣

文例 15 では、ルール①にもとづいて「苗」から「助」へ、ルール④にもとづいて「長」から「助」へ、ルール⑧にもとづいて「助」から「不」へ、それぞれ返り点を打つべきところ、「長」から「助」への返り点は削除している。加えて、「助」から「不」への返り点は、返り元を「長」へ移動している。

天下之不助苗長者寡矣

ルール⑯ ルール④で打った返り点において、返り先が「勸」であり、かつ、xcompリンクによる場合、返り点を削除する。加えて、その「勸」が他のルールによる返り点の返り元である場合、返り元を、削除した返り点の返り元へ移動する。

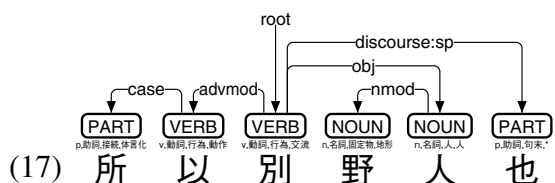


(16) 勸齊伐燕有諸

文例 16 では、「伐」から「勸」への返り点を削除している。

勸齊伐燕有諸

ルール⑰ ルール⑥で打った「以」から「所」への返り点において、その「以」に左向きの advmod リンクが入っている場合、返り元を、advmod リンクのリンク元へ移動する。

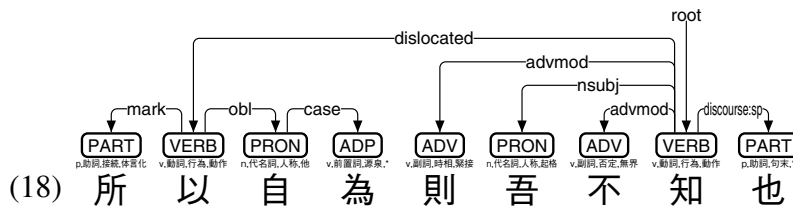


(17) 所以別野人也

文例 17 では、ルール①にもとづいて「人」から「別」へ、ルール⑥にもとづいて「以」から「所」へ、それぞれ返り点を打つべきところ、後者の返り元を「別」へ移動している。このルール⑰は、「所以」を「ゆゑん」と訓読するためのものである。

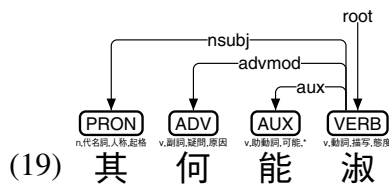
所以別野人也

ルール⑬ ルール⑥で打った「以」から「所」への返り点において、その「以」に左向きの **advmod** リンクが入っていない場合、返り点を削除する。加えて、その「以」が他のルールによる返り点の返り先・返り元である場合、それらの返り先・返り元を「所」へ移動する。



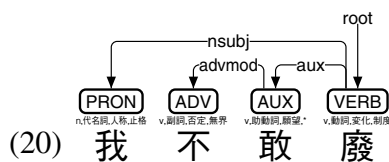
文例 18 では、ルール②にもとづいて「自」から「以」へ、ルール⑥にもとづいて「以」から「所」へ、ルール⑧にもとづいて「知」から「不」へ、それぞれ返り点を打つべきところ、「以」から「所」への返り点は削除している。加えて、「自」から「以」への返り点は、返り先を「所」へ移動した上で、さらにルール⑳にもとづいて返り元を「為」へ移動している。このルール⑱も、「所以」を「ゆゑん」と訓読するためのものである。

ルール⑱ ルール⑦で打った返り点において、返り先が「能」であり、かつ、その「能」が他のルールによる返り点の返り元でない場合、「能」を返り先とする返り点を削除する。



文例 19 では、「淑」から「能」への返り点を削除している。このルール⑱は、助動詞である「能」を、「よく～す」と訓読するためのものである。ただし、「不能」を「あたはず」と訓読する場合は、返り点を削除しない。

ルール⑳ ルール⑦で打った返り点において、返り先が「敢」の場合、返り点を削除する。加えて、その「敢」が他のルールによる返り点の返り元である場合、返り元を、削除した返り点の返り元へ移動する。



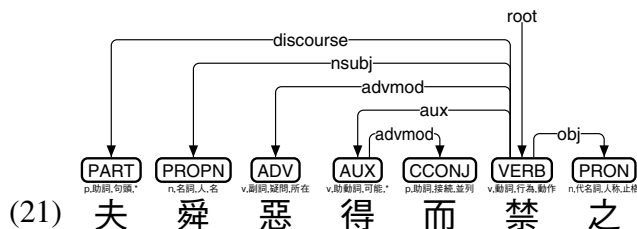
文例 20 では、ルール⑦にもとづいて「廢」から「敢」へ、ルール⑧にもとづいて「敢」から「不」へ、それぞれ返り点を打つべきところ、削除と移動で「廢」から「不」への返り点としている。このルール⑳は、助動詞である「敢」を、「あへて～す」と訓読するためのものである。

所
以
自
為
則
吾
不
知
也

其
何
能
淑

我
不
敢
廢

ルール⑳ ルール㉑で打った返り点において、返り先が「得」であり、かつ、その「得」から「而」へ右向きの **advmod** リンクが出ている場合、返り点を削除する。加えて、その「得」が他のルールによる返り点の返り元である場合、返り元を、削除した返り点の返り元へ移動する。



