四書を学んだMeCab＋UDPipeは センター試験の漢文を読めるのか

安岡孝一*

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 の構造的統語論 [2] に源を発し、Мельчук の有向グラフ記述 [3] 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を可能としているのである。

しかしながら、古典中国語の白文は、単語の間にも文の間にも区切りを持たず、その点で、欧米諸語とは異なる処理手法を必要とする。すなわち、白文の解析においては、まず、単語の区切りを認識することが必須であり、そのためには形態素解析をおこなわねばならない。入力された白文に対し、形態素解析 [4-13] により、単語切りをおこなうと同時に、各単語の品詞を得る。その後に、依存文法解析 [14-19] により、係り受け関係 (単語間のリンク) を解析すると同時に、文の切れ目を得る、という手順になる。

これまでにわれわれは、古典中国語の形態素解析に MeCab [20] を用いてきた。MeCab は、もともとは日本語向けの形態素解析エンジンだったが、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。一方、古典中国語の依存文法解析には、UDPipe [21] が向いているようである。UDPipe は、チェコ語の係り受け解析エンジンが土台となっているが、UD への設計拡張をおこなった際に、単語切り・品詞付与・文切りの機能が追加され、形態素解析と依存文法解析の両方を兼ね備えたものとなっている。ただ、UDPipe の形態素解析機能が、MeCab の形態素解析性能を凌駕しているのかどうかは、多少、疑問がある。

この問題に対し、われわれは、MeCab の形態素解析機能を、どのように UDPipe の依存文法解析機能と接続していくかについて、いくつかの比較実験をおこなった。学習用のコーパスは四書 (孟子・論語・大学・中庸) を用いた。実験対象は、この 5 年間に実施された大学入試センター試験『国語』の本試験から、第 4 問の本文を用いた。また、比較実験の結果に対して、多少の改善を試みた。以下、実験と改善の概要を、順に述べる。

*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

2　MeCab + UDPipe によるセンター漢文の解析

白文の形態素解析および依存文法解析は、単語切り・品詞付与・係り受け解析・文切りの 4 段階から成る。MeCab 0.996 と UDPipe 1.2.0 を、これらの 4 段階に適用するにあたって、われわれは、以下の 3 つの手法を比較することにした。

手法① 単語切り＋品詞付与＋係り受け解析＋文切りの全てに UDPipe を用いる

手法② 単語切りに MeCab を、品詞付与＋係り受け解析＋文切りに UDPipe を用いる

手法③ 単語切り＋品詞付与に MeCab を、係り受け解析＋文切りに UDPipe を用いる

MeCab と UDPipe は、いずれも、Kanripo [22] をもとに手作業で作成した孟子・論語・大学・中庸の全文 UD コーパス (合計 56767 字、異なり字数 2333 字、文字コードは UTF-8) を学習済みである。全文 UD コーパスの一部を、以下に示す。

```
# newdoc id = KR1h0001_001
# text = 梁惠王上
1	梁	梁	PROPN	n,名詞,主体,国名	Case=Loc|NameType=Nat	3	nmod	_	Gloss=[country-name]|SpaceAfter=No
2	惠	惠	PROPN	n,名詞,人,その他の人名	NameType=Prs	3	compound	_	Gloss=Hui|SpaceAfter=No
3	王	王	NOUN	n,名詞,人,役割	_	0	root	_	Gloss=king|SpaceAfter=No
4	上	上	NOUN	n,名詞,固定物,関係	Case=Loc	3	list	_	Gloss=up|SpacesAfter=\n

# newpar
# text = 孟子見梁惠王
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

# text = 王曰
1	王	王	NOUN	n,名詞,人,役割	_	2	nsubj	_	Gloss=king|SpaceAfter=No
2	曰	曰	VERB	v,動詞,行為,伝達	_	0	root	_	Gloss=say|SpaceAfter=No
```

MeCab 辞書には、これら四書に出現する単語 (形態素) に加え、図 1 左のように、固有名詞などを大量に追加 [11] している。さらに、この MeCab 辞書を変形し、UDPipe の外部辞書 (図 1 右) にも用いることで、できるだけ実験条件を揃えるようにした。UDPipe の文切りは、`--tokenizer=joint_with_parsing` オプションにより、係り受け解析と連動させている。上記の 3 手法の比較実験対象としては、過去 5 年間のセンター試験『国語』

```
江,0,0,0,n,名詞,人,名,*,*,江,*,*,[given-name]
江,0,0,0,n,名詞,人,姓氏,*,*,江,*,*,[surname]
江,0,0,0,n,名詞,固定物,地名,*,*,江,*,*,[place-name]
江中,0,0,0,n,名詞,固定物,地名,*,*,江中,*,*,[place-name]
江北,0,0,0,n,名詞,固定物,地名,*,*,江北,*,*,[place-name]
江南,0,0,0,n,名詞,固定物,地名,*,*,江南,*,*,[place-name]
江壅,0,0,0,n,名詞,固定物,地名,*,*,江壅,*,*,[place-name]
江夏,0,0,0,n,名詞,固定物,地名,*,*,江夏,*,*,[place-name]
江寧,0,0,0,n,名詞,固定物,地名,*,*,江寧,*,*,[place-name]
江州,0,0,0,n,名詞,固定物,地名,*,*,江州,*,*,[place-name]
江戸,0,0,0,n,名詞,固定物,地名,*,*,江戸,*,*,[place-name]
江戸,0,0,0,n,名詞,固定物,地名,*,*,江戸,*,*,[place-name]
江東,0,0,0,n,名詞,固定物,地名,*,*,江東,*,*,[place-name]
江浙,0,0,0,n,名詞,固定物,地名,*,*,江浙,*,*,[place-name]
江淮,0,0,0,n,名詞,固定物,地名,*,*,江淮,*,*,[place-name]
江西,0,0,0,n,名詞,固定物,地名,*,*,江西,*,*,[place-name]
江進,0,0,0,n,名詞,人,名,*,*,江進,*,*,[given-name]
江陵,0,0,0,n,名詞,固定物,地名,*,*,江陵,*,*,[place-name]
江都,0,0,0,n,名詞,固定物,地名,*,*,江都,*,*,[place-name]
江都,0,0,0,n,名詞,固定物,地名,*,*,江都,*,*,[place-name]
```

```
江	江	PROPN	n,名詞,人,名	NameType=Giv
江	江	PROPN	n,名詞,人,姓氏	NameType=Sur
江	江	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江中	江中	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江北	江北	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江南	江南	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江壅	江壅	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江夏	江夏	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江寧	江寧	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江州	江州	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江戸	江戸	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江戸	江戸	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江東	江東	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江浙	江浙	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江淮	江淮	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江西	江西	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江進	江進	PROPN	n,名詞,人,名	NameType=Giv
江陵	江陵	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江都	江都	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
江都	江都	PROPN	n,名詞,固定物,地名	Case=Loc|NameType=Geo
```

図 1: MeCab 辞書 (左) と UDPipe の外部辞書 (右) における固有名詞の例

の漢文 (図 3・23・43・63・83) から、句読点や返り点などを除去し、完全な白文を準備した。なお、センター試験の漢文は常用漢字で書かれているが、白文の字体を変更するようなことはせず、その代わり、MeCab 辞書と UDPipe 外部辞書に常用漢字の単語を追加した。処理結果を図 4〜6・24〜26・44〜46・64〜66・84〜86 に示す。

これらの処理結果を定量的に比較すべく、LAS (Labeled Attachment Score)・MLAS (Morphology-aware Labeled Attachment Score)・BLEX (Bi-LEXical dependency score) の 3 つの指標 [23] を用いることにした。ただし、これらの指標を用いるためには、問題に対する正しい答が必要となることから、手作業で「正解」UD (図 22・42・62・82・102) を準備した。比較結果を表 1 に示す。

表 1: 手法①②③の LAS/MLAS/BLEX

	2019 年	2018 年	2017 年	2016 年	2015 年
手法①	45.48/39.87/43.09	42.86/37.54/38.77	49.74/40.40/43.05	50.66/45.96/47.83	38.74/35.73/36.80
手法②	41.96/38.10/38.73	45.33/38.98/40.26	51.78/43.23/46.45	46.28/43.48/44.10	42.03/40.96/41.49
手法③	44.14/39.62/42.14	48.73/42.68/43.95	49.24/42.72/44.66	47.34/43.61/44.24	42.03/40.43/41.49

この結果を見る限り、手法③が他よりは良さそうに見える。しかしながら、LAS・MLAS・BLEX はいずれも 100 点満点の指標であり、その意味で表 1 は、全体に低調と言わざるを得ない。一体、どこに問題があるのか。

3 文切りの改善

手法①②③の処理結果 (図 4〜6・24〜26・44〜46・64〜66・84〜86) を見る限り、筆者としては、どうも文切りがうまくいってない印象を受ける。たとえば、2019 年の手法① (図 4) を正解 (図 22) と見比べてみると、正しく文が切り出せているのは「非敢当是也」「亦為報也」「県君有焉」の 3 文だけであり、他はズレてしまっている。文切りに失敗して、泣き別れになってしまった単語間のリンクは、必ず不正解になってしまうことから、これが LAS・MLAS・BLEX を引き下げてしまっているのだと考えられる。

この仮説を検証すべく、文切りを改善することを考えた。具体的には、センター試験の問題本文 (図 3・23・43・63・83) の句読点をもとに、文切り情報を事前に準備し、UDPipe の文切りに代えることとした。実際に比較したのは、以下の 3 つの手法である。

手法1 文切りを事前に与え、単語切り＋品詞付与＋係り受け解析に UDPipe を用いる

手法2 文切りを事前に与え、単語切りに MeCab を、品詞付与＋係り受け解析に UDPipe を用いる

手法3 文切りを事前に与え、単語切り＋品詞付与に MeCab を、係り受け解析に UDPipe を用いる

処理結果を図 7〜9・27〜29・47〜49・67〜69・87〜89 に、比較結果を表 2 に示す。手法2と手法3が競っているものの、表 1 に比べると、LAS・MLAS・BLEX ともに改善され

ている。UDPipe の文切りオプション`--tokenizer=joint_with_parsing`は、非常に先進的な機能ではあるものの、現時点では、まだ試験的実装の域を出ていないということだろう。

表 2: 手法1 2 3の LAS/MLAS/BLEX

	2019 年	2018 年	2017 年	2016 年	2015 年
手法1	71.58/67.72/70.89	51.10/45.40/47.85	60.05/50.49/54.37	59.47/53.87/56.97	58.11/55.70/56.23
手法2	70.84/67.51/70.03	61.76/56.96/56.33	62.28/52.73/56.59	63.83/58.31/61.44	58.45/56.08/56.61
手法3	68.12/63.95/67.08	67.42/61.39/62.66	57.22/49.20/52.40	57.98/54.21/55.45	57.00/55.03/55.03

4　返り点をヒントにした品詞改善

センター試験の問題本文 (図 3・23・43・63・83) には、返り点が打たれている。これを元にして、手法1 2 3の処理結果 (図 7〜9・27〜29・47〜49・67〜69・87〜89) を改善できないだろうか。

たとえば、2018 年 (図 23) の 5 行目には、「所」←「沢」←「民」という返り点 (上中下点) が打たれている。すなわち「沢」は、「民」からの返り先であると同時に、「所」への返り元である。古典中国語 UD と返り点の関係 (図 2) を考慮すると、この場合、「沢」の品詞は VERB・AUX・PART のいずれかにしか成り得ず、図 27〜29 の「沢」の品詞 NOUN は誤りである。NOUN は、返り元には現れ得るが、返り先には現れ得ないので、返り先かつ返り元である「沢」は NOUN では有り得ない。

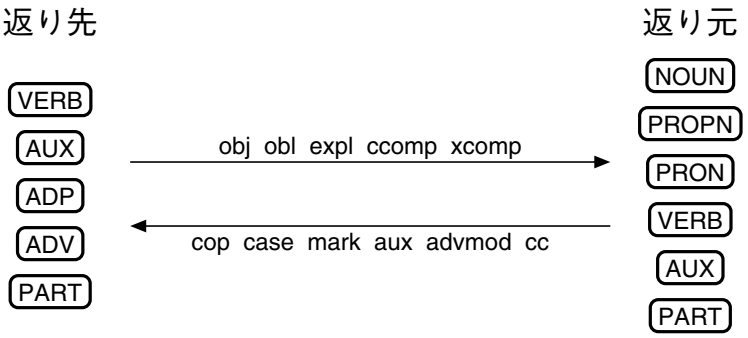


図 2: 古典中国語 UD と返り点の関係 [18]

このようなやり方で、問題本文中の全ての返り点をヒントに、手法1 2 3を、それぞれ改善するような手法を試した。

手法❶ 文切りを事前に与え、単語切り＋品詞付与に UDPipe を用いた後、返り点による品詞改善をおこない、係り受け解析に UDPipe を用いる

手法❷ 文切りを事前に与え、単語切りに MeCab を、品詞付与に UDPipe を用いた後、返り点による品詞改善をおこない、係り受け解析に UDPipe を用いる

手法❸ 文切りを事前に与え、単語切り＋品詞付与に MeCab を用いた後、返り点による品詞改善をおこない、係り受け解析に UDPipe を用いる

処理結果を図 10〜12・30〜32・50〜52・70〜72・90〜92 に、比較結果を表 3 に示す。2019 年・2018 年・2017 年は、表 2 より微妙な改善が見られるものの、2016 年と 2015 年は逆に悪化している。非常に残念だが、手法❶❷❸は、かなりの作業量を必要とする割に、効果が薄い手法のようである。

表 3: 手法❶❷❸の LAS/MLAS/BLEX

	2019 年	2018 年	2017 年	2016 年	2015 年
手法❶	72.68/69.21/72.38	55.49/48.93/52.60	64.12/55.19/59.74	48.42/41.49/42.72	55.69/51.46/54.11
手法❷	71.93/68.99/71.52	65.16/59.94/59.94	66.33/57.61/62.14	52.13/45.14/46.39	55.07/50.79/53.44
手法❸	68.66/64.78/67.92	68.56/62.03/63.29	61.27/53.55/57.42	48.40/42.37/42.99	55.56/51.85/53.44

5　おわりに

大学入試センター試験の漢文を題材に、四書 (孟子・論語・大学・中庸) を学んだ MeCab と UDPipe を用いて、古典中国語の単語切り・品詞付与・係り受け解析・文切りをおこなった。結論としては、単語切りに関しては MeCab の能力が高く、品詞付与では MeCab と UDPipe がトントンで、係り受け解析と文切りは UDPipe に任せるしかない。ただし、UDPipe の文切りは、古典中国語では良い結果が得られず、可能であれば、文切り情報を事前に準備した方がよい。一方、返り点をヒントにした品詞改善は、労おおくして功すくなし、という感じである。

なお、本稿で学習に用いた古典中国語 UD コーパスや、MeCab・UDPipe などの学習済みモデルは

`http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/2019-03-08/`

で公開している。古典中国語 (漢文) に関する様々な用途に、どしどし使ってほしい。

付録　MeCab ＋ StanfordNLP によるセンター漢文の解析

本稿の執筆中 (2019 年 1 月 31 日) に、StanfordNLP 0.1.1 [24] がリリースされた。本稿の〆切まで時間が残されておらず、十分な解析は出来なかったものの、四書の学習、および MeCab との接続は実現できたので、以下の 9 つの手法を比較した。

手法Ⓐ 単語切り＋品詞付与＋係り受け解析＋文切りの全てに StanfordNLP を用いる

手法Ⓑ 単語切りに MeCab を、品詞付与＋係り受け解析＋文切りに StanfordNLP を用いる

手法Ⓒ 単語切り＋品詞付与に MeCab を、係り受け解析＋文切りに StanfordNLP を用いる

手法🅐 文切りを事前に与え、単語切り＋品詞付与＋係り受け解析に StanfordNLP を用いる

手法🅑 文切りを事前に与え、単語切りに MeCab を、品詞付与＋係り受け解析に Stanford NLP を用いる

手法🅒 文切りを事前に与え、単語切り＋品詞付与に MeCab を、係り受け解析に Stanford NLP を用いる

手法❹ 文切りを事前に与え、単語切り＋品詞付与に StanfordNLP を用いた後、返り点による品詞改善をおこない、係り受け解析に StanfordNLP を用いる

手法❺ 文切りを事前に与え、単語切りに MeCab を、品詞付与に StanfordNLP を用いた後、返り点による品詞改善をおこない、係り受け解析に StanfordNLP を用いる

手法❻ 文切りを事前に与え、単語切り＋品詞付与に MeCab を用いた後、返り点による品詞改善をおこない、係り受け解析に StanfordNLP を用いる

処理結果を図 13〜21・33〜41・53〜61・73〜81・93〜101 に、比較結果を表 4 に示す。外部辞書を StanfordNLP に接続する方法を見つけられなかったため、手法ⒶⒷ🅐🅑では、常用漢字などの品詞に間違いが多く、MLAS が非常に低くなっている。

表 4: 手法ⒶⒷⒸ🅐🅑🅒❹❺❻の LAS/MLAS/BLEX

	2019 年	2018 年	2017 年	2016 年	2015 年
手法Ⓐ	31.40/28.12/28.12	39.34/33.54/34.16	43.04/32.90/36.13	32.28/27.27/28.48	36.01/32.88/31.27
手法Ⓑ	33.79/29.02/30.28	42.49/34.50/36.42	43.04/32.90/36.13	34.57/28.75/30.58	35.27/32.09/30.48
手法Ⓒ	40.33/36.19/38.10	43.06/35.03/37.58	46.08/38.83/40.78	42.55/37.42/39.26	43.48/41.69/41.69
手法🅐	46.70/38.34/40.26	51.38/43.48/45.34	48.61/37.54/40.78	39.68/33.33/33.33	45.26/41.71/41.18
手法🅑	47.41/39.24/40.51	60.62/50.32/55.41	48.61/37.54/40.78	42.02/34.89/36.14	45.41/40.96/40.96
手法🅒	62.67/55.70/60.13	66.29/59.94/58.68	54.68/47.74/50.32	51.60/45.34/47.83	52.66/49.74/50.79
手法❹	50.00/42.81/42.81	51.38/43.48/45.34	49.11/38.19/40.78	41.27/34.46/35.08	45.26/41.71/41.18
手法❺	50.68/43.67/43.04	61.76/51.59/56.69	49.11/38.19/40.78	43.62/36.02/37.89	45.41/40.96/40.96
手法❻	62.67/55.70/60.13	66.86/60.57/59.31	55.70/49.03/51.61	51.06/45.34/47.83	52.66/49.74/50.79

また、ところどころにリンクの交差が散見される。われわれの古典中国語 UD では、リンクの交差は起こらない [15, 16] のだが、StanfordNLP は、リンクの交差をターゲットにしたアルゴリズム [25] を採用している。これが結果的にミスマッチとなっており、LAS・MLAS・BLEX を引き下げてしまっているように思われる。

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] 守岡知彦: MeCab を用いた古典中国語の形態素解析の試み, 情報処理学会研究報告, Vol.2008-CH-79 (2008 年 7 月), pp.17-22.

[5] 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 (2009 年 10 月), No.3, pp.1-5.

[6] Tomohiko Morioka: A Prototype of a Classical Chinese Morphological Analyzer based on MeCab, Proceedings of Osaka Symposium on Digital Humanities 2011 (September 2011), p.36.

[7] 守岡知彦: 古典中国語形態素コーパス編集システムの開発, 東洋学へのコンピュータ利用, 第 23 回研究セミナー (2012 年 3 月), pp.75-83.

[8] 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.

[9] Tomohiko Morioka, Christian Wittern, Koichi Yasuoka, Naoki Yamazaki: A Study of Linguistic Analysis for Classical Chinese Texts, Proceedings 2013 International Conference on Culture and Computing (September 2013), pp.143-144.

[10] Koichi Yasuoka, Naoki Yamazaki, Christian Wittern, Yoshihiro Nikaido, Tomohiko Morioka: A Morphological Analysis of Classical Chinese Texts, Proceedings of Digital Humanities 2014 (July 2014), pp.410-412.

[11] 安岡孝一, 守岡知彦, Christian Wittern, 山崎直樹, 二階堂善弘, 鈴木慎吾: 古典中国語形態素解析による地名の自動抽出, 人文科学とコンピュータシンポジウム「じんもんこん 2014」論文集 (2014 年 12 月), pp.63-68.

[12] 安岡孝一, Christian Wittern, 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, 東洋学へのコンピュータ利用, 第 27 回研究セミナー (2016 年 3 月), pp.3-14.

[13] 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.

[14] 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.

[15] 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.

[16] 安岡孝一: 古典中国語 (漢文) の依存文法解析と直接構成素解析, 漢字文献情報処理研究, 第 18 号 (2018 年 10 月), pp.56-62.

[17] 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).

[18] 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.

[19] 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 (2019 年 3 月).

[20] 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.

[21] 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.

[22] 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 (2016 年 3 月).

[23] Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov: CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Proceedings of the CoNLL 2018 Shared Task (October 2018), pp.1-21.

[24] 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.

[25] Timothy Dozat, Peng Qi, Christopher D. Manning: Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task, Proceedings of the CoNLL 2017 Shared Task (August 2017), pp.20-30.

嗚呼哀哉。有兄子曰甫、制服於斯、紀徳於斯、刻石於斯。

或曰、「豈孝童之猶子与、奚孝義之勤若此。」甫泣而対曰、「非敢当是也、亦為報也。甫昔臥病於我諸姑、姑之子又病。問女巫、巫曰、『処楹之東南隅者吉。』姑遂易子之地以安我。我用是存、而姑之子卒。後乃知之於走使。甫嘗有説於人、客将出涕、感者久之、相与定謚曰義。」

君子以為魯義姑者、遇暴客於郊、抱其所携、棄其所抱、以割私愛。県君有焉。

是以挙茲一隅、昭彼百行。銘而不韻、蓋情至無文。其詞曰「嗚呼、有唐義姑、京兆杜氏之墓。」

図 3: 大学入試センター試験『国語』(2019 年 1 月 19 日) 第 4 問本文

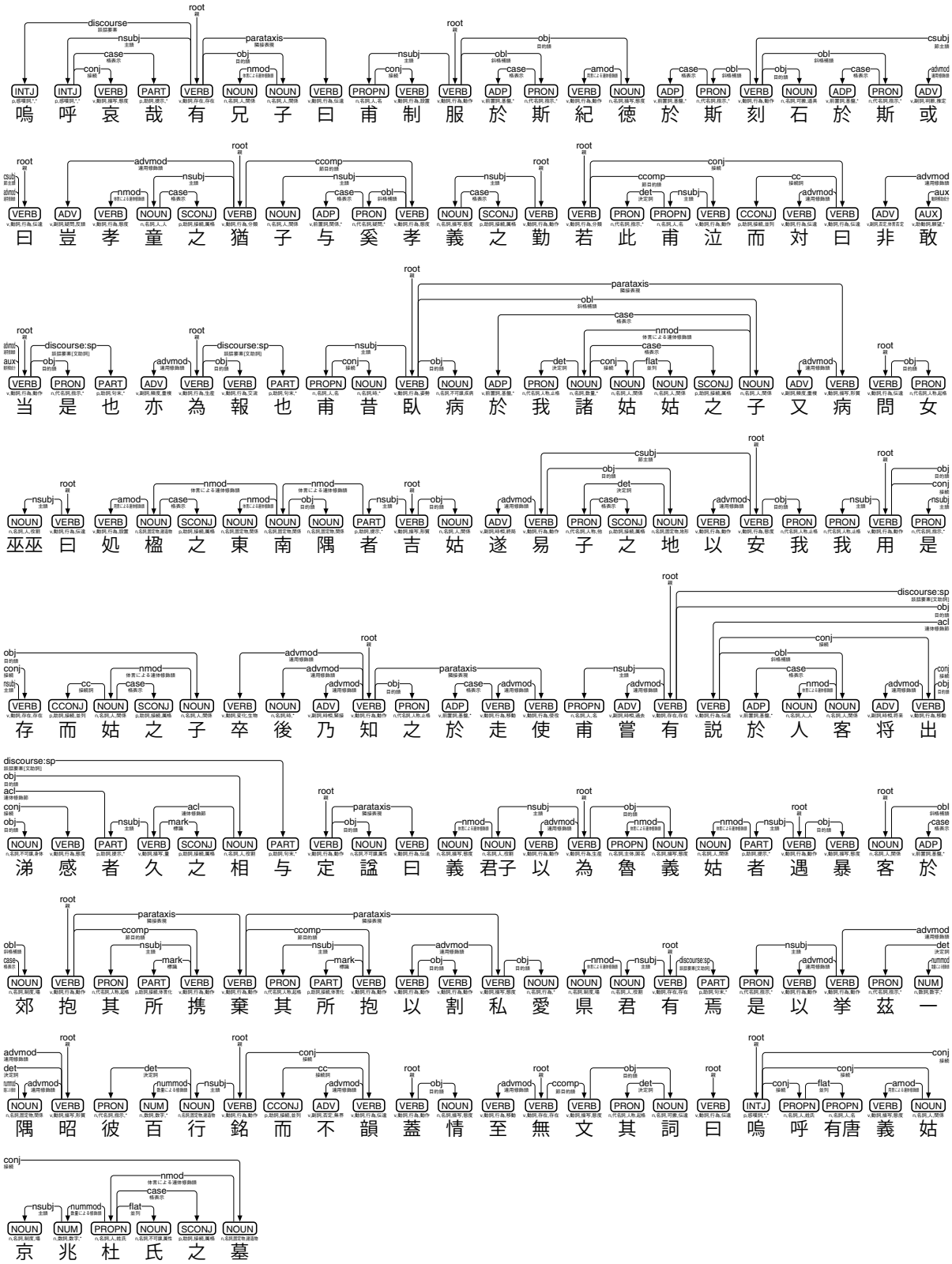


図 4: 手法①の処理結果 (2019 年)

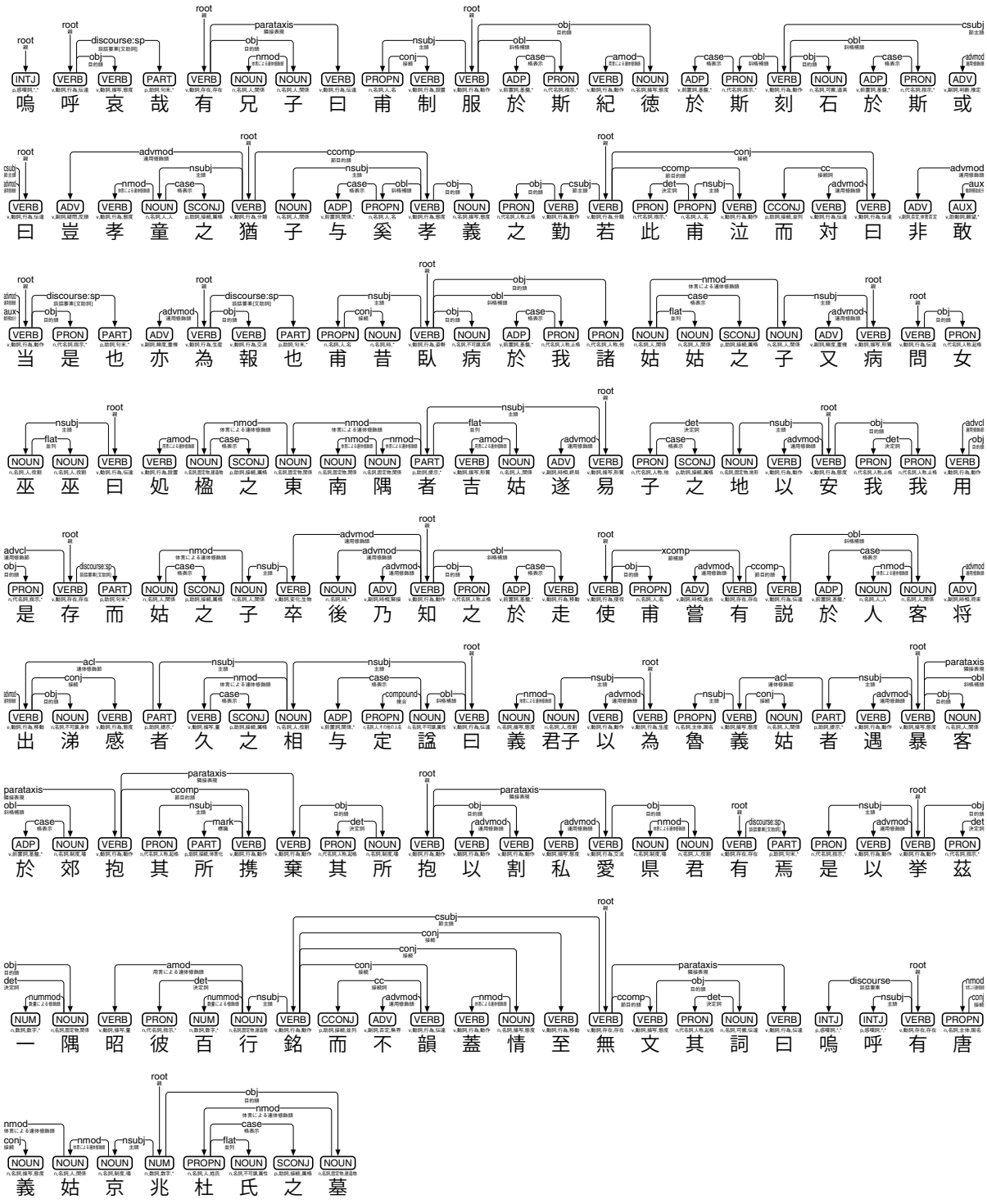


図 5: 手法②の処理結果 (2019 年)

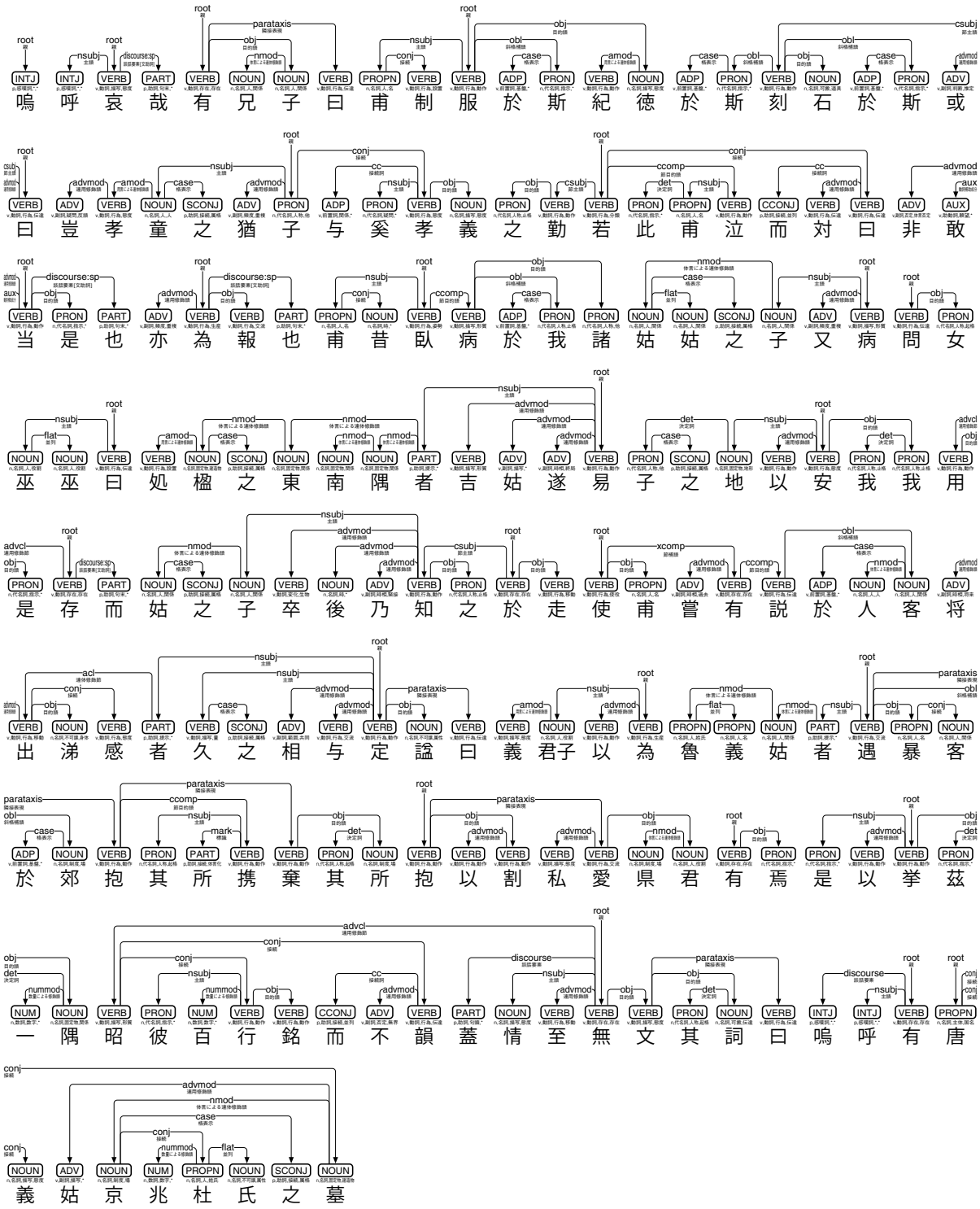


図 6: 手法③の処理結果 (2019 年)

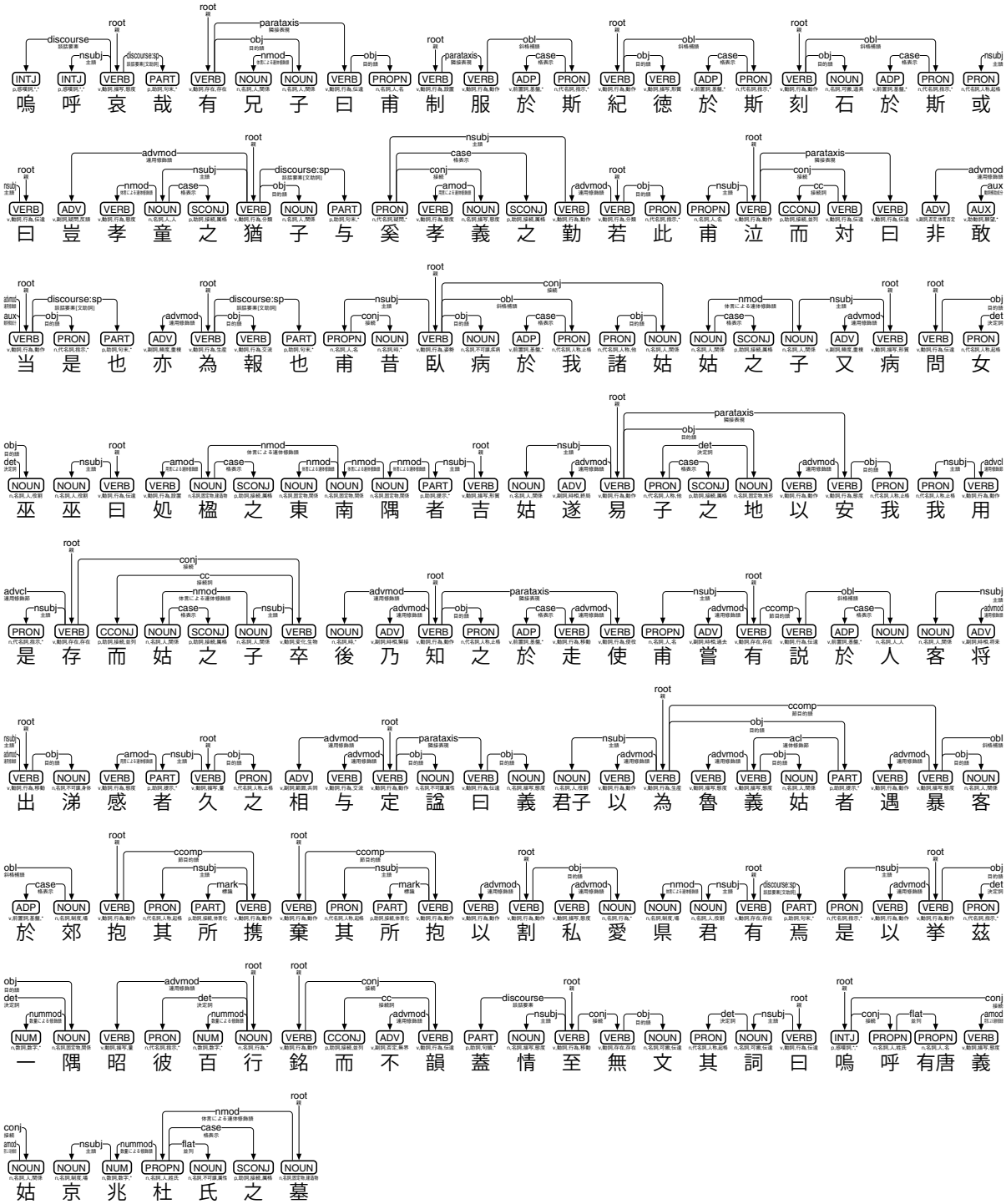


図 7: 手法1の処理結果 (2019 年)

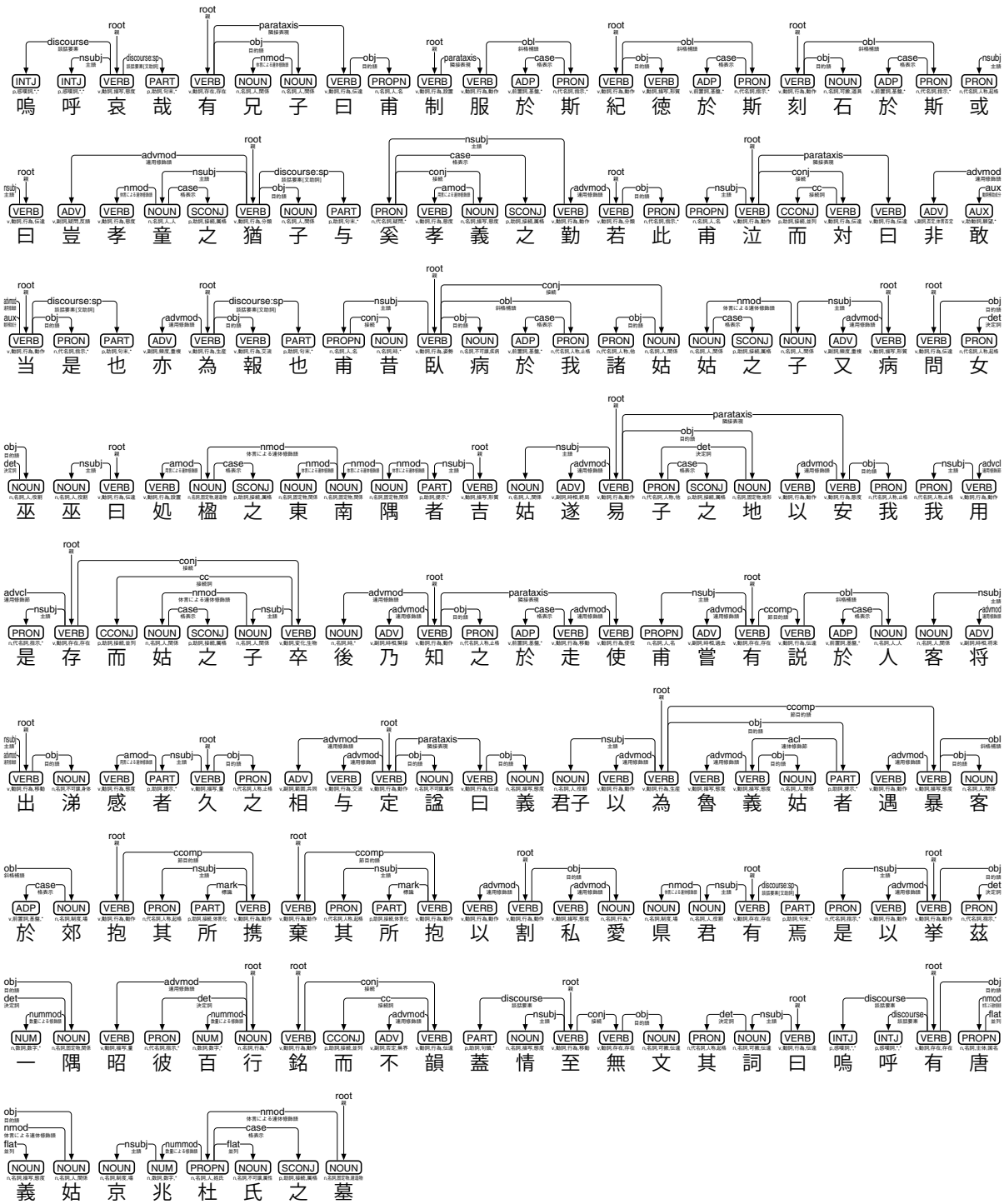


図 8: 手法2の処理結果 (2019 年)

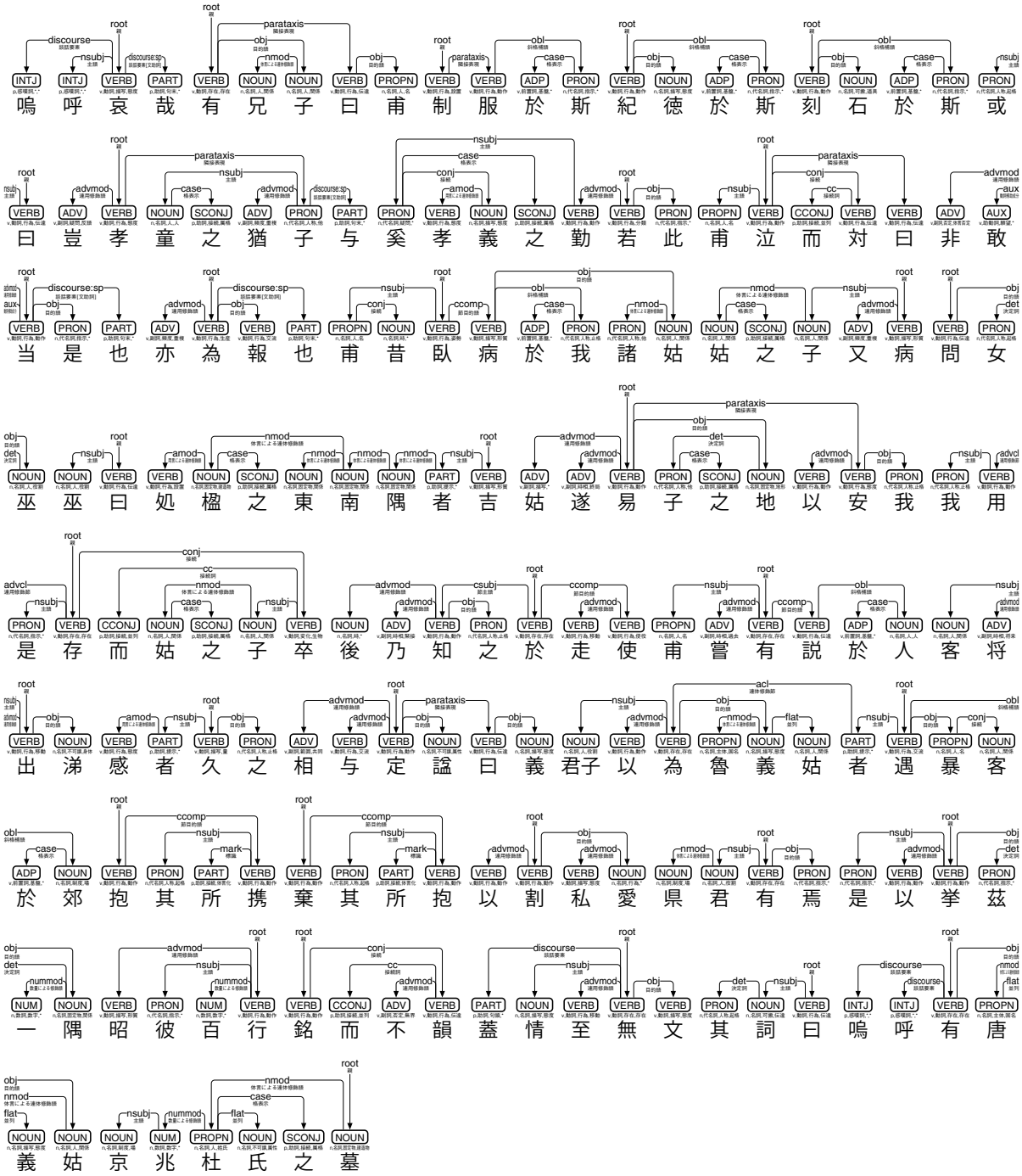


図 9: 手法3の処理結果 (2019 年)