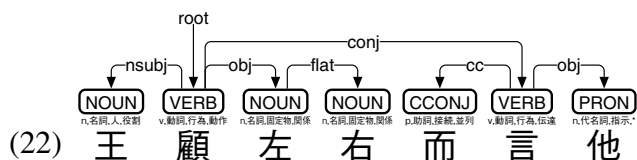
夫舜惡得而禁之

文例 21 では、「禁」から「得」への返り点を削除している。このルール㉑は、助動詞である「得而」を、動詞として訓読するためのものである。

3.4 返り点の付け替えに対するルール

ここまでのルールで打った返り点に対し、以下では、返り点を削除・移動・追加するルールを考える。

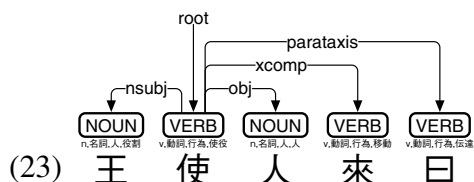
ルール㉒ ルール①～㉑で打った返り点 (ルール⑧を除く) の返り元から、右向きの **conj**・**clf**・**flat**・**case** リンクが出ている場合、それらのリンク先のうち文末に最も近いものへ、返り元を移動する。



王顧左右而言他

文例 22 では、ルール①にもとづいて「左」から「顧」へ、ルール①にもとづいて「他」から「言」へ、それぞれ返り点を打つべきところ、「左」から「顧」への返り元を「右」へ移動している。

ルール㉓ ルール①～㉒で打った返り点において、1つの返り先へ複数の返り元から返り点が集中している場合、それらの返り元のうち文末に最も近いものを残し、他の返り元は削除する。

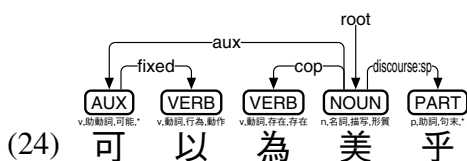


王使人來曰

文例 23 では、ルール①にもとづいて「人」から「使」へ、ルール④にもとづいて「來」から「使」へ、それぞれ返り点を打つべきところ、「人」からの返り点は削除している。

ルール②④ ルール①～③で打った返り点において、1つの返り元から複数の返り先へ返り点がある場合、それらの返り先のうち文頭に最も近いもの以外に返り元を追加し、返り点を後ろから順に辿る形とする。

可₃以₂為₁美₄乎₅



文例 24 では、ルール⑤にもとづいて「美」から「為」へ、ルール⑦にもとづいて「美」から「可」へ、それぞれ返り点を打つべきところ、「為」に返り元を追加して、「美」「為」「可」を順に辿る返り点としている。

- (1) 孟子見₁梁惠王₂〔卷一 1〕
- (2) 王立₁於₂沼上₃〔卷一 2〕
- (3) 女子之₁嫁也₂母命₃之₄〔卷六 2〕
- (4) 何由知₁吾可₂也〔卷一 7〕
- (5) 滕文公為₁世子₂〔卷五 1〕
- (6) 古之人與₁民偕樂₂〔卷一 2〕
- (7) 是則可₁憂也〔卷八 28〕
- (8) 非₁此之謂₂也〔卷四 2〕
- (9) 吾子與₁子路₂孰賢₃〔卷三 1〕
- (10) 而況於₁王乎₂〔卷四 9〕
- (11) 乃所₁謂善也〔卷十一 6〕
- (12) 王請大₁之〔卷二 3〕
- (13) 去之₁岐山之下₂居焉〔卷一 14〕
- (14) 如₁之何其可₂也〔卷二 11〕
- (15) 天下之不₁助₂苗長₃者寡矣〔卷三 2〕
- (16) 勸₁齊伐₂燕有₃諸〔卷四 8〕
- (17) 所₁以別₂野人也〔卷五 3〕
- (18) 所以自為₁則吾不₂知也〔卷四 5〕
- (19) 其何能淑〔卷七 9〕
- (20) 我不₁敢廢〔卷八 24〕
- (21) 夫舜惡得而禁₁之〔卷十三 35〕
- (22) 王顧₁左右而言₂他〔卷二 6〕
- (23) 王使₁人來₂曰〔卷四 2〕
- (24) 可₁以為₂美乎〔卷十一 8〕

図 3: 富山房『漢文大系』の「孟子定本」における返り点の文例

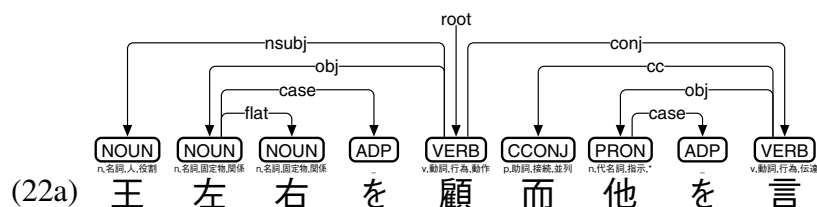
4 訓読における送り仮名の自動付与

返り点によって語順を入れ換えた古典中国語 UD に対し、送り仮名を付与して、書き下し文にしたい。訓読における送り仮名は、助詞と活用語尾に分けられることから、それぞれについて考えてみよう。

4.1 助詞の付与

目的語を表す格助詞

obj リンクおよび ccomp リンクについては、リンク先の単語を始点とする構成鎖の末尾に、目的語を表す格助詞を付与する。目的語を表す格助詞は「を」が代表的である。たとえば文例 22 では、返り点によって語順を入れ換えた上で、「左右」(「左」を始点とする構成鎖)と「他」の後に「を」を付与する。