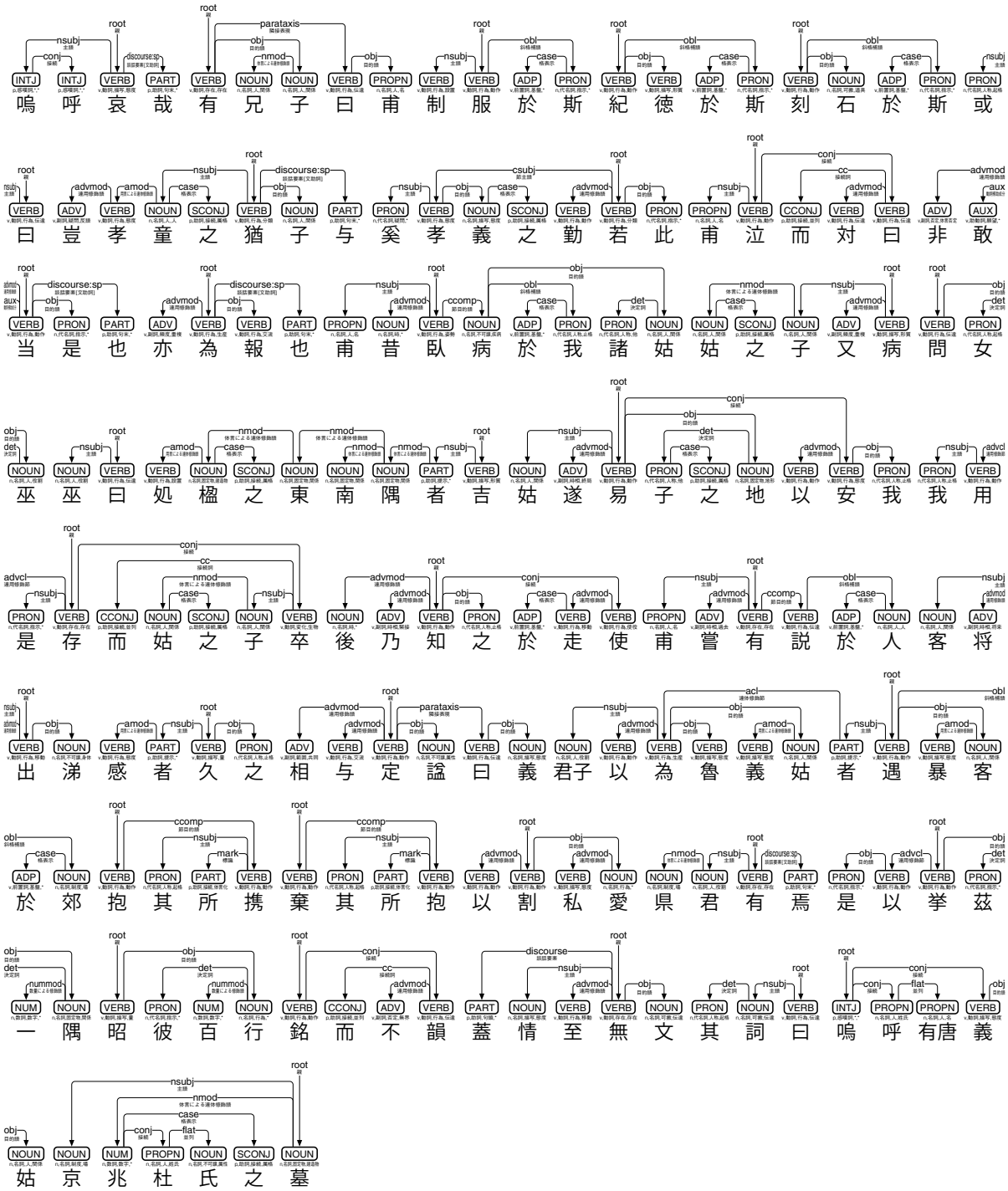


図 10: 手法❶の処理結果 (2019 年)

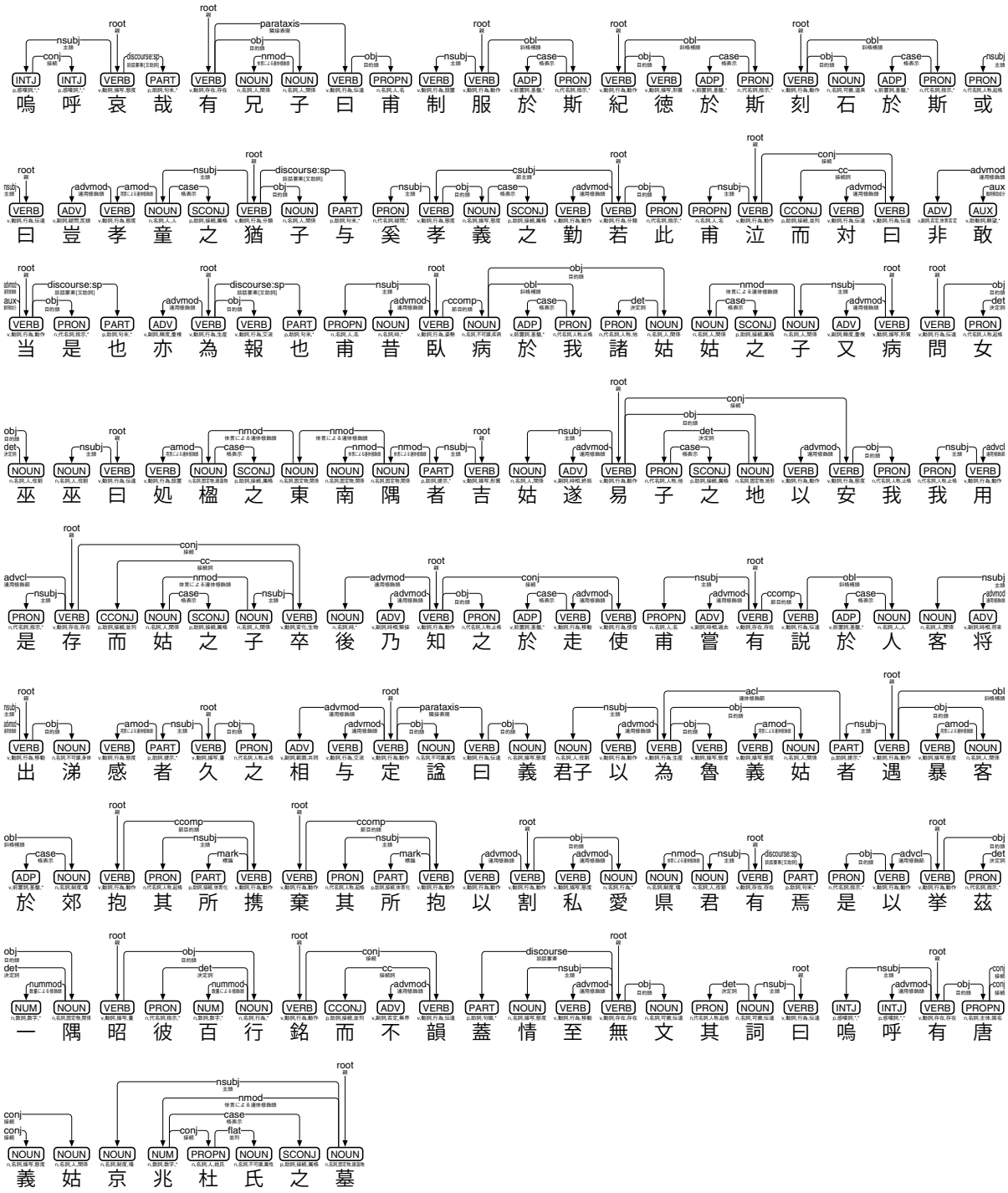


図 11: 手法❷の処理結果 (2019 年)

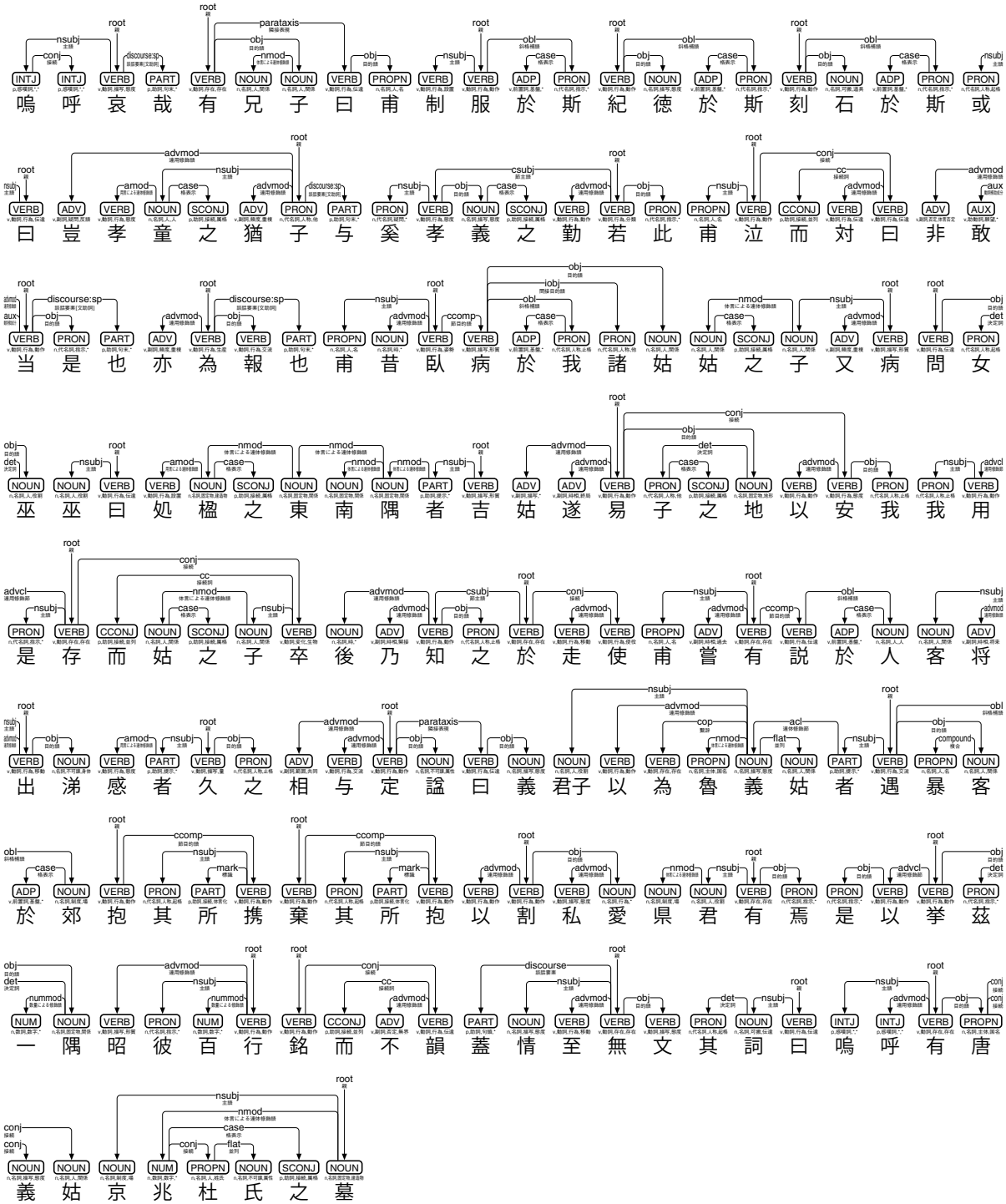


図 12: 手法❸の処理結果 (2019 年)

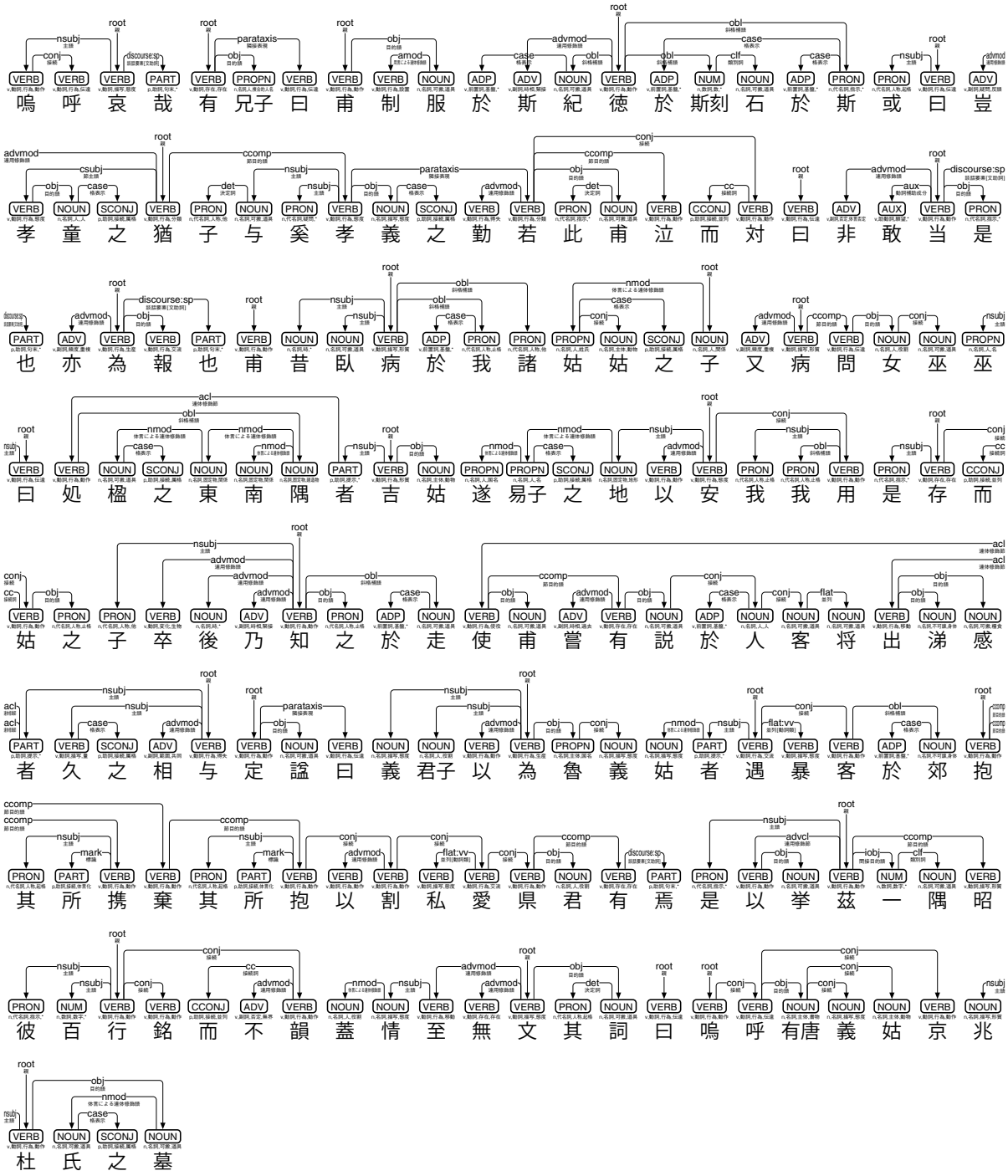


図 13: 手法Ⓐの処理結果 (2019 年)

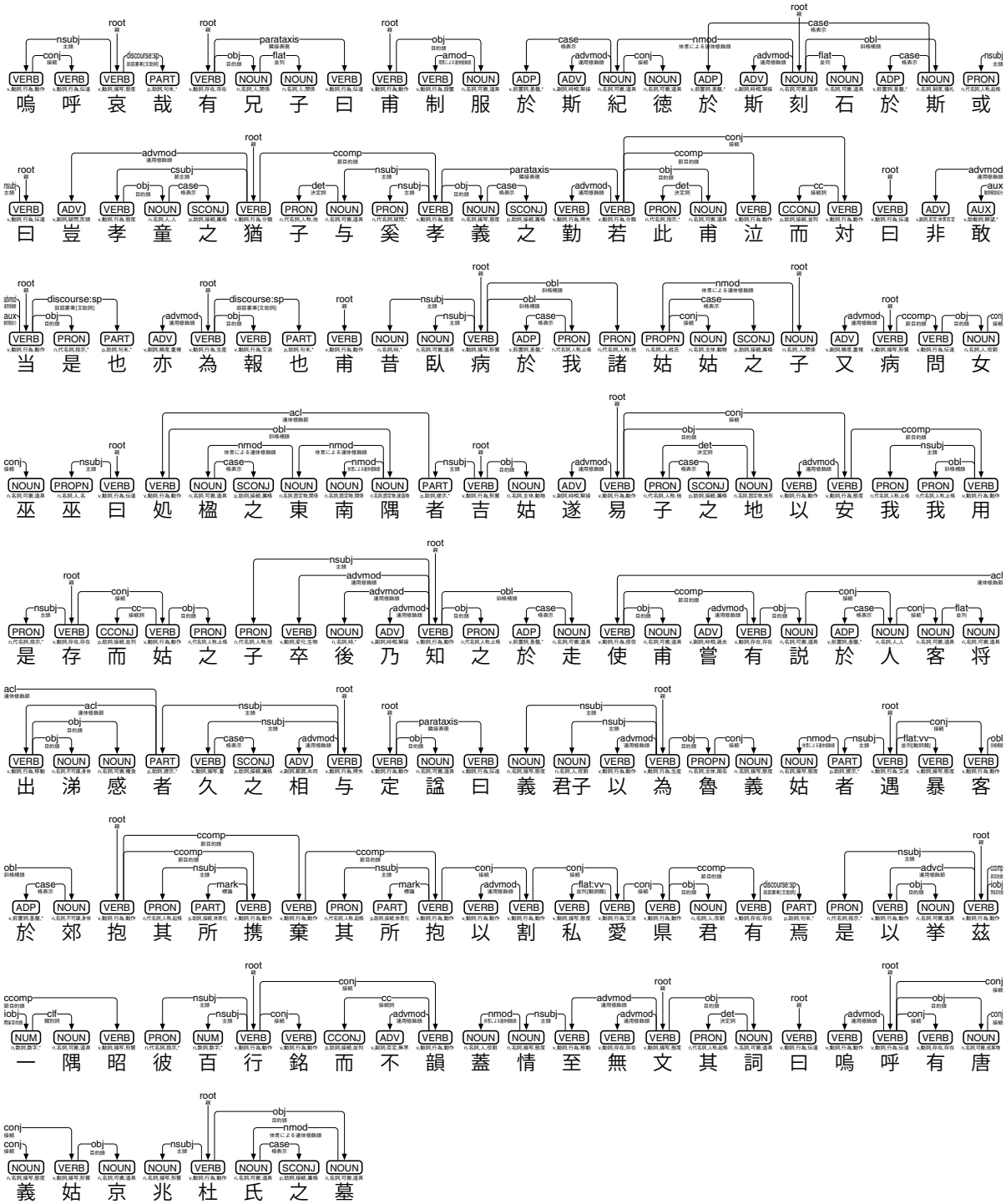


図 14: 手法Ⓑの処理結果 (2019 年)