ただし、動詞の種類によっては、目的語を表す格助詞に「を」以外が適切な場合もある。筆者の考えを表2に示すが、あくまで現時点での実装のために作成した表であり、今後さらなる改良をおこなう可能性がある。

表 2: 目的語を表す格助詞に「を」以外が適切な場合

格助詞	4 階層品詞	備考
に	v, 動詞, 行為, 移動	
に	v, 動詞, 行為, 伝達	obj リンク先が「n, 代名詞, 人称」「n, 名詞, 人」の場合
と	v, 動詞, 行為, 伝達	助動詞の「可」を伴う場合
の	v, 動詞, 行為, 分類	
が	v, 動詞, 行為, 分類	obj・ccomp リンク先が動詞の場合
をして	v, 動詞, 行為, 使役	
	v, 動詞, 存在, 存在	「有」「無」は格助詞を伴わない

斜格補語を表す格助詞

obl リンクについては、リンク先の単語を始点とする構成鎖の末尾に、斜格補語を表す格助詞を付与する。斜格補語を表す格助詞は「に」が代表的だが、case リンクを伴う場合は、その先にある前置詞の種類に応じて格助詞を変更する(表 3)とともに、前置詞を格助詞に同化させる。一方、case リンクを伴わない代名詞の「自」に対しては、格助詞を付与せずに「自」を「自ら」(みずから)とする。

表 3: 斜格補語を表す格助詞に「に」以外が適切な場合

格助詞	4 階層品詞	備考
より	v, 前置詞, 経由	
より	v, 前置詞, 基盤	obl リンク元が「v, 動詞, 描写, 量」の場合
ゆえ	v, 前置詞, 源泉	
と	v, 前置詞, 関係	

主語を表す格助詞

nsubj リンクおよび csubj リンクについては、リンク先の単語を始点とする構成鎖の末尾に、主語を表す格助詞「は」を付与する。ただし、主語を表す格助詞が複数並ぶ場合は、2つ目以降は「が」を用いる。一方、case リンクに「之」(p, 助詞, 接続, 属格)を伴っている場合は、格助詞を付与せずに「之」を「の」とする。

所有を表す格助詞

nmod リンクおよび det リンクについては、リンク先の単語を始点とする構成鎖の末尾に、所有を表す格助詞「の」を付与する。ただし、case リンクに「之」(p, 助詞, 接続, 属格)を伴っている場合は、格助詞を付与せずに「之」を「の」とする。

接続助詞

左向きの advcl リンクが「則」(v, 副詞, 時相, 緊接)をまたいでいる場合は、リンク先の単語を始点とする構成鎖の末尾に、接続助詞「ば」を付与する。「而」(p, 助詞, 接続, 並列)は「而して」とする。

句末の助詞

句末の助詞(p, 助詞, 句末)のうち、「乎」は「や」とする。「已」と「耳」は「のみ」とする。「兮」は「よ」とする。「也」は「なり」とした上で、ナリ活用(表7参照)をおこなう。その他の句末の助詞は「か」とするが、句末の助詞が複数並ぶ場合は読まない。また、「也」(p, 助詞, 提示)は「なる」とする。

句頭の助詞

句頭の助詞(p, 助詞, 句頭)のうち、「蓋」は「けだし」とする。その他の句頭の助詞は「それ」とする。

4.2 活用語尾の付与

文中の動詞(助動詞や一部の副詞を含む)には、以下に示す手順により活用語尾を付与する。なお、活用語尾の付与順序は、文末の動詞から文頭の動詞へと、順におこなう。

動詞の活用語尾

動詞(日本語では形容詞となる場合を含む)は、表4にもとづき、次の単語との間に活用語尾を付与する。ただし、動詞と次の単語の間に、左向きの **amod** リンクが繋がっている場合、あるいは右向きの **flat:vv** リンクが繋がっている場合は、活用語尾を付与しない。動詞と次の単語の間に、左向きの **advmod**・**advcl** リンクが繋がっている場合、右向きの **parataxis** リンクが繋がっている場合、あるいは直接のリンクが無い場合は、連用形に「て」を付与する。動詞と次の単語の間に、右向きの **aux** リンクが繋がっている場合は、以下のとおりとする。

- 次の単語が「得」であれば、連体形に「を」を付与する。
- 次の単語が「欲」であれば、未然形に「んと」を付与する。
- 次の単語が「被」「見」であれば、未然形とする。
- その他は、連体形とする。

表4に載っていない動詞については、旧仮名口語 UniDic^[9]に活用の種類(表7)を問い合わせる。旧仮名口語 UniDicに問い合わせても活用の種類が判明しない動詞は、文語サ行変格とみなす。**flat:vv** リンクのリンク先にあたる動詞も、文語サ行変格とみなす。

ただし、「況」(v, 動詞, 行為, 動作)は活用せず「況んや」とする。「所以」は「以てする所」ではなく「ゆゑん」とする。「所謂」は「いはゆる」とする。「如何」「奈何」「若何」は「いかん」とする。

助動詞の活用語尾

助動詞は、表5に基づいて、次の単語との間に活用語尾を付与する。「須」「宜」「儀」については再読文字とせず、「～すべし」とだけ読んでいる。なお、「敢」(v, 助動詞, 願望)は活用せず「敢へて」とし、直後が動詞以外の場合に限って「敢へてす」(文語サ行変格)で活用する。同様に、「肯」(v, 助動詞, 願望)は活用せず「肯へて」とし、直後が動詞以外の場合に限って「肯へてす」(文語サ行変格)で活用する。「能」(v, 助動詞, 可能)は活用せず「能く」とし、直後が否定の場合に限って「能ふ」(文語四段-ハ行)で活用する。