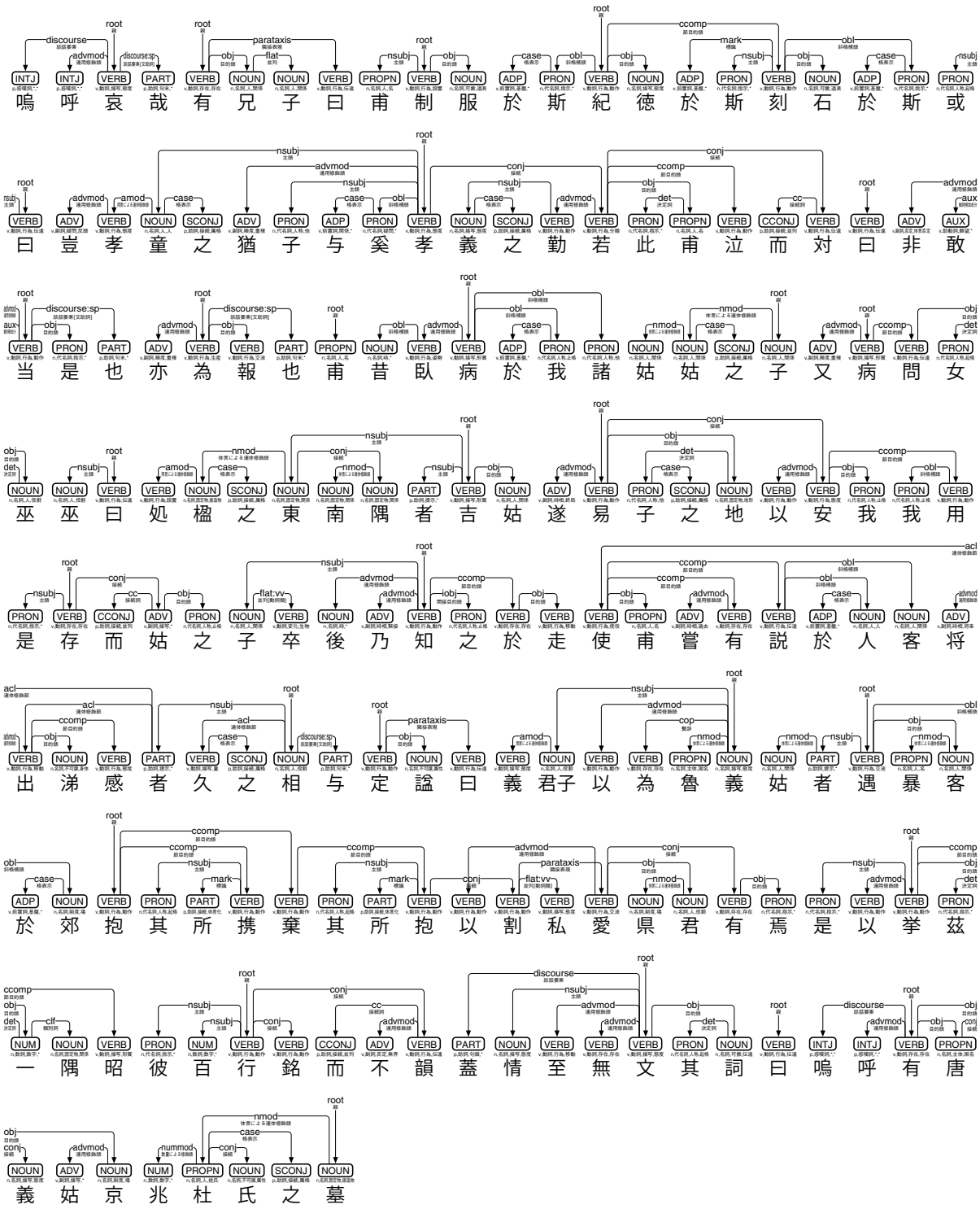


図 15: 手法Ⓒの処理結果 (2019 年)

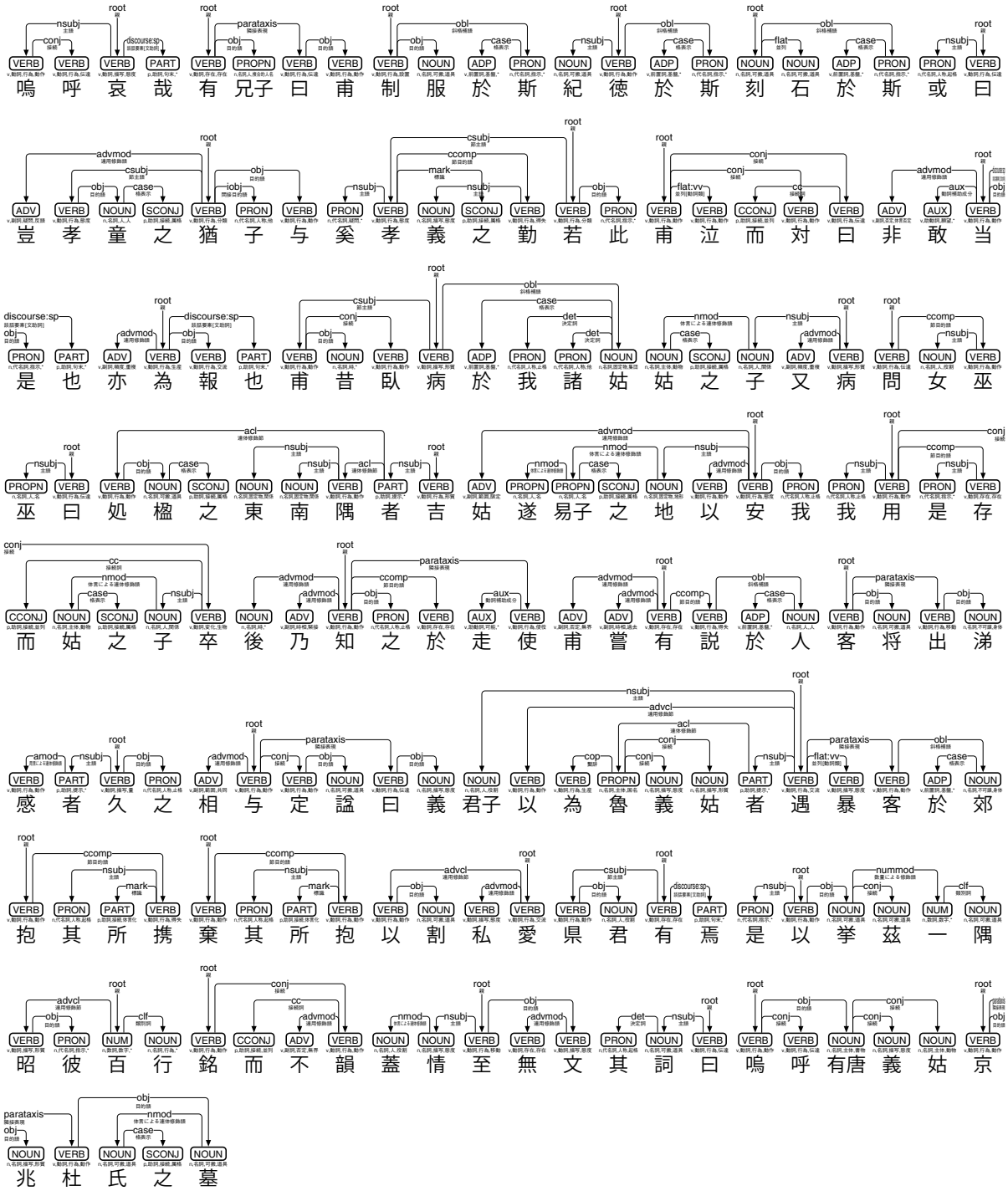


図 16: 手法Aの処理結果 (2019 年)

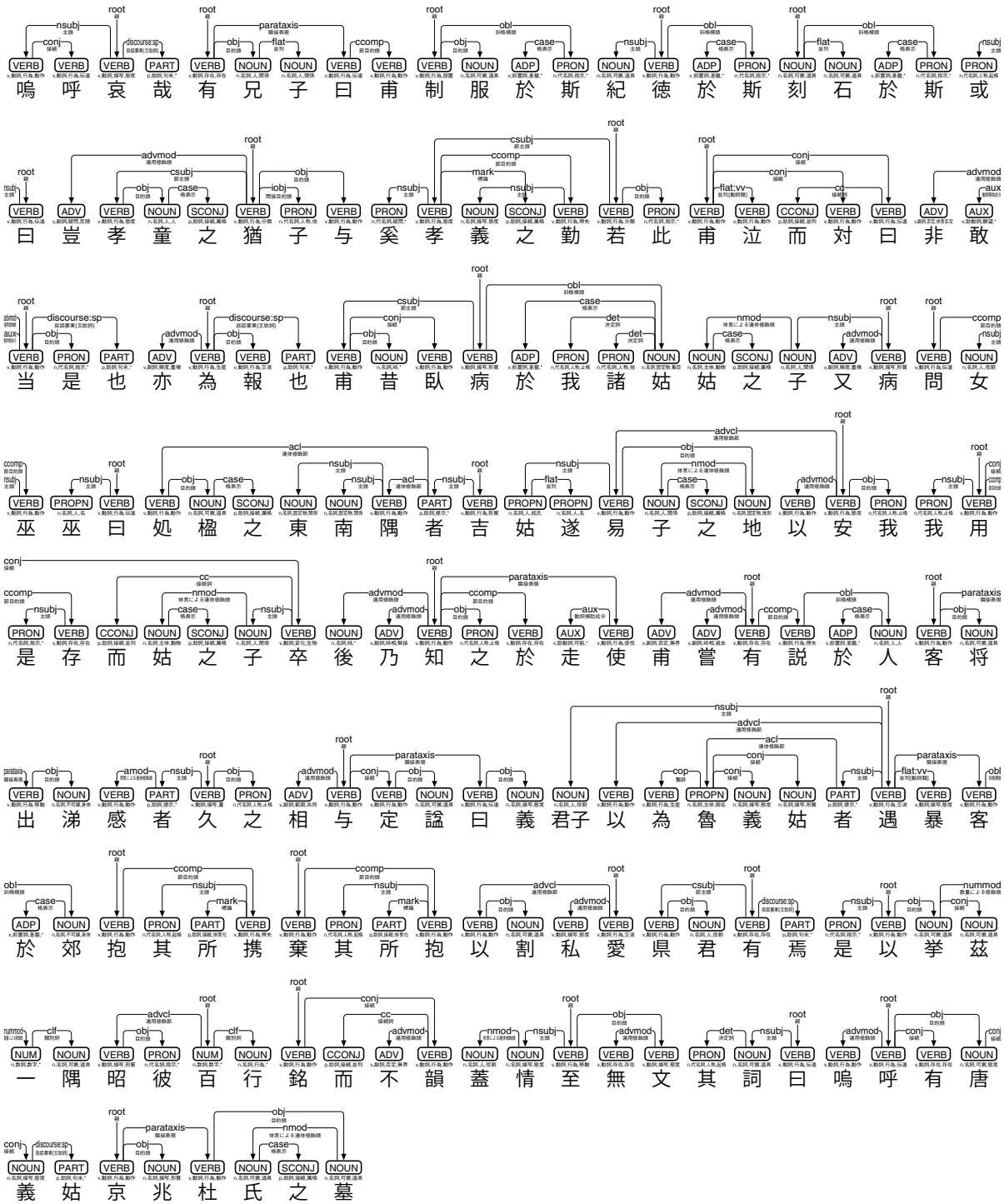


図 17: 手法Bの処理結果 (2019 年)

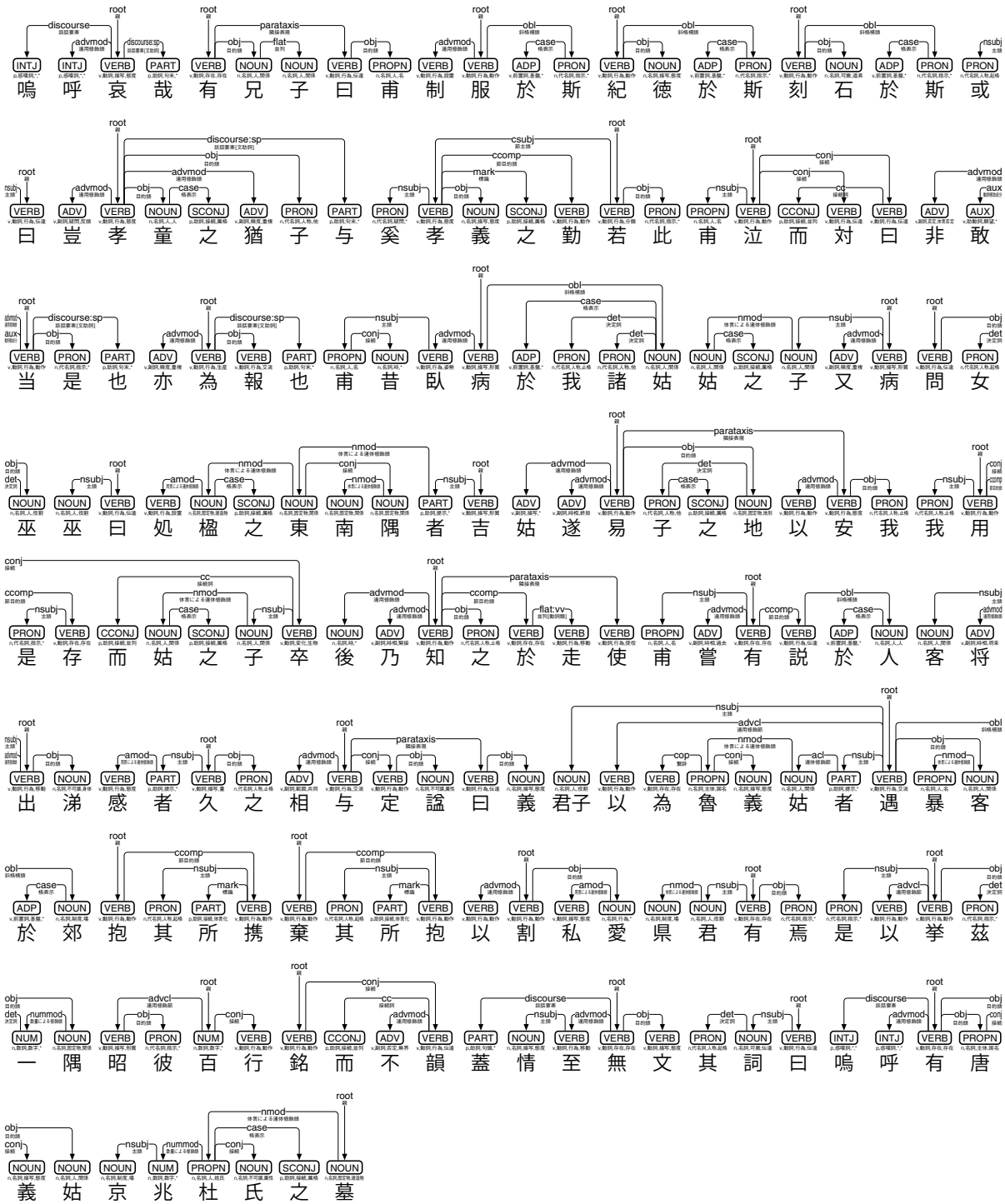


図 18: 手法Ⓒの処理結果 (2019 年)

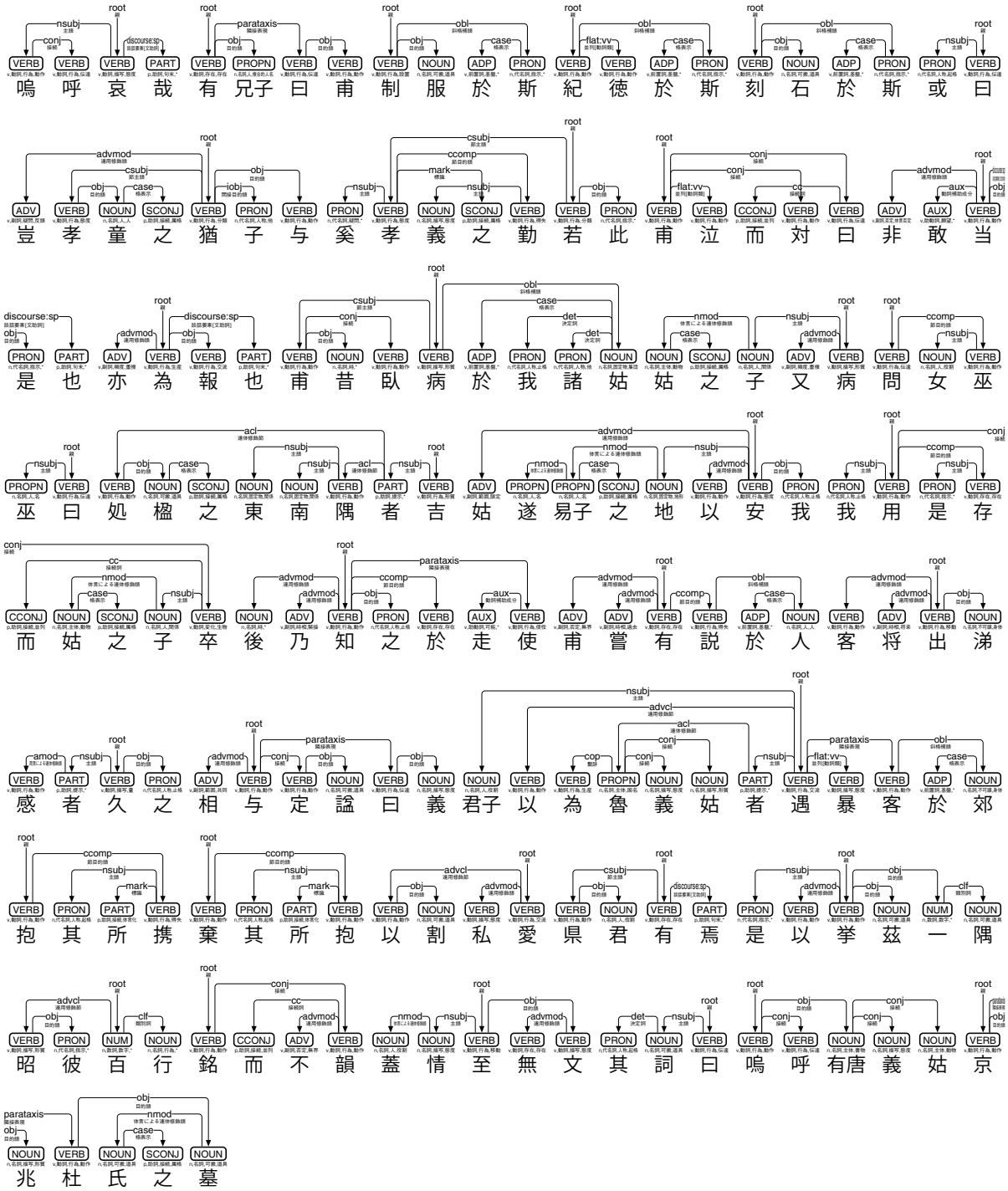


図 19: 手法Ⓐの処理結果 (2019 年)

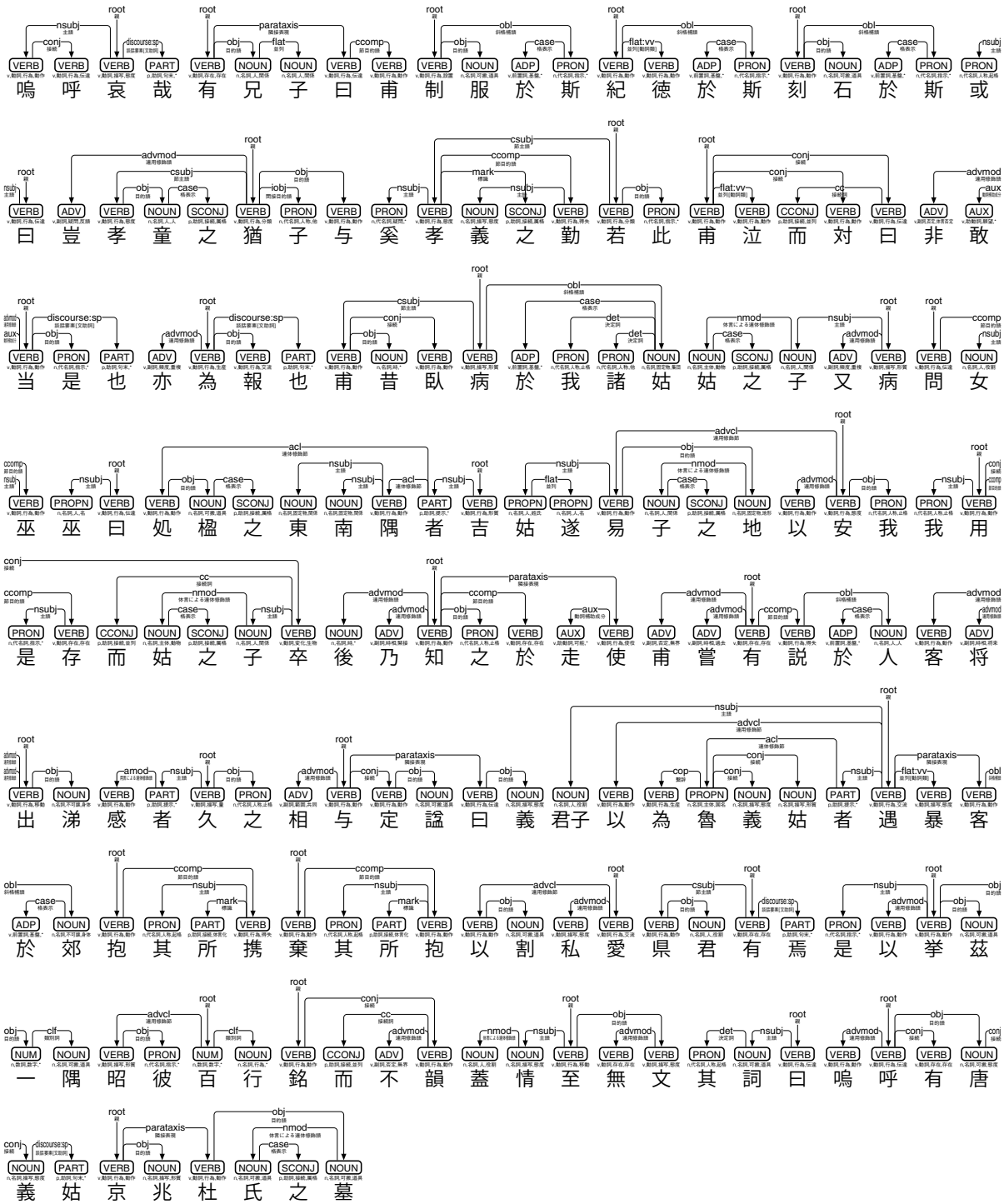


図 20: 手法Ⓑの処理結果 (2019 年)

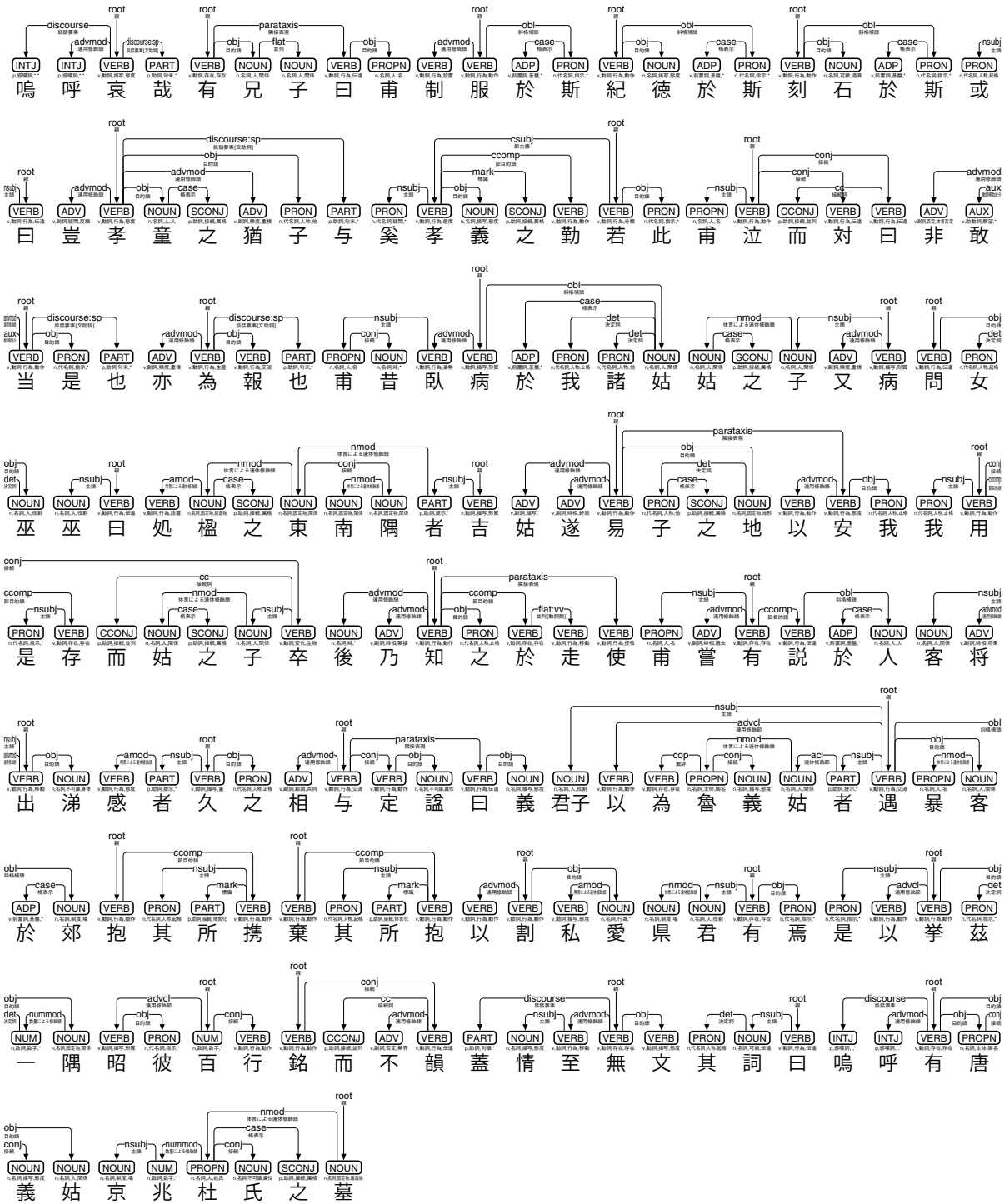


図 21: 手法Ⓒの処理結果 (2019 年)

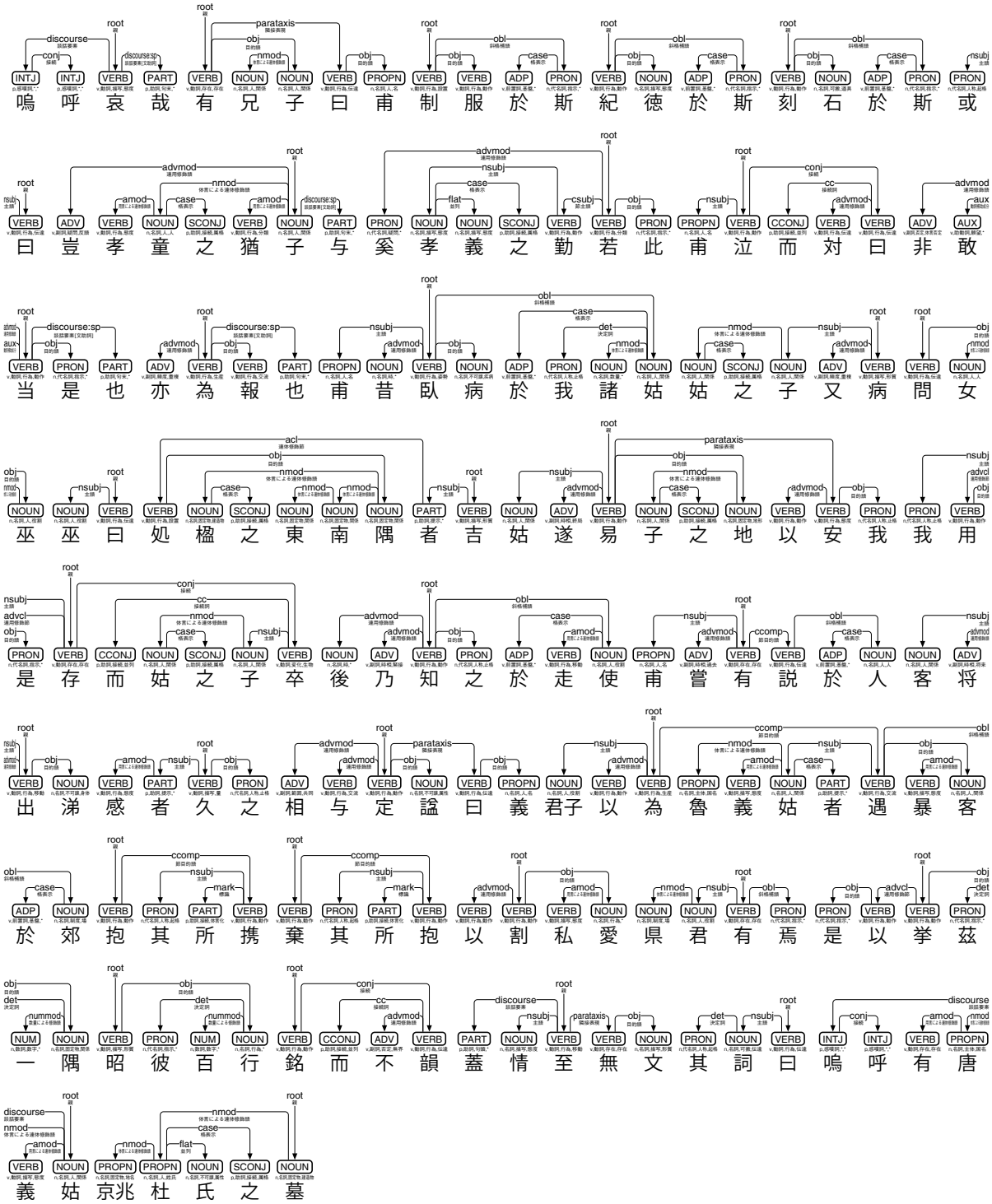


図 22: 手作業で作成した「正解」UD (2019 年)

嘉祐、禹偁子也。嘉祐平時若愚騃、独寇準知之。準知開

封府、一日、問嘉祐曰、「外間議準云何。」嘉祐曰、「外人皆云丈人旦夕入相。」準曰、「於吾子意何如。」嘉祐曰、「以愚観之、丈人不若未為相。為相則誉望損矣。」準曰、「何故。」嘉祐曰、「自古賢相所以能建功業沢生民者、其君臣相得皆如魚之有水。故言聴計従、而功名倶美。今丈人負天下重望、相則中外以太平責焉。丈人之于明主、能若魚之有水乎。嘉祐所以恐誉望之損也。」準喜、起執其手曰、「元之雖文章冠天下、至於深識遠慮、殆不能勝吾子也。」

図 23: 大学入試センター試験『国語』(2018 年 1 月 13 日) 第 4 問本文

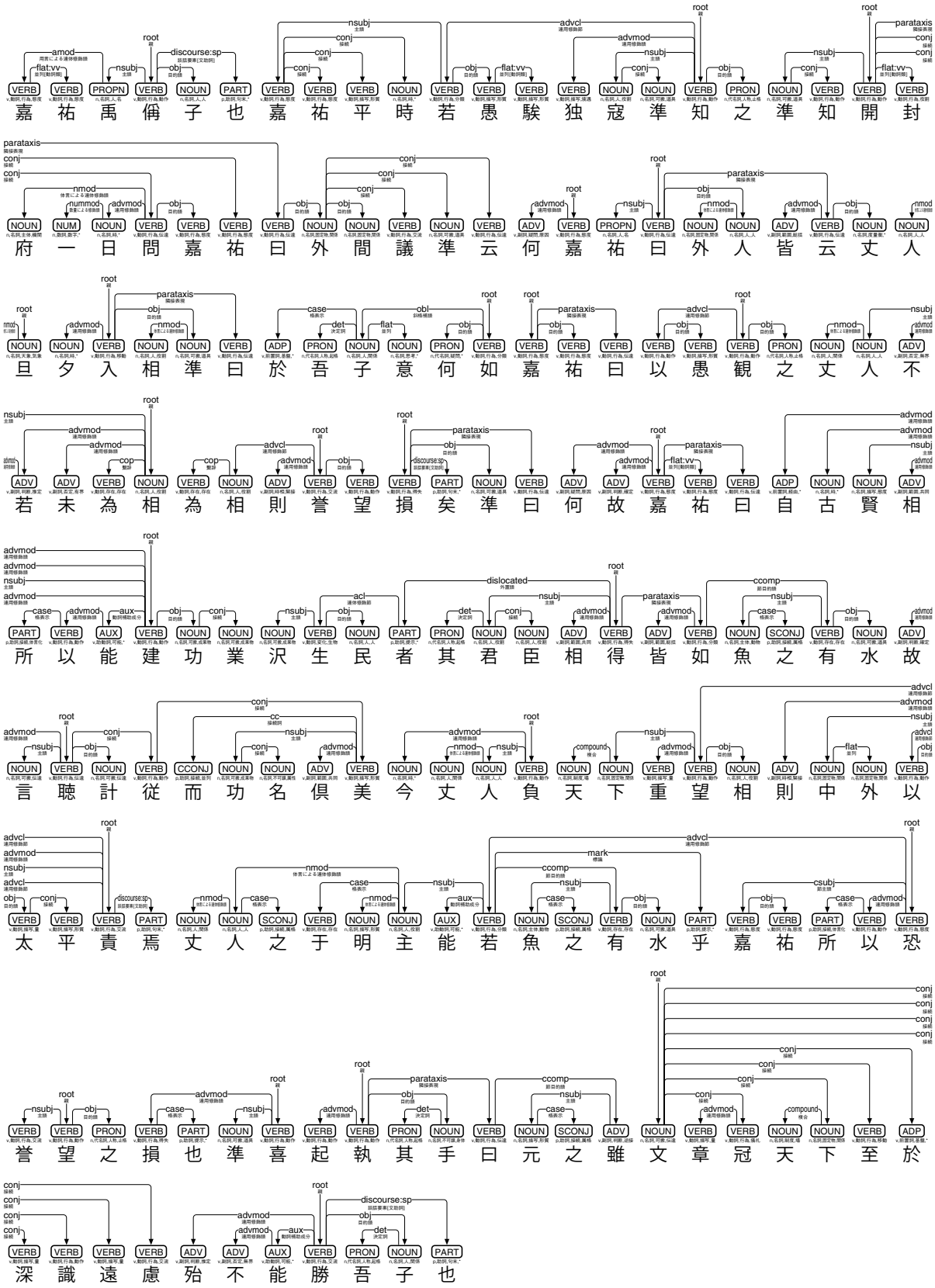


図 24: 手法①の処理結果 (2018 年)

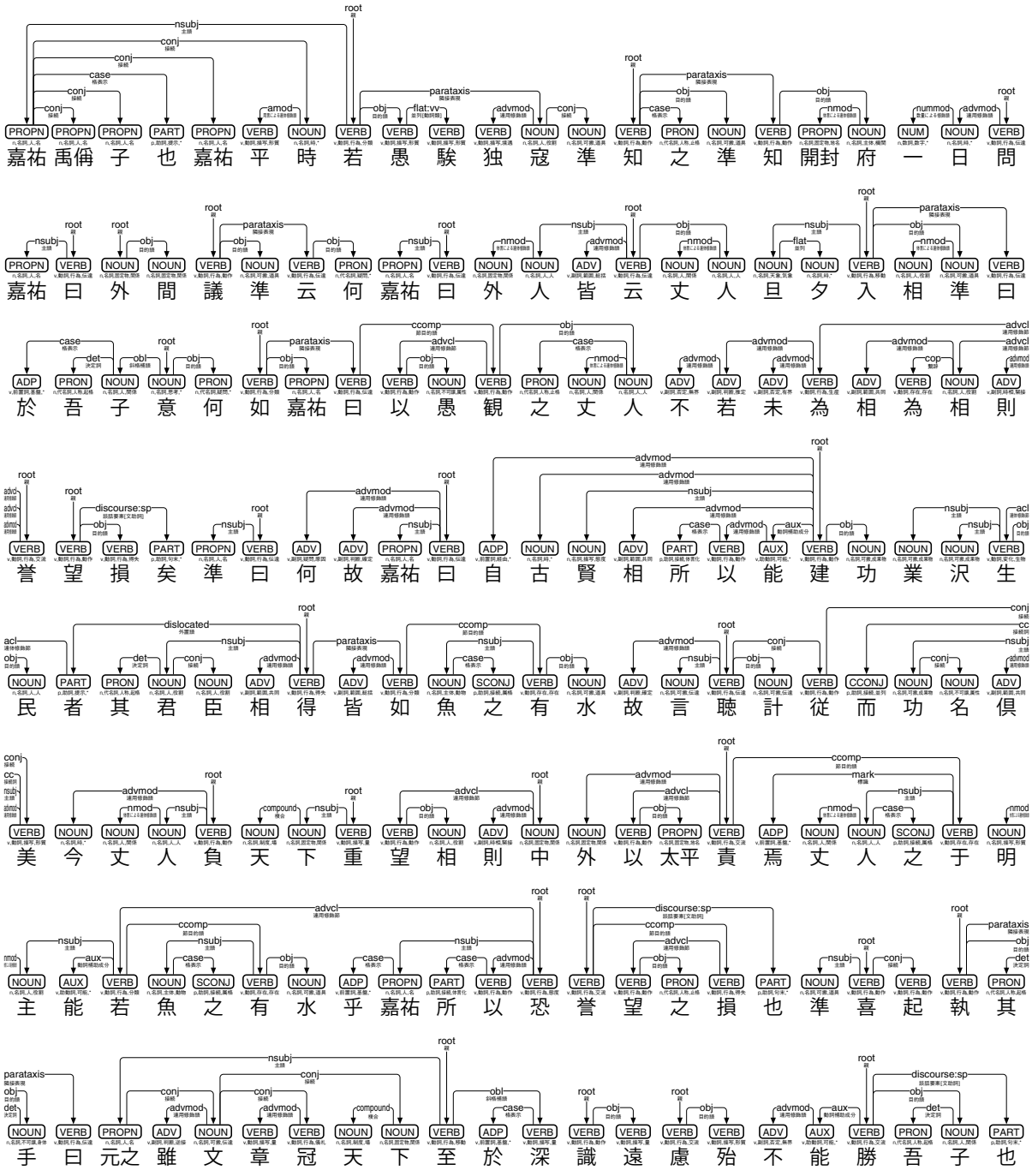


図 25: 手法②の処理結果 (2018 年)

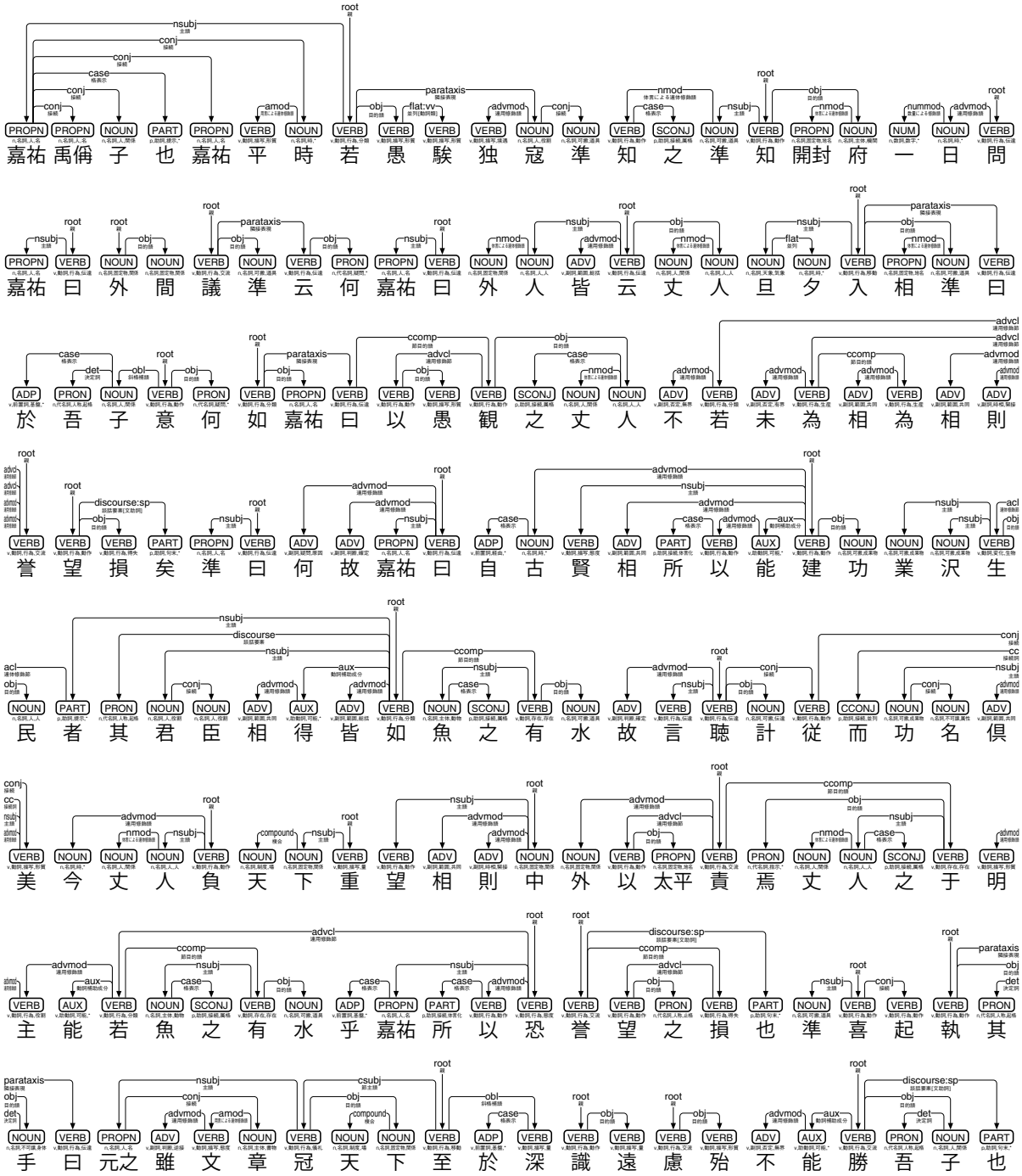


図 26: 手法③の処理結果 (2018 年)