特殊な副詞の活用語尾

副詞のうち「v, 副詞, 否定」と「v, 副詞, 時相, 将来」については、表6に基づいて活用する。ただし、「未」(v, 副詞, 否定, 有界)は「未だ」と「ず」で動詞を挟み込み、「ず」を表6に基づいて活用することで、いわゆる再読文字とする。他の副詞については、ひらがなに置き換えるか「に」を付与することで、日本語の副詞に近づける。

^[9] 小木曾智信: 旧仮名遣いの口語文を対象とした形態素解析辞書, 人文科学とコンピュータ「じんもんこん 2012」 論文集 (2012年11月), pp.25-32.

表 4: 動詞活用表 (一部)

動詞		未然形	連用形	終止形	連体形	已然形	命令形
有	v, 動詞, 存在, 存在	有ら	有り	有り	有る	有れ	有れ
無	v, 動詞, 存在, 存在	無から	無く	無し	無き	無けれ	無かれ
於	v, 動詞, 存在, 存在	於てせ	於	於てす	於てする	於てせ	於てせよ
為	v, 動詞, 存在, 存在	たら	たり	たり	たる	たれ	たれ
為	v, 動詞, 行為, 生産	為さ	為し	為す	為す	為せ	為せ
來	v, 動詞, 行為, 移動	來	來	來る	來る	來れ	來よ
之	v, 動詞, 行為, 移動	ゆか	ゆき	ゆく	ゆく	ゆけ	ゆけ
行	v, 動詞, 行為, 移動	行か	行き	行く	行く	行け	行け
行	v, 動詞, 行為, 動作	行は	行ひ	行ふ	行ふ	行へ	行へ
以	v, 動詞, 行為, 動作	以てせ	以	以てす	以てする	以てせ	以てせよ
易	v, 動詞, 行為, 動作	易へ	易へ	易へる	易へる	易へれ	易へよ
易	v, 動詞, 描写, 形質	易から	易く	易し	易き	易けれ	易くせよ
重	v, 動詞, 描写, 量	重ね	重ね	重ねる	重ねる	重ねれ	重ねよ
如	v, 動詞, 行為, 分類	如くあら	如く	如し	如き	如くあれ	如くあれ
曰	v, 動詞, 行為, 伝達	曰は	曰ひ	曰く	曰ふ	曰へ	曰へ
説	v, 動詞, 行為, 伝達	説か	説き	説く	説く	説け	説け
説	v, 動詞, 行為, 態度	説ば	説び	説ぶ	説ぶ	説べ	説べ
可	v, 動詞, 描写, 態度	可とせ	可とし	可とす	可とする	可とすれ	可とせよ

表 5: 助動詞活用表

助動詞		未然形	連用形	終止形	連体形	已然形	命令形
得	v, 助動詞, 可能	得	得	得る	得る	得れ	得よ
可	v, 助動詞, 可能	べから	べく	べし	べき	べけれ	べけれ
須	v, 助動詞, 必要	べから	べく	べし	べき	べけれ	べけれ
宜	v, 助動詞, 必要	べから	べく	べし	べき	べけれ	べけれ
儀	v, 助動詞, 必要	べから	べく	べし	べき	べけれ	べけれ
欲	v, 助動詞, 願望	欲さ	欲し	欲す	欲する	欲せ	欲せよ

表 6: 特殊な副詞の活用表

副詞		未然形	連用形	終止形	連体形	已然形	命令形
不	v, 副詞, 否定, 無界	ざら	ずし	ず	ざる	ざれ	ざれ
弗	v, 副詞, 否定, 無界	ざら	ずし	ず	ざる	ざれ	ざれ
未	v, 副詞, 否定, 有界	ざら	ずし	ず	ざる	ざれ	ざれ
毋	v, 副詞, 否定, 禁止	なから	なく	なかれ	なき	なけれ	なかれ
勿	v, 副詞, 否定, 禁止	なから	なく	なかれ	なき	なけれ	なかれ
莫	v, 副詞, 否定, 禁止	なから	なく	なかれ	なき	なけれ	なかれ
非	v, 副詞, 否定, 体言否定	非ざら	非ずし	非ず	非ざる	非ざれ	非ざれ
將	v, 副詞, 時相, 将来	んとせ	んとし	んとす	んとする	んとすれ	んとせよ
且	v, 副詞, 時相, 将来	んとせ	んとし	んとす	んとする	んとすれ	んとせよ

表 7: 動詞活用表 (旧仮名口語 UniDic 問い合わせ用)