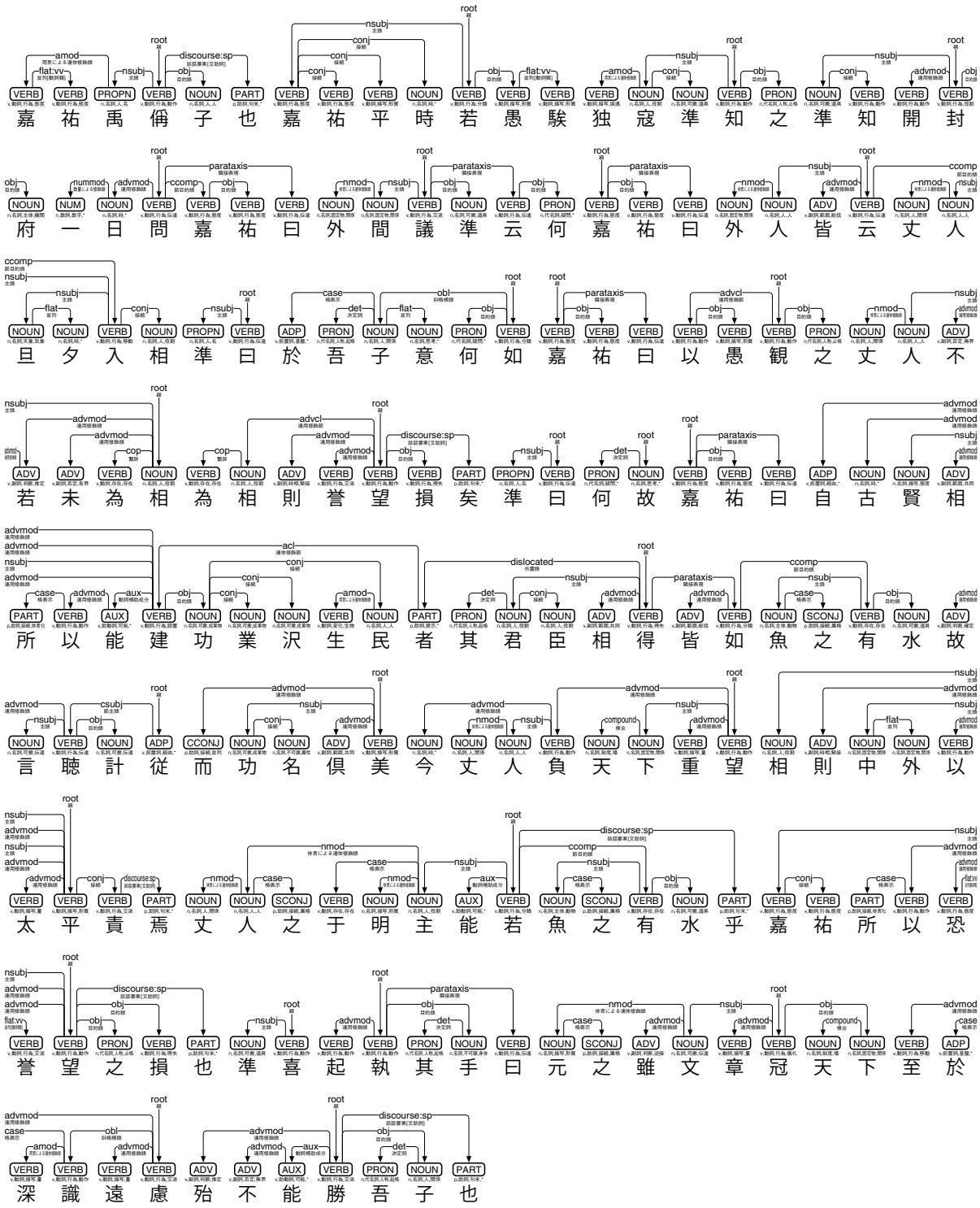


図 27: 手法1の処理結果 (2018 年)

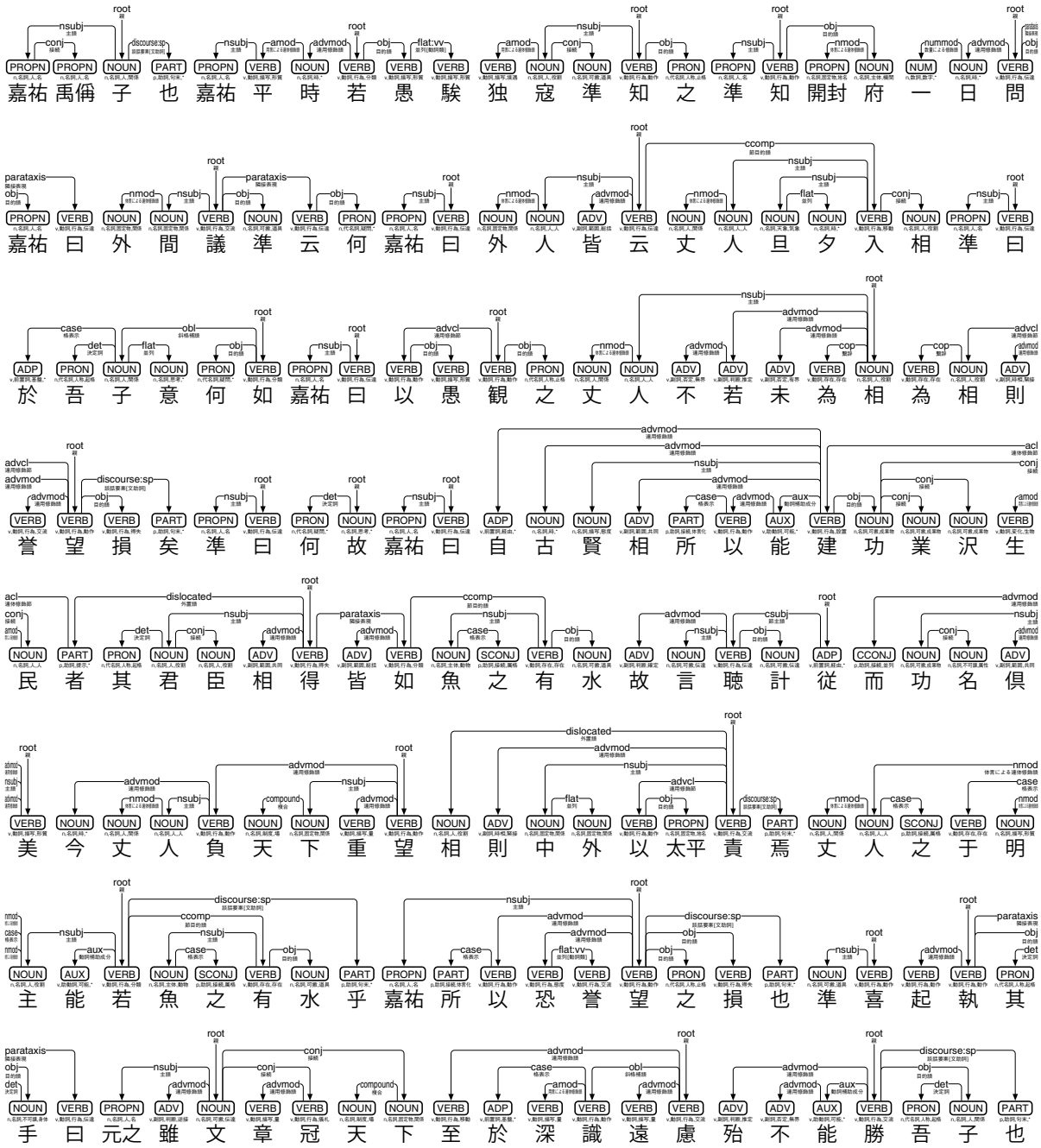


図 28: 手法2の処理結果 (2018 年)

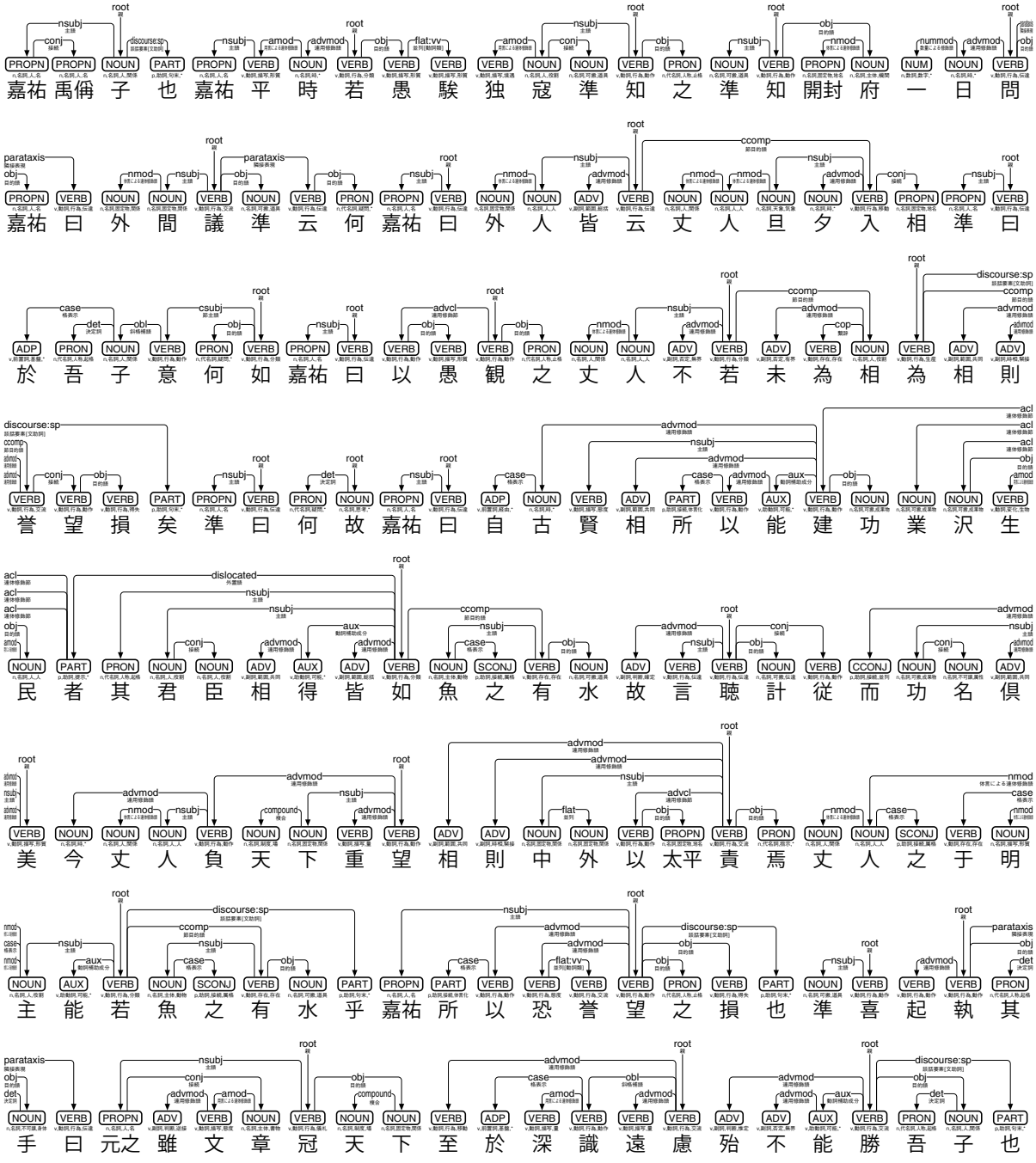


図 29: 手法3の処理結果 (2018 年)

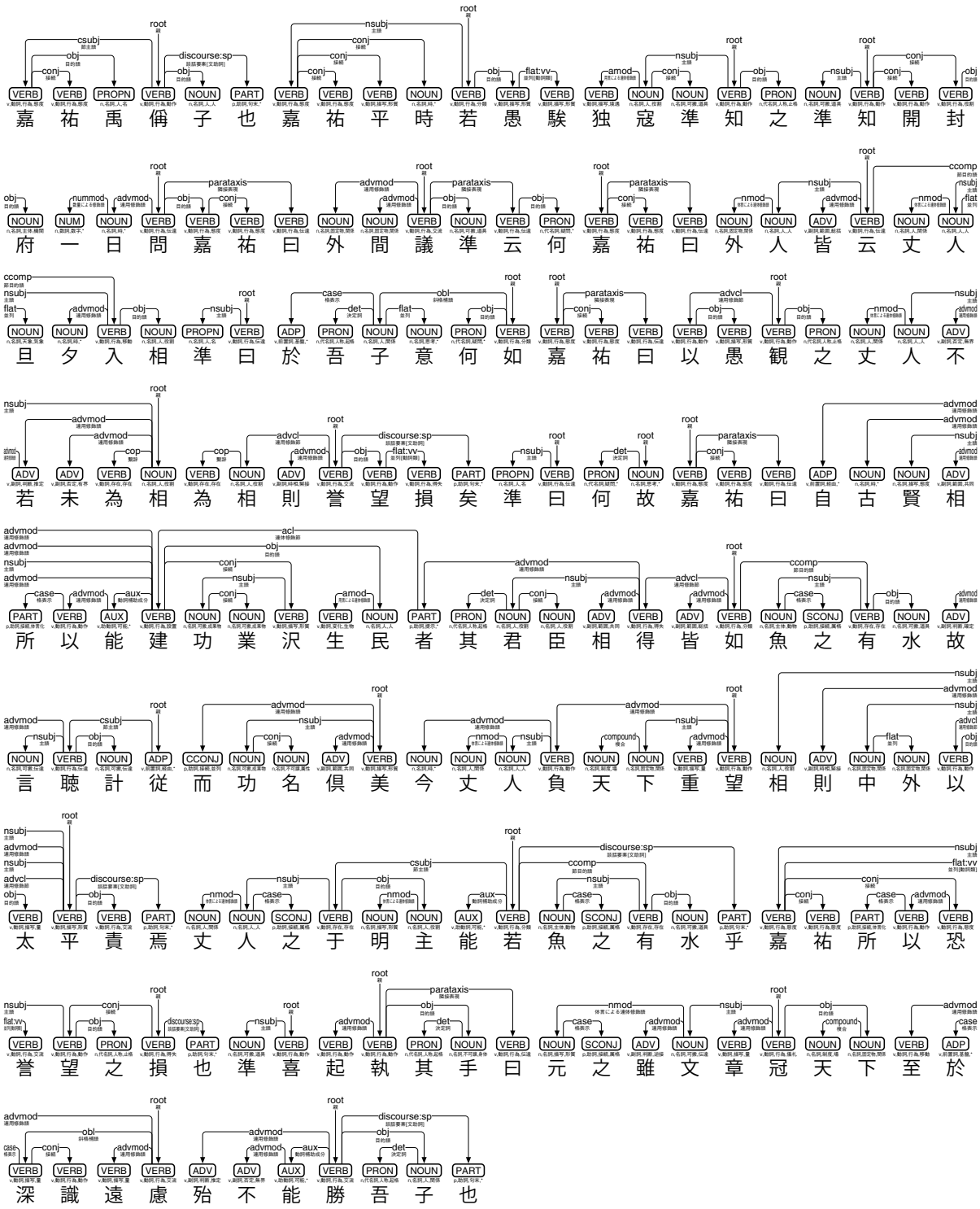


図 30: 手法❶の処理結果 (2018 年)

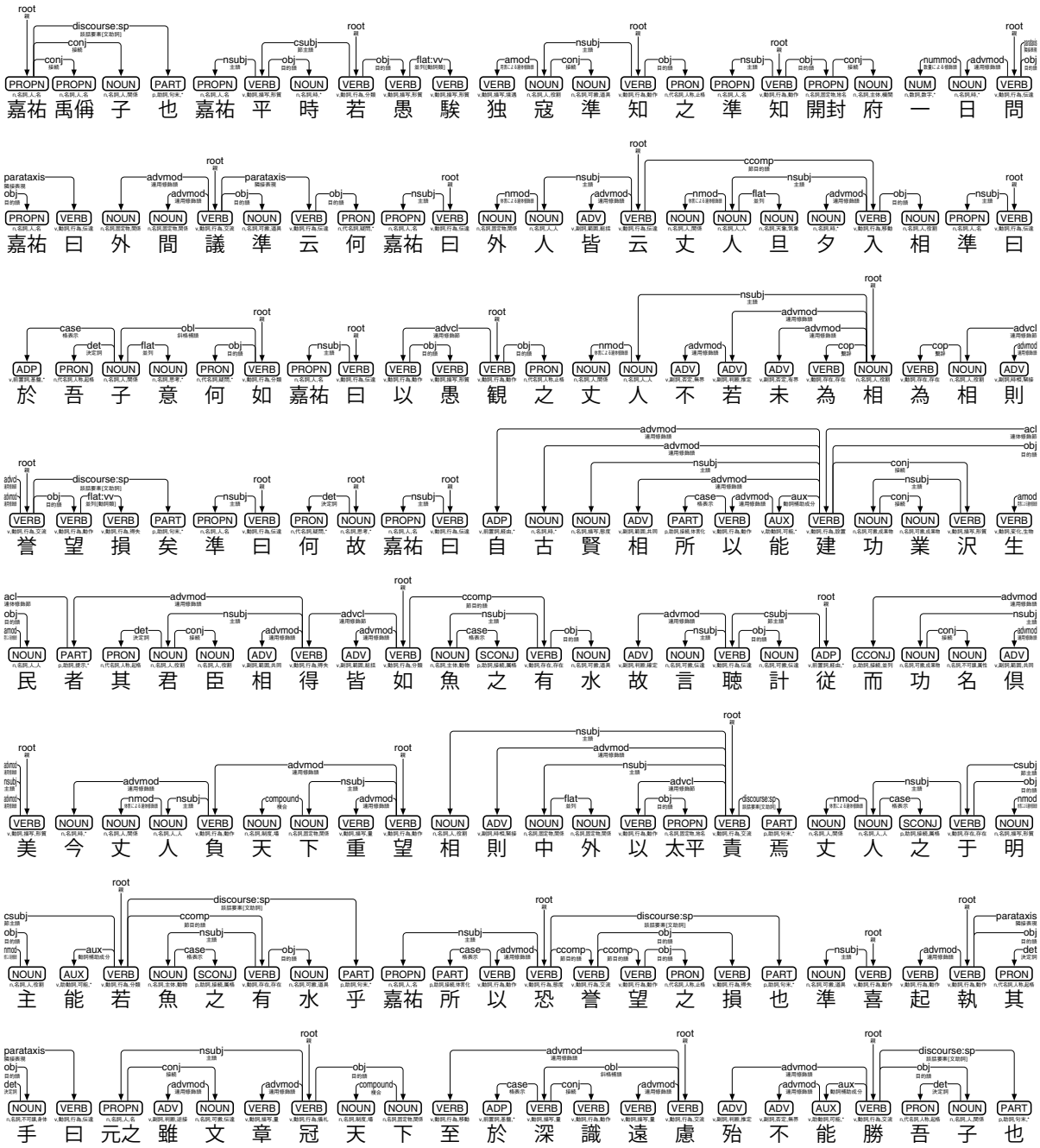


図 31: 手法❷の処理結果 (2018 年)

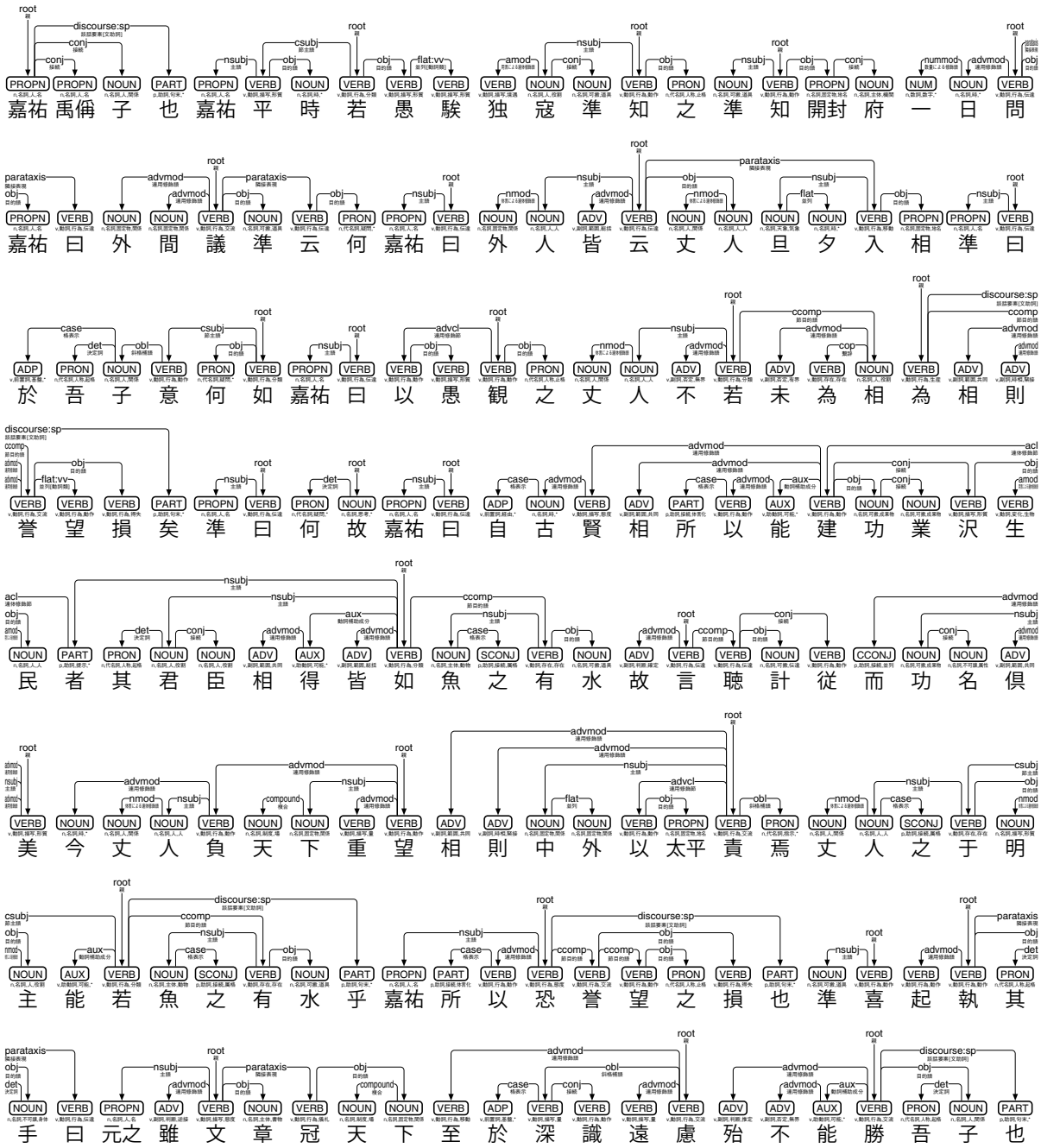


図 32: 手法❸の処理結果 (2018 年)