活用の種類	未然形	連用形	終止形	連体形	已然形	命令形
五段-バ行	□ば	□び	□ぶ	□ぶ	□べ	□べ
文語上二段-バ行	□び	□び	□ぶ	□びる	□びれ	□びよ
文語下二段-バ行	□べ	□べ	□ぶ	□べる	□べれ	□べよ
五段-ダ行	□だ	□ぢ	□づ	□づ	□で	□で
文語上二段-ダ行	□ぢ	□ぢ	□づ	□ぢる	□ぢれ	□ぢよ
文語下二段-ダ行	□で	□で	□づ	□でる	□でれ	□でよ
五段-サ行	□さ	□し	□す	□す	□せ	□せ
文語ザ行変格	□ぜ	□じ	□ず	□ずる	□じれ	□ぜよ
五段-マ行	□ま	□み	□む	□む	□め	□め
文語上二段-マ行	□み	□み	□む	□みる	□みれ	□みよ
文語下二段-マ行	□め	□め	□む	□める	□めれ	□めよ
五段-ワア行	□は	□ひ	□ふ	□ふ	□へ	□へ
文語四段-ハ行	□は	□ひ	□ふ	□ふ	□へ	□へ
文語上二段-ハ行	□ひ	□ひ	□ふ	□ひる	□ひれ	□ひよ
文語下二段-ハ行	□へ	□へ	□ふ	□へる	□へれ	□へよ
五段-タ行	□た	□ち	□つ	□つ	□て	□て
文語上二段-タ行	□ち	□ち	□つ	□ちる	□ちれ	□ちよ
文語下二段-タ行	□て	□て	□つ	□てる	□てれ	□てよ
五段-ガ行	□が	□ぎ	□ぐ	□ぐ	□げ	□げ
文語上二段-ガ行	□ぎ	□ぎ	□ぐ	□ぎる	□ぎれ	□ぎよ
文語下二段-ガ行	□げ	□げ	□ぐ	□げる	□げれ	□げよ
文語形容詞-ク	□から	□く	□し	□き	□けれ	□くせよ
文語形容詞-シク	□しから	□しく	□し	□しき	□しけれ	□しくせよ
五段-カ行	□か	□き	□く	□く	□け	□け
文語上二段-カ行	□き	□き	□く	□きる	□きれ	□きよ
文語下二段-カ行	□け	□け	□く	□ける	□けれ	□けよ
五段-ラ行	□ら	□り	□る	□る	□れ	□れ
上一段	□	□	□る	□る	□れ	□よ
下一段	□	□	□る	□る	□れ	□よ
ナリ活用	□なら	□なり	□なり	□なる	□なれ	□なれ
文語サ行変格	□せ	□し	□す	□する	□すれ	□せよ

5 漢文自動訓読ツール UD-Kundoku

ここまで述べてきた手法を用いて、漢文自動訓読ツール UD-Kundoku を python3 モジュールとして実装した。UD-Kundoku は、いわゆる encode-reorder-decode モデル^[10]を採用しており、以下の3つのステップから構成される。

1. 白文を依存文法解析して、古典中国語 UD を生成する (encode)
2. 返り点に基づいて、語順を入れ替える (reorder)
3. 送り仮名を付与する (decode)

「孟子見梁惠王王曰叟不遠千里而來」という白文に対する UD-Kundoku の動作の概要を、図4に示す。UD-Kundoku は、内部に UD-Kanbun^[11] と UniDic2UD^[12] と旧仮名口語 UniDic を内蔵しており、encode と返り点の生成は UD-Kanbun が、decode のうち活用語尾の生成は UniDic2UD + 旧仮名口語 UniDic が、それぞれ分担している。なお、入力された白文が長い (10 文字以上) 場合には、古詩文断句^[13] で文切りを試し、それが上手くいかない場合は、UD-Kanbun 内蔵の UDPipe^[14] で文切りをおこなっている。

UD-Kundoku の評価をおこなうべく、大学入試センター試験『国語』の令和2年度本試験から、第4問本文(図5)の返り点や送り仮名を除去し、**C**に「窓」を入れ(図8問4の正解①)、以下の白文として準備した。

樵隱俱在山 由来事不同 不同非一事 養痾亦園中 園中屏氛雜
清曠招遠風 卜室倚北阜 啓扉面南江 激澗代汲井 挿槿当列墉
群木既羅戸 衆山亦对窓 靡迤趨下田 迢遞瞰高峰 寡欲不期勞
即事罕人功 唯開蔣生徑 永懷求羊蹤 賞心不可忘 妙善冀能同

この20句を、UD-Kundoku 1.0.6 で自動訓読させてみたところ、図6が結果として得られた。各句ごとに書き下し文を準備し、一文字を一単語とみなして BLEU^[15] および RIBES^[16] で評価した結果を、表8に示す。BLEU は、NLTK 3.4.5 の method3 でスムージング (NIST geometric sequence smoothing) をおこなったものの、評価値の高低が訓読の良し悪しに連動していない。一方、RIBES は、句の構造を取り違えた際に評価値がグンと下がるという点で、全体として直感に合う評価となっている。筆者の感触としては、RIBESの方が自動訓読の評価に適しているようである。

^[10]Josep M. Crego and José B. Mariño: Integration of POS-tag-based Source Reordering into SMT Decoding by an Extended Search Graph, AMTA2006: 7th Conference of the Association for Machine Translation in the Americas (August 2006), pp.29-36.

^[11]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 (2019年3月8日), pp.3-110.

^[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 (2019年12月), pp.43-50.

^[13]胡鞠奮, 李紳, 諸雨辰: 基於深層語言模型的古漢語知識表示及自動斷句研究, CCL2019: 18th China National Conference on Computational Linguistics (2019年10月).

^[14]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.

^[15]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu: BLEU: a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (July 2002), pp.311-318.

^[16]平尾努, 磯崎秀樹, 須藤克仁, Duh Kevin, 塚田元, 永田昌明: 語順の相関に基づく機械翻訳の自動評価法, 自然言語処理, 第21巻, 第3号 (2014年6月), pp.421-444.

孟子見梁惠王王曰叟不遠千里而來

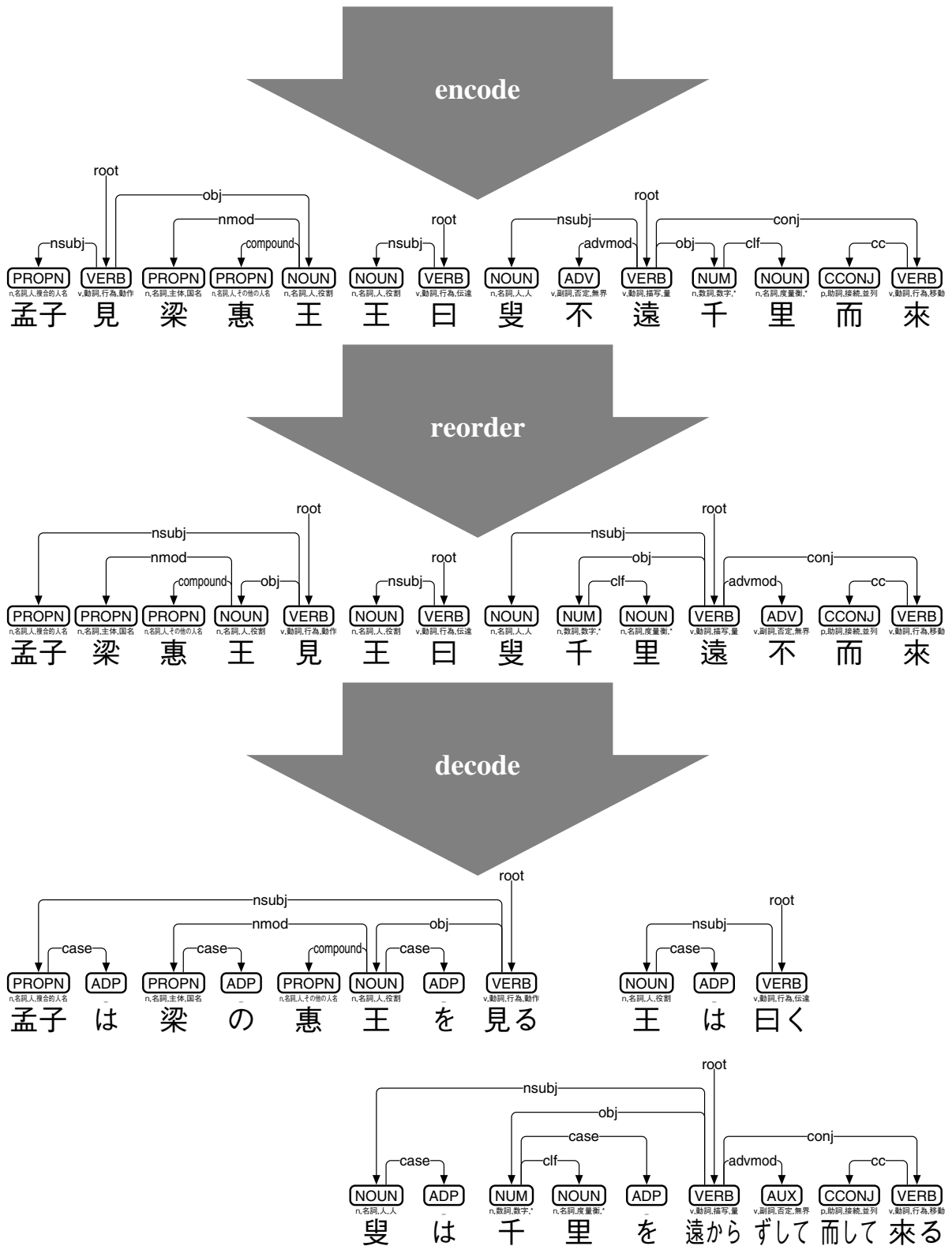


図 4: 「孟子見梁惠王王曰叟不遠千里而來」に対する自動訓読の流れ

<p>E</p> <p>賞^(注13) 心 不 可^{カラ} 忘^ル</p> <p>妙^(注14) 善 冀^{こひねが} 能^{よク} 同^{ともニセン}</p>	<p>(イ)</p> <p>唯^タ 開^キ 蔣^(注11) 生^{セイ} 徑^{みちヲ}</p> <p>永^ク 懷^{おもフ} 求^(注12) 羊^{やウ} 蹤^{あとヲ}</p>	<p>(イ)</p> <p>寡^カ 欲^ヨ 不^レ 期^セ 勞^ヲ</p> <p>即^{シテ} 事^ニ 罕^(注10) 人^ノ 功^ニ</p>	<p>D</p> <p>靡^(注8) 迤^{イトシテ} 趨^{おもむキ} 下^ニ 田^ニ</p> <p>迢^(注9) 遙^{てうていトシテ} 瞰^{みル} 高^ニ 峰^ヲ</p>	<p>群^ニ 木^ニ 既^ニ 羅^{つらなり} 戸^ニ</p> <p>衆^ニ 山^ニ 亦^タ 對^ス C^ニ</p>	<p>激^{せきと} 澗^{たにがは} 代^ヘ 汲^{くム} 井^ニ</p> <p>挿^{ウエテ} 槿^{むくげヲ} 當^ツ 列^{つらな} 壙^{ルニかき} 壙^ニ</p>	<p>B</p> <p>ト^(注7) 室^ヲ 倚^{より} 北^ノ 阜^{をかニ}</p> <p>啓^{ひらキテ} 扉^ヲ 面^ス 南^ノ 江^{かはニ}</p>	<p>園^ニ 中^{しりぞケ} 屏^ニ 氛^(注5) 雜^ヲ</p> <p>清^(注6) 曠^{くわう} 招^ク 遠^ニ 風^ヲ</p>	<p>不^レ 同^ニ 非^レ 一^ニ 事^ニ</p> <p>養^(注3) 痾^{やまひヲ} 亦^タ 園^(注4) 中^ニ</p>	<p>樵^(注1) 隱^(ア) 俱^ニ 在^{ルモ} 山^ニ</p> <p>由^(注2) 来^ル 事^ニ 不^レ 同^ニ</p>
---	--	---	---	--	---	---	--	--	--