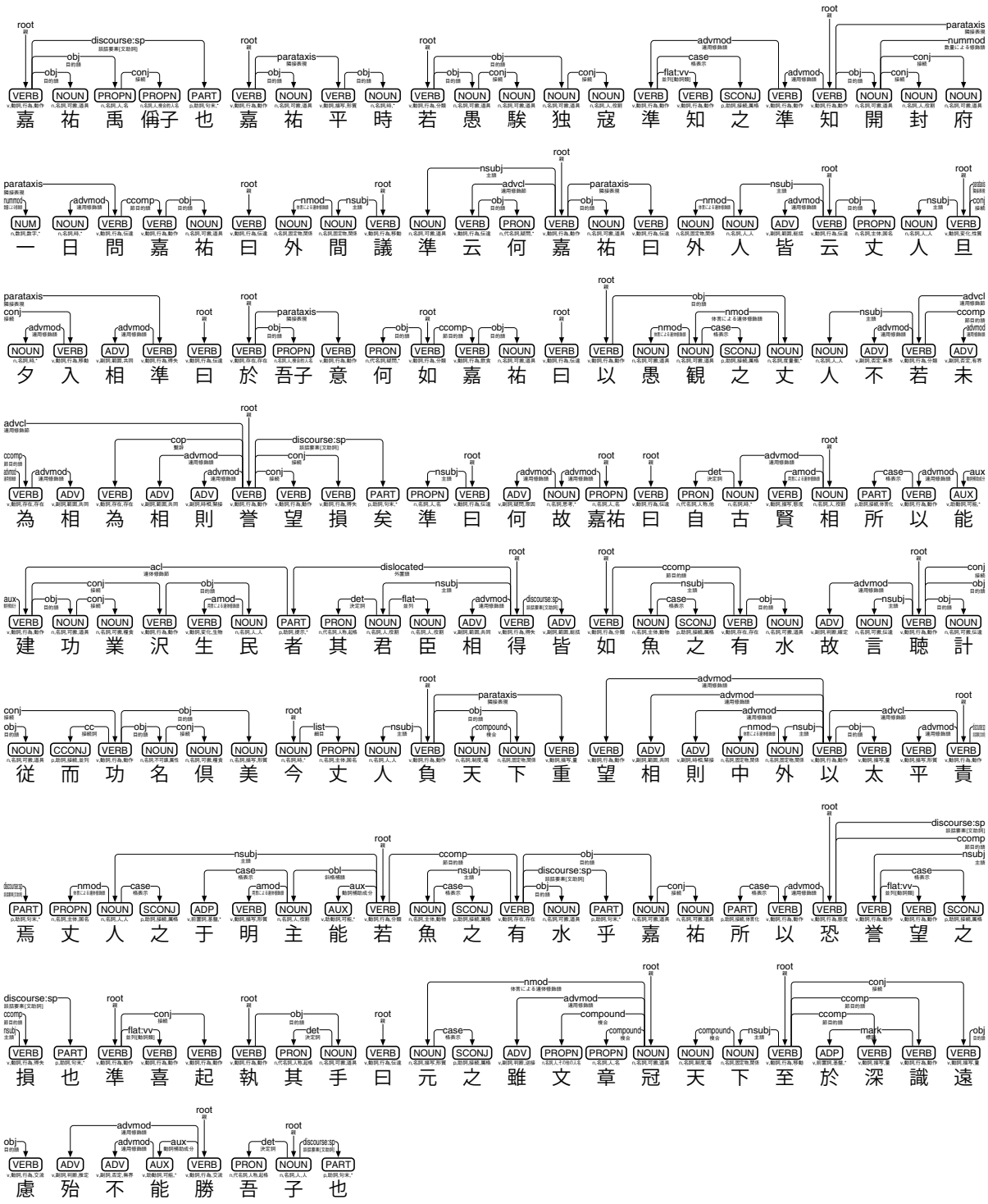


図 33: 手法Ⓐの処理結果 (2018 年)

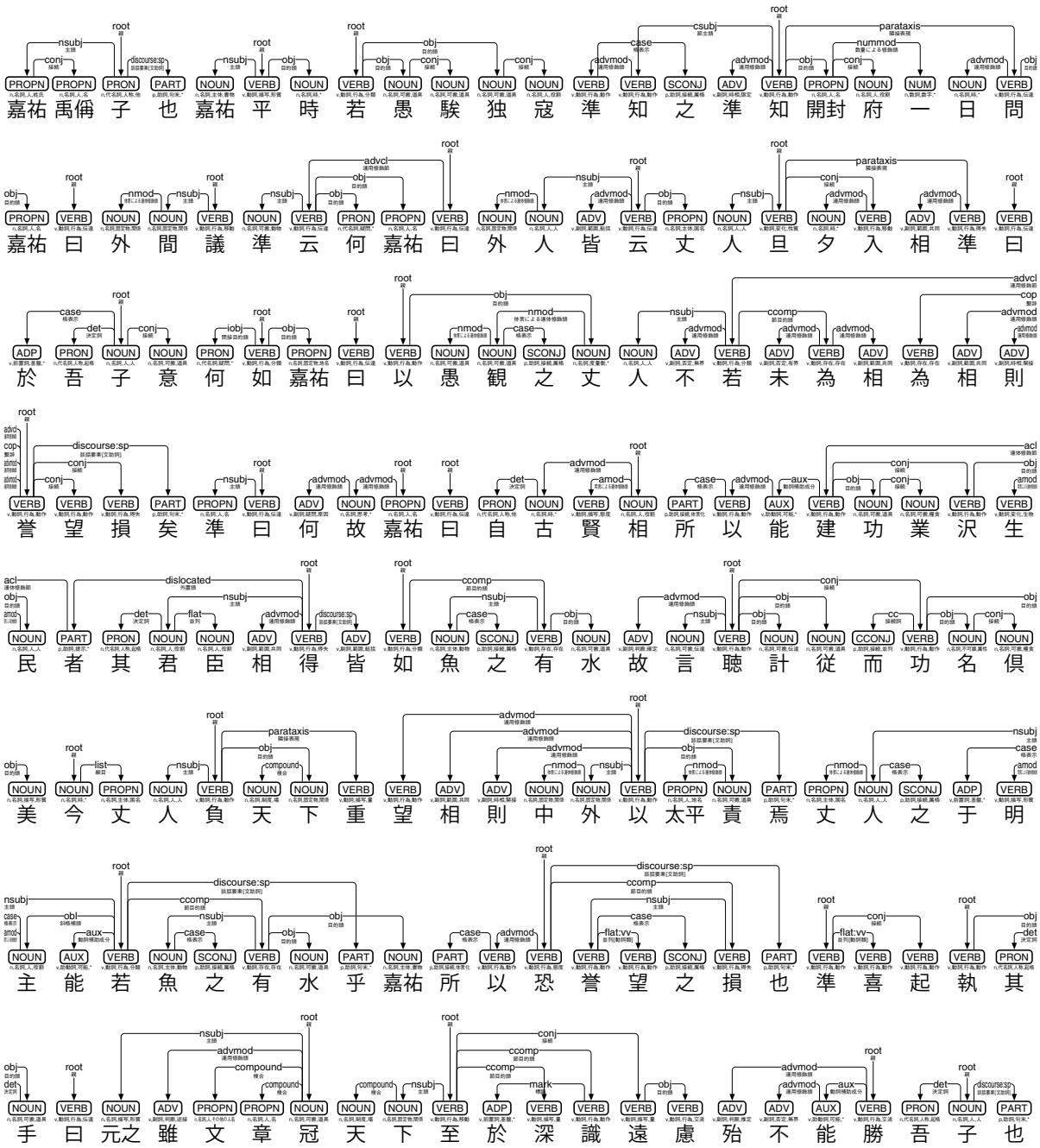


図 34: 手法Ⓑの処理結果 (2018 年)

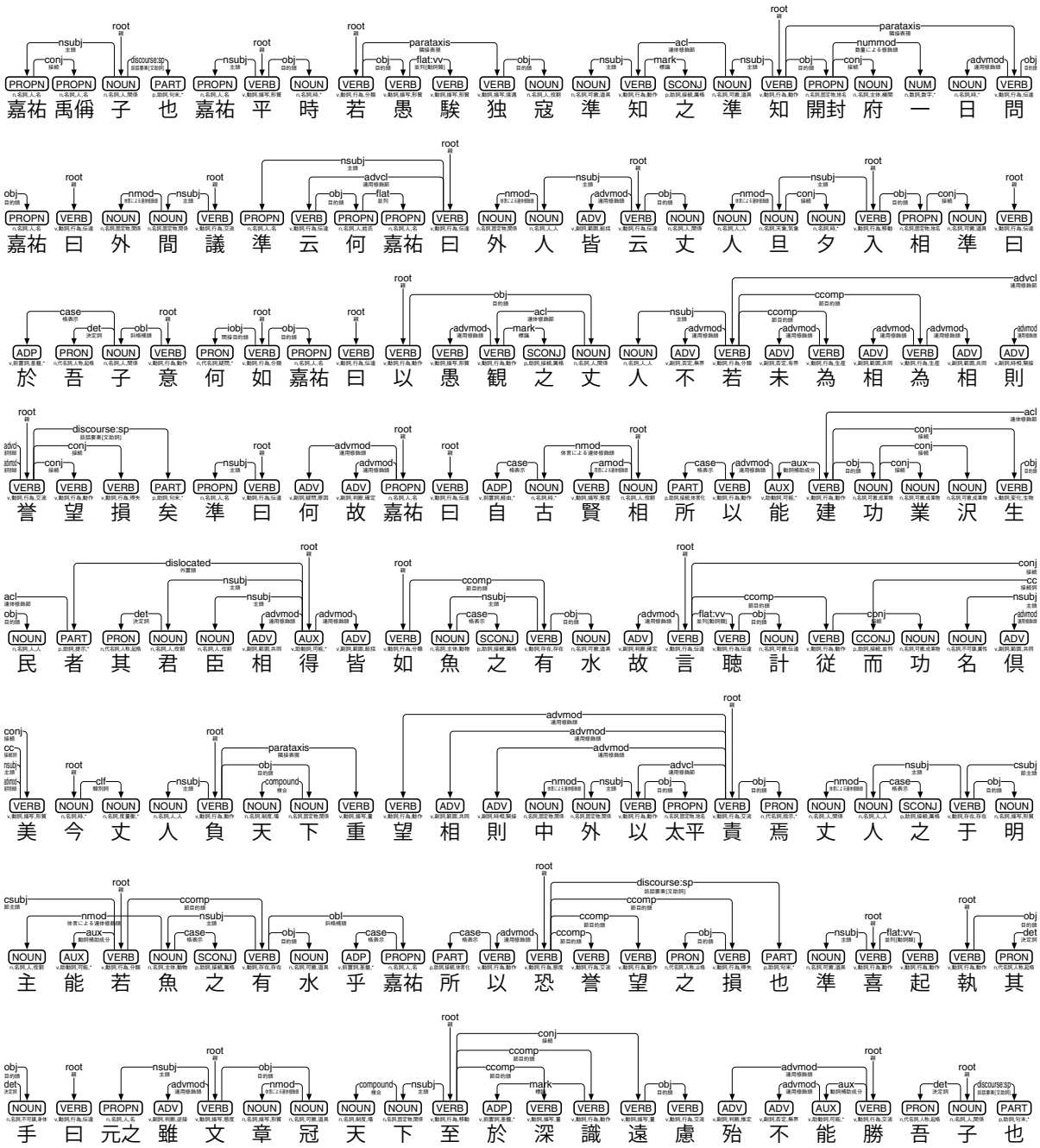


図 35: 手法Ⓒの処理結果 (2018 年)

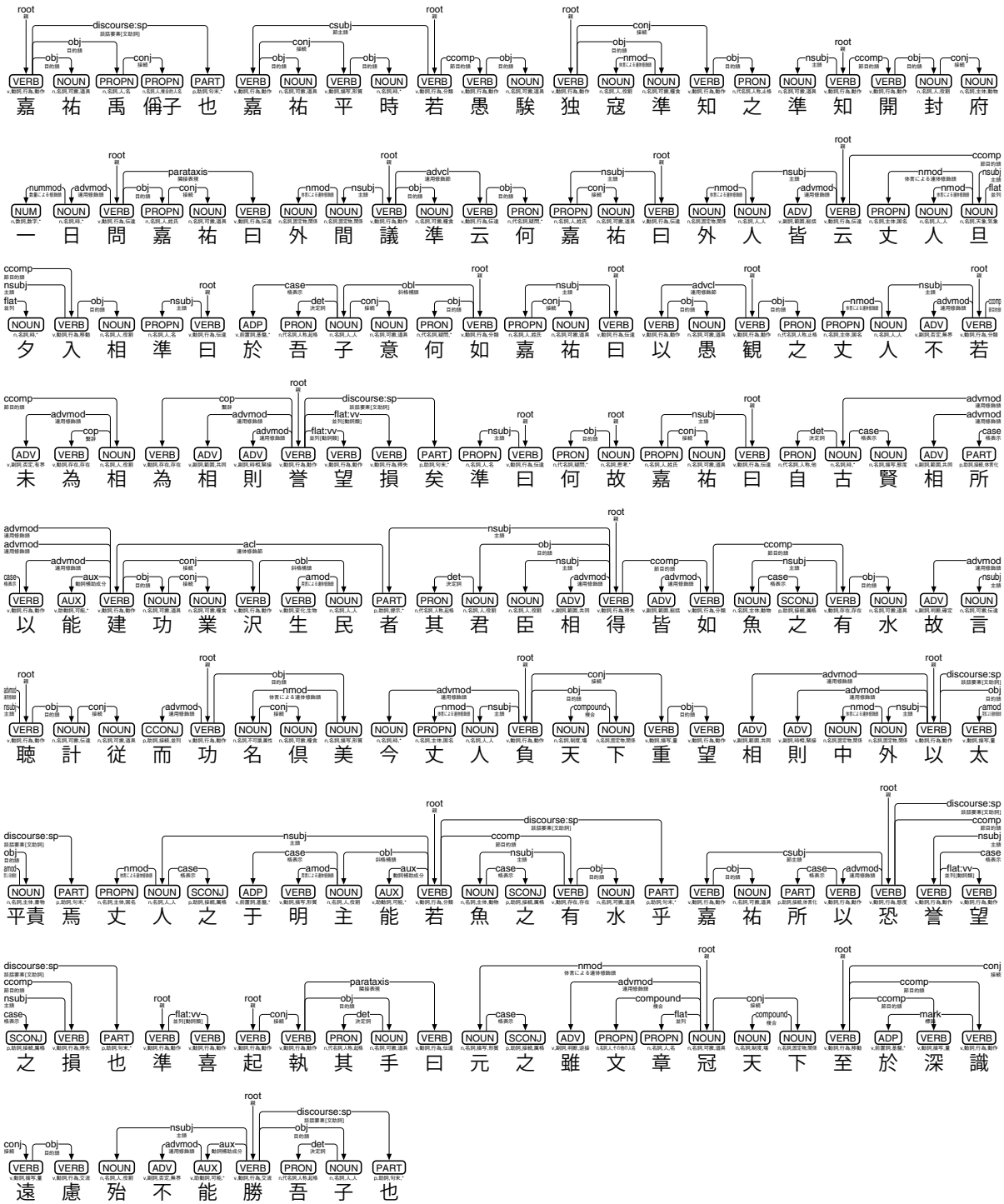


図 36: 手法Aの処理結果 (2018 年)

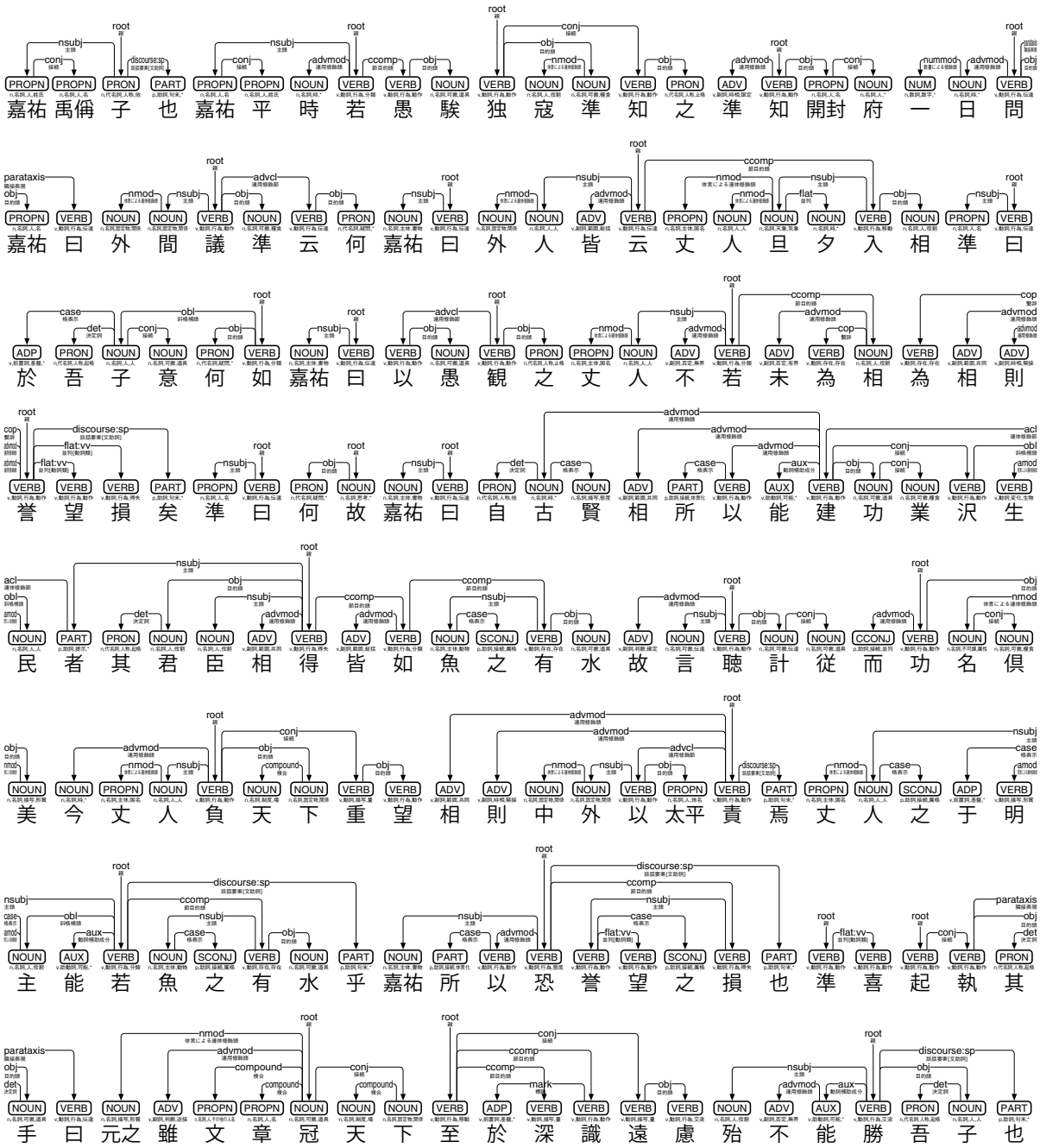


図 37: 手法Bの処理結果 (2018 年)

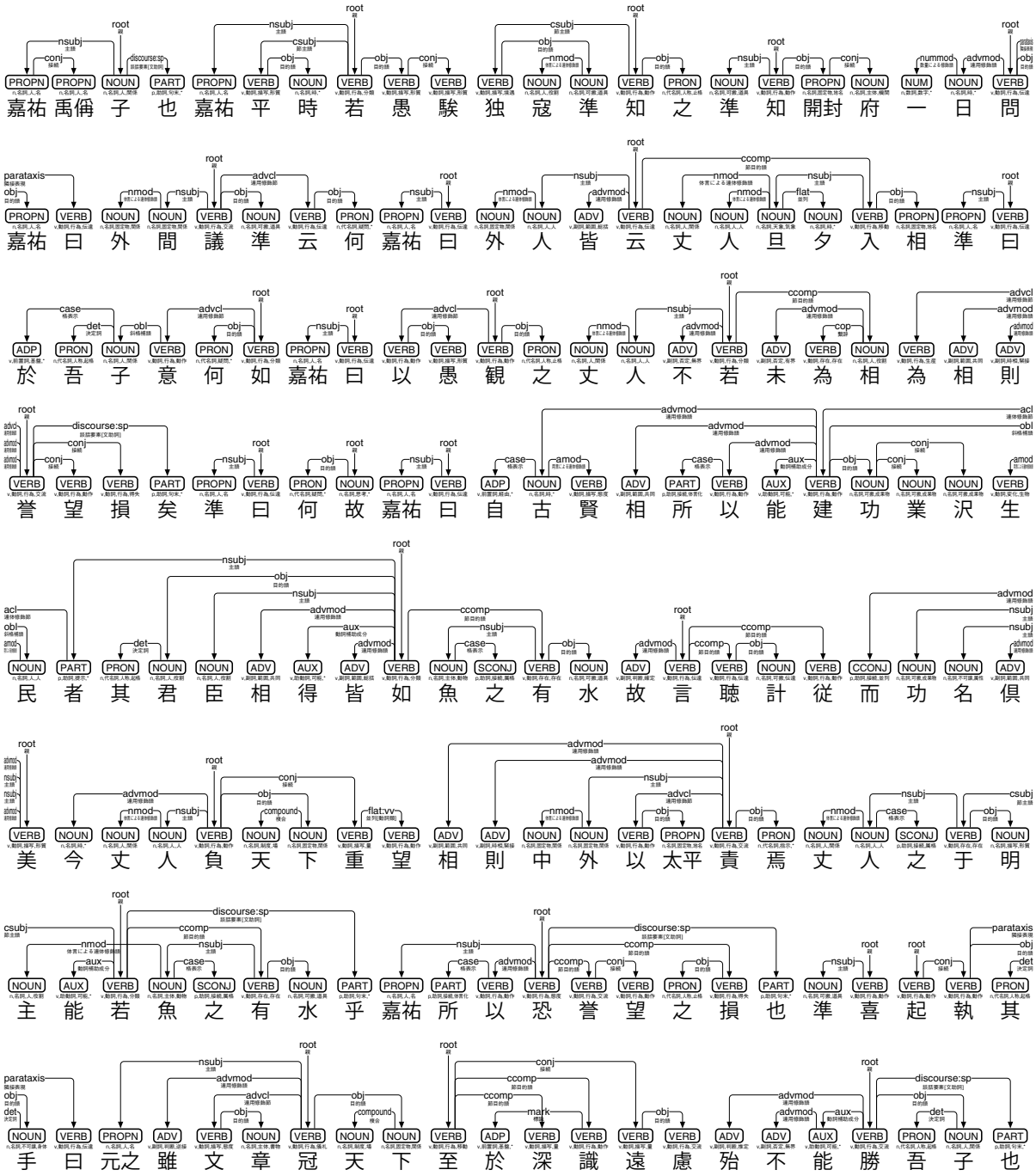


図 38: 手法Cの処理結果 (2018 年)