図 5: 大学入試センター試験『国語』(2020年1月18日)第4問本文

表 8: 自動訓読結果(図6)の評価

自動訓読結果	BLEU	RIBES	書き下し文
樵隱して俱に山に在り	0.451801	0.914691	樵隱俱に山に在るも
来ゆえ事は同じとせず	0.290715	0.880112	由来事は同じからず
一事に非ざるを同じとせず	0.295023	0.413602	同じからざるは一事に非ず
痾を養ひてまた園の中	0.205567	0.880112	痾を養ふも亦た園中
園の中は氛雑へるを屏く	0.117312	0.859389	園中氛雑を屏け
清曠して遠風を招く	0.513345	0.939104	清曠遠風を招く
ト室は北の阜を倚る	0.175638	0.699783	室をトして北の阜に倚り
啓は扉に南の江を面す	0.205848	0.723234	扉を啓きて南の江に面す
澗を激して代へて井を汲む	0.465954	0.743122	澗を激めて井を汲むに代へ
槿を挿して列壙を当る	0.163420	0.860799	槿を挿ゑて壙に列るに当つ
群の木はすでに戸を羅す	0.094252	0.821097	群木既に戸に羅り
衆山はまた窓を對ふ	0.142588	0.863340	衆山亦た窓に対す
靡迤なるは下の田に趨く	0.183603	0.859389	靡迤として下田に趨き
迢なりて高峰を遙瞰す	0.302138	0.740464	迢遙として高峰を瞰る
寡く欲して勞を期さず	0.435315	0.910739	欲を寡くして勞を期せず
即ち事へて人の功を罕し	0.279016	0.725586	事に即して人の功罕なり
ただ蔣は生徑を開く	0.233569	0.903602	唯だ蔣生の徑を開き
永く羊の蹤を求めるを懷ふ	0.341723	0.849282	永く求羊の蹤を懷ふ
心を賞でて忘ざるべからず	0.155801	0.832319	賞心忘るべからず
妙く善く冀ふは能く同じ	0.193956	0.861259	妙善冀はくは能く同にせんことを
全体	0.262779	0.814051	

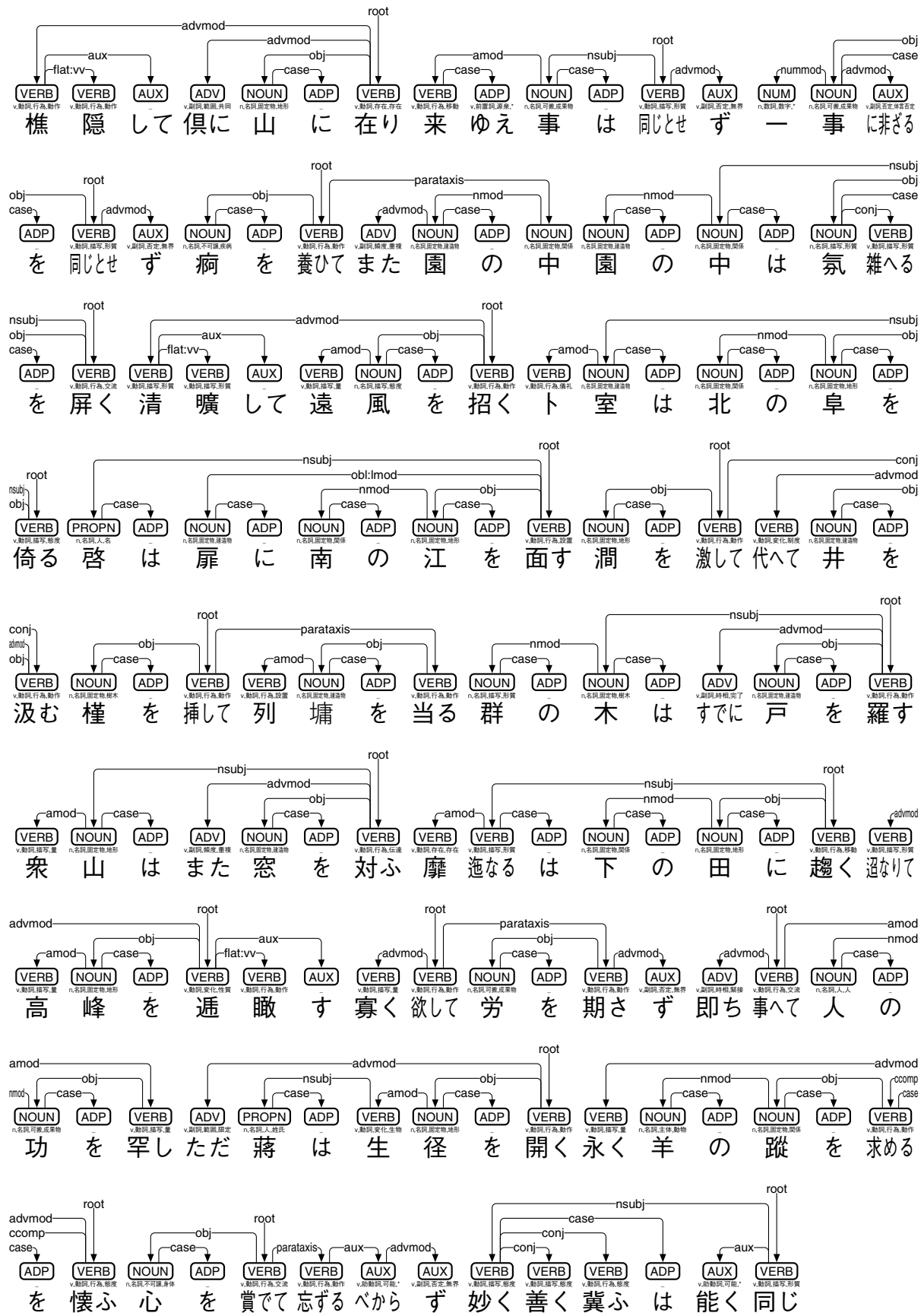


図 6: UD-Kundoku 1.0.6 による自動訓読結果

表8中、RIBES 評価値が最も低い「不同非一事」(第3句)に対し、UD-Kundokuの encode-reorder-decode モデルが、どのステップで評価値を下げてしまっているのか、さらに検討をおこなった。そうしたところ「不同非一事」に関しては、encode ステップで致命的な誤りが発生していることがわかった。「不同非一事」は、正しくは「不同」を節主語(csubj)とするコピュラ文(図7(b))なのだが、UD-Kundoku 1.0.6の encode は、「非一事」を「同」の目的語(obj)だと認識してしまう(図7(a))。この結果、reorder ステップでは、UD-Kundoku 1.0.6は「不同非一事」と返り点を打ってしまう。図8の間2によれば、正解は②「不同非一事」である。encode ステップでの誤りが、結果としてRIBES 評価値を下げてしまうと同時に、間2にも正解できない。



図7: 「不同非一事」に対する古典中国語 UD

UD-Kundoku 1.0.6の encode ステップは、現状ではUD-Kanbun 1.9.0に依拠しており、『孟子』『論語』『禮記』などのUDを機械学習したものである。これらの文例における「不同」の用法は、たとえば「不同榘榘不同巾櫛」のように「同」が目的語を取る例が多く、「不同」を節主語とする用例は一つもない。結果として「不同非一事」の「不同」を、図7(b)のように正しく encode することは、現時点のUD-Kundokuでは困難と言わざるを得ない。

6 おわりに

古典中国語 Universal Dependencies にもとづく漢文自動訓読ツール UD-Kundoku を、encode-reorder-decode モデルによる python3 モジュールとして実装し、PyPI で公開^[17]した。encode(依存文法解析)は機械学習、reorder(返り点による語順入れ替え)と decode(送り仮名の付与)はルールベースである。

ただ、筆者の感触としては、decodeのうち格助詞の付与は、機械学習の方が良いのではないかと、思える。というのも、日本語の格助詞は、われわれの4階層品詞やUD依存構造タグと今一つ連動しておらず、手作業でルールを書いても、なかなか上手くいかないからだ。一方で、decodeのうち活用語尾の付与は、どう考えてもルールベースでおこなうべきだし、reorderもルールベースの方が良い(気がする)^[18]ので、そのあたりをどうハイブリッドにするか、今後の課題になると考えられる。

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。

[17]<https://pypi.org/project/udkundoku/>

[18]Jinhua Du, Andy Way: Pre-Reordering for Neural Machine Translation: Helpful or Harmful?, Prague Bulletin of Mathematical Linguistics, No.108 (June 2017), pp.171-182.

問2 傍線部A「由来事不同、不同非一事」について、(a)返り点の付け方と、(b)書き下し文との組合せとして最も適当な

ものを、次の①～⑤のうちから一つ選べ。解答番号は 31。

- | | | | | |
|---|-----|--|-----|--------------------------|
| ⑤ | (a) | 由来事 _レ 不同、不同 _レ 非 _二 一事 _一 | (b) | 由来事は同じからず、一事を非とするを同じうせず |
| ④ | (a) | 由来事 _レ 不同、不同 _レ 非 _一 事 _二 | (b) | 由来事は同じうせず、非を同じうせずんば事を一にす |
| ③ | (a) | 由来事 _レ 不同、不同 _二 非 _一 事 _二 | (b) | 由来事は同じうせず、一に非ざる事を同じうせず |
| ② | (a) | 由来事 _レ 不同、不同 _二 非 _一 事 _二 | (b) | 由来事は同じからず、同じからざるは一事に非ず |
| ① | (a) | 由来事 _レ 不同、不同 _レ 非 _二 一事 _一 | (b) | 由来事は同じからず、一事を非とするを同じうせず |

問4

空欄

C

に入る文字として最も適当なものを、次の①～⑤のうちから一つ選べ。解答番号は

33。

- | | | | | |
|---|---|---|---|---|
| ⑤ | ④ | ③ | ② | ① |
| 月 | 門 | 虹 | 空 | 窓 |

図 8: 大学入試センター試験『国語』(2020年1月18日)第4問の問4・問2