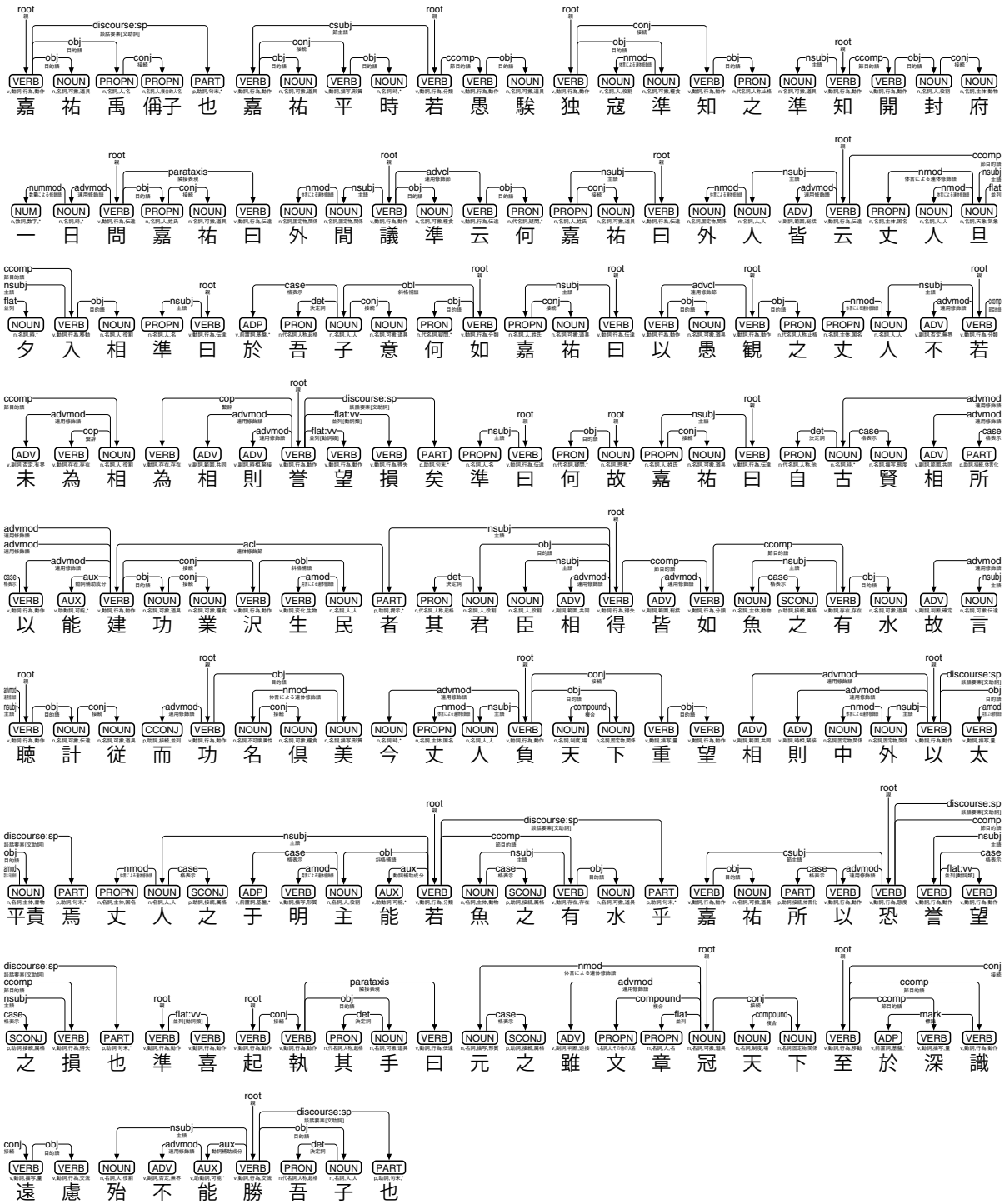


図 39: 手法Ⓐの処理結果 (2018 年)

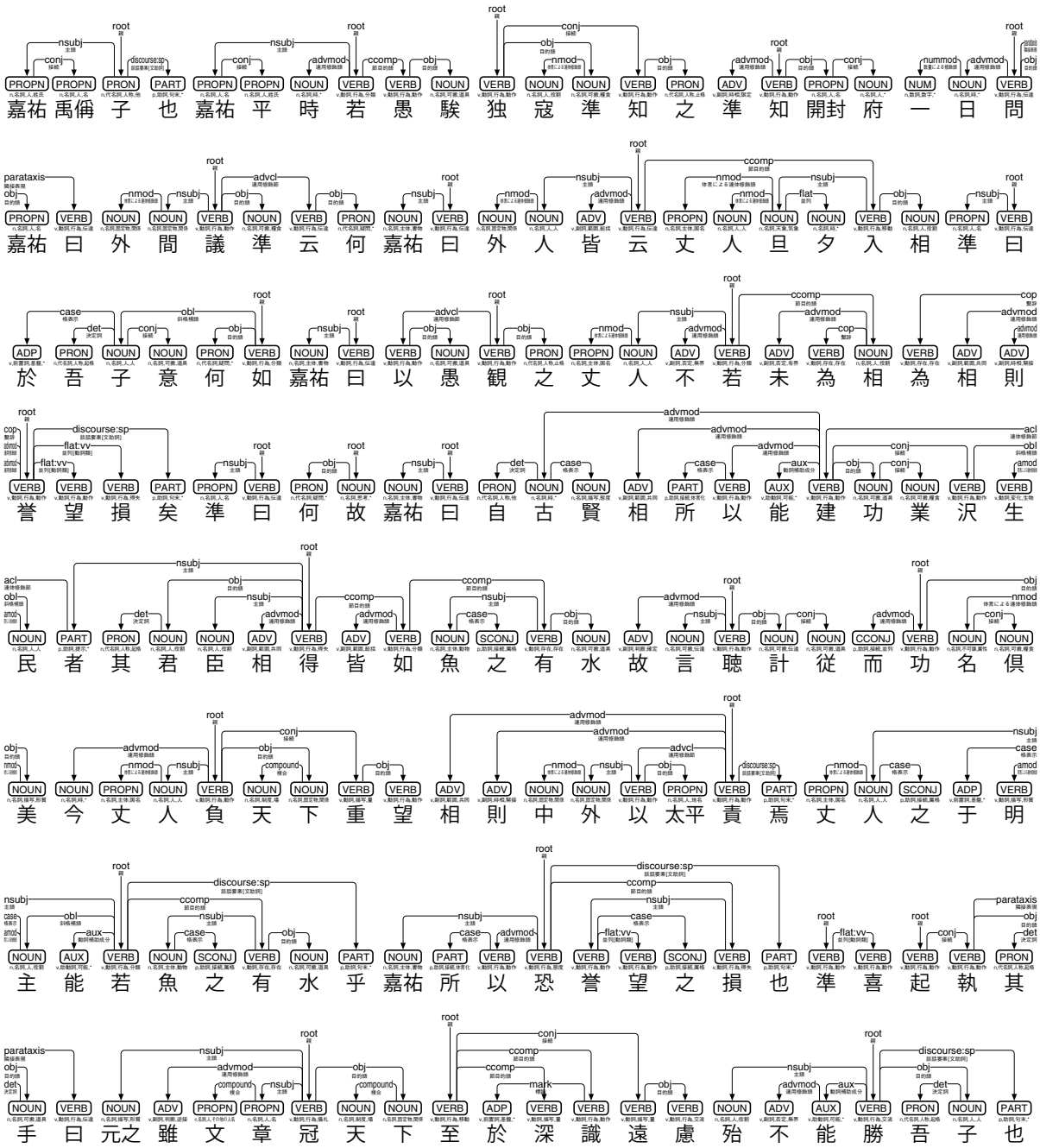


図 40: 手法Ⓑの処理結果 (2018 年)

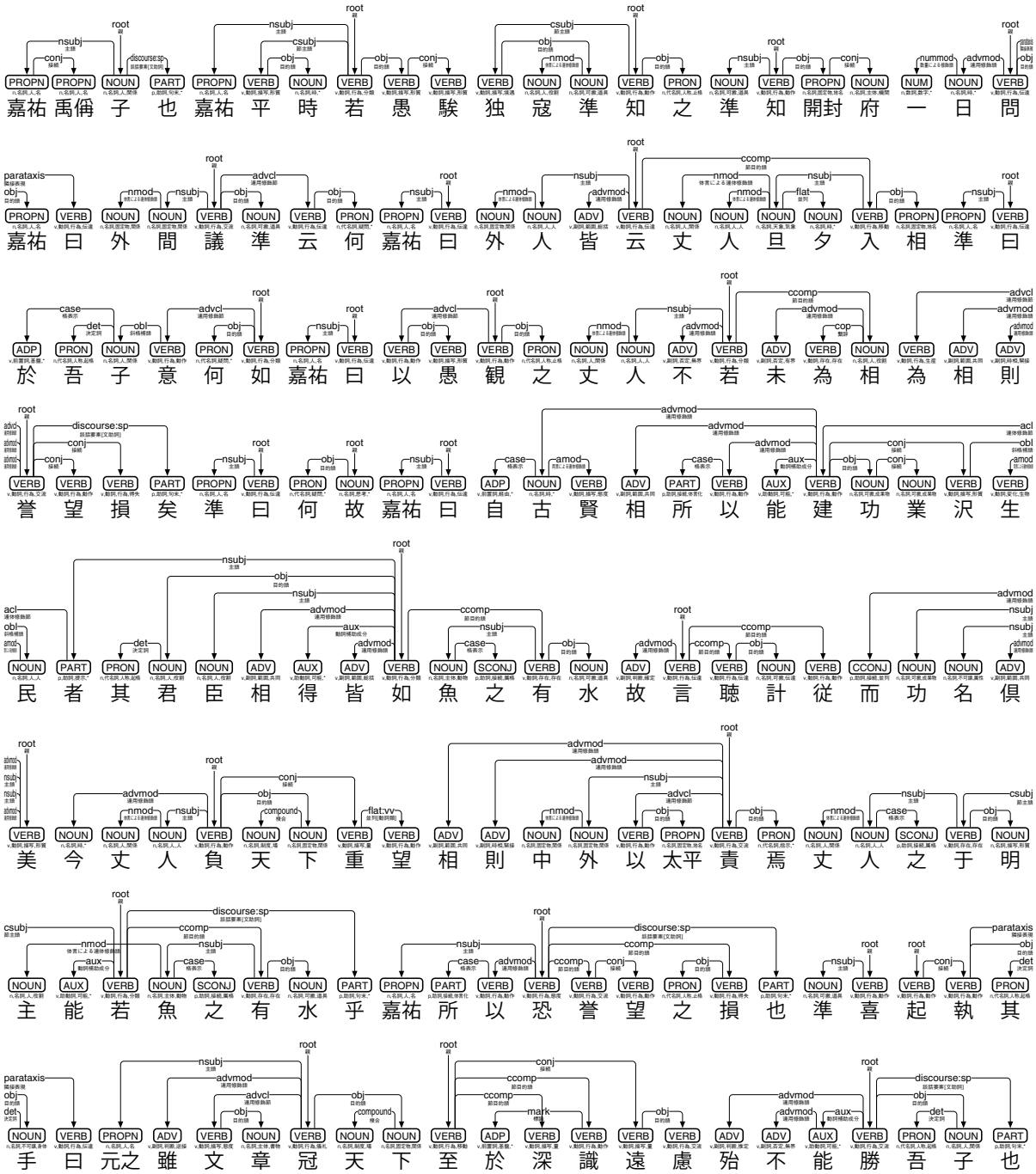


図 41: 手法Ⓒの処理結果 (2018 年)

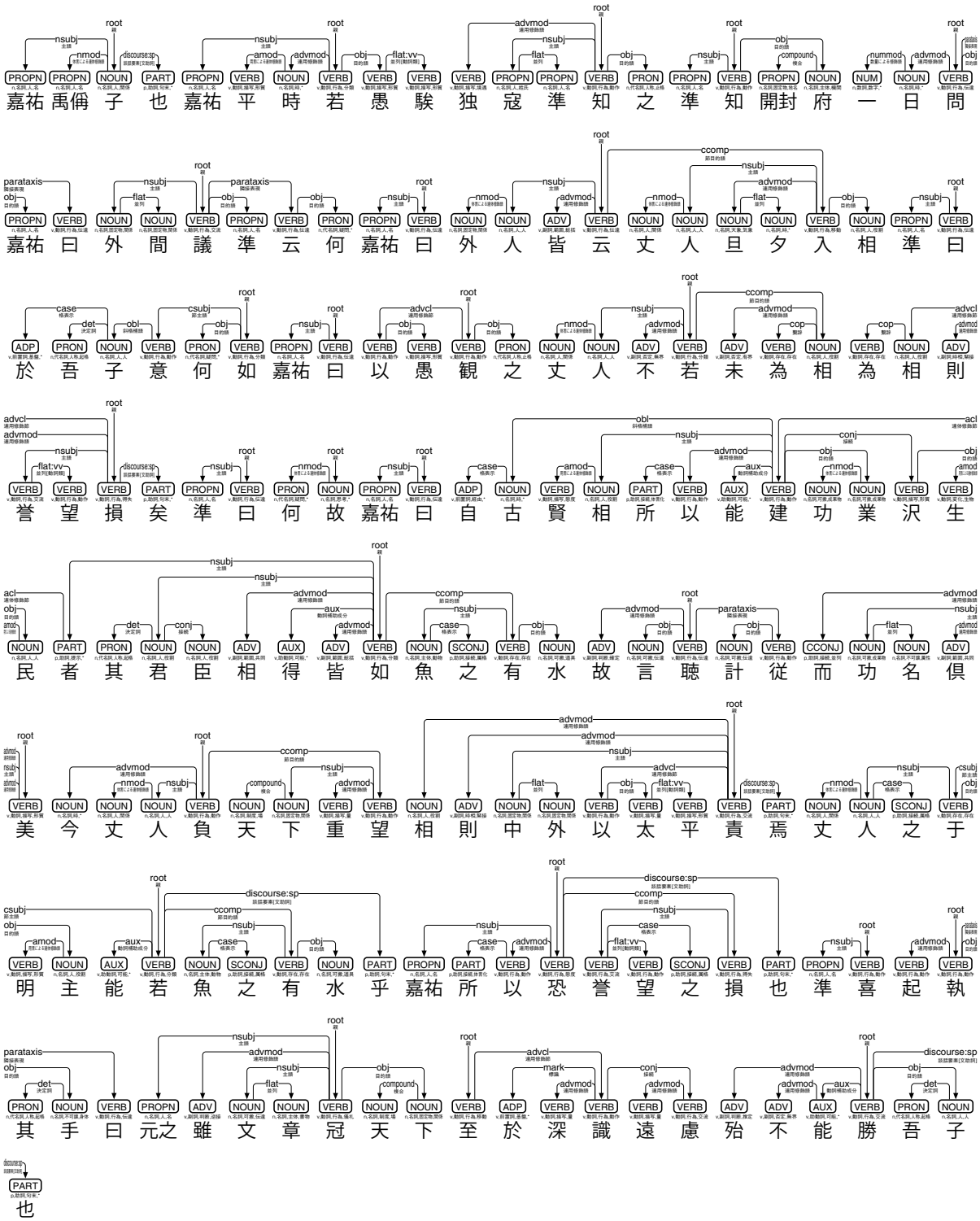


図 42: 手作業で作成した「正解」UD (2018 年)

A

聴ニ雷霆(注1)於百里之外一者、如レ鼓レ盆(注2)、望ニ江河ヲ於千里之間一者、

如レ縈レ帯、以ニ其ノ相ヒ去ル之遠キヲ一也。故ニ居ニ于千載之下一而求ムルニ之ヲ于千載之上ニ、以ニ相ヒ去ル之遠キヲ一而不レ知ラ有ルヲ其ノ変、則チ猶ホ刻レ舟ニ(注3)求ムルガ剣ヲ。今之所ハ求ムル、非ニ往ザルモ者ノ所ニ失フ一而謂ヒ其ノ刻ミシハリ在レ此ニ、是レ所ニ従ヒテ墜ツ一也。

B

豈ニ不レ惑ヒナラ乎。

今夫レ江戸者、世之所ノ称スル名都大邑(注4)、冠(注5)蓋之所レ集マル(2)、舟車之所ニシテ湊マル、実ニ為タル天下之大都会一也。而レドモ其地之為名、訪之於古、未之聞。豈ニ非ズ古今相ヒ去ルコト日ビニ遠ク、而事物之変モ亦マタ在ルニ于其ノ間一耶。(ア)蓋知ル、後之於ケルモ今ニ、世之相ヒ去ルコト愈遠ク、事之相ヒ変ズルコト愈多ク、求ムルニ其ノ所ヲ欲スレ聞カント(1)而不レ可カラ得、亦タ猶ホ今之於ケルガ古ニ也。

C

D

吾窃ニ有リ感ズレ焉。『遺聞』(注6)之書、所ニ由リテ作ル一也。

図 43: 大学入試センター試験『国語』(2017 年 1 月 14 日) 第 4 問本文

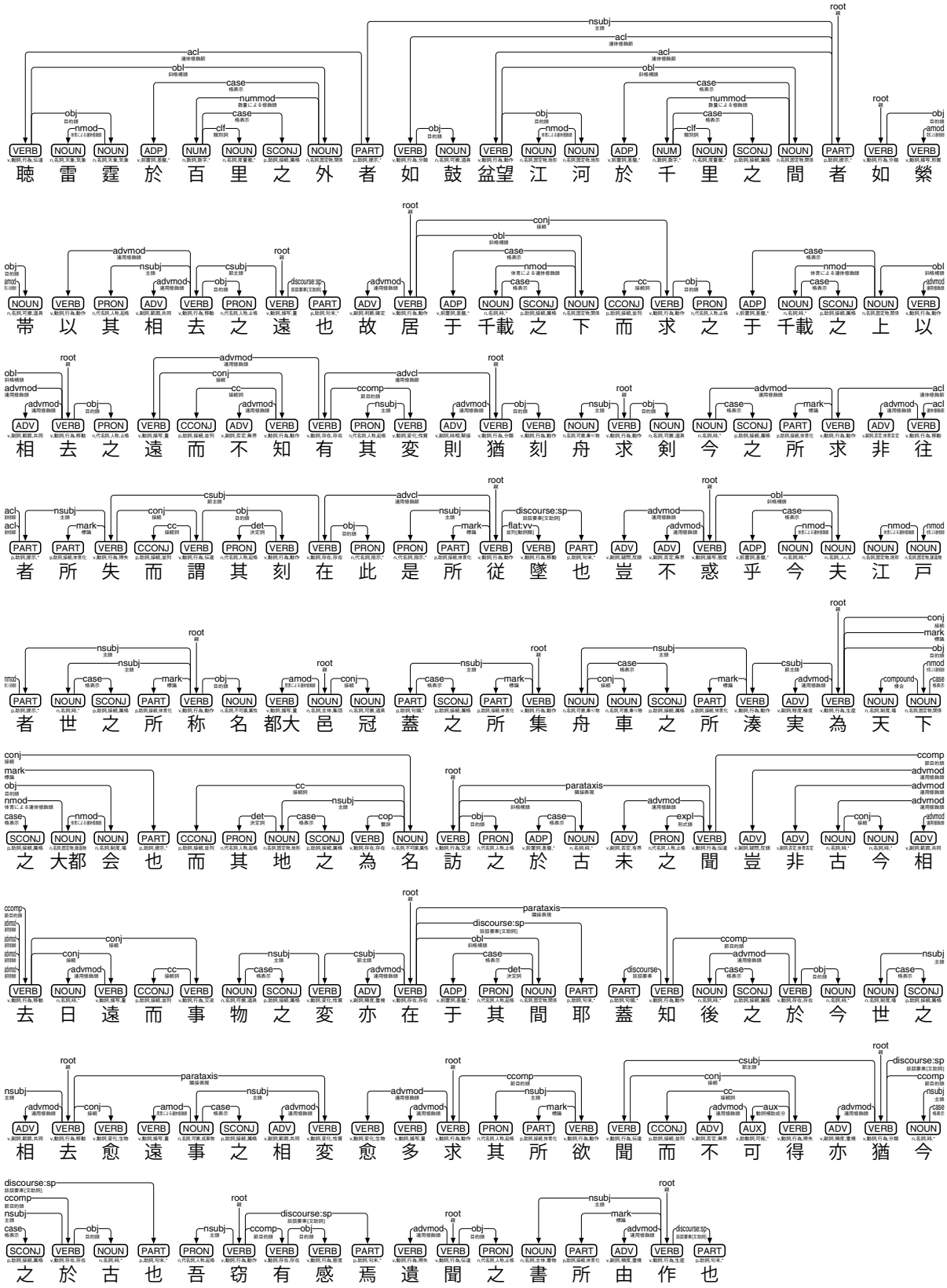


図 44: 手法①の処理結果 (2017 年)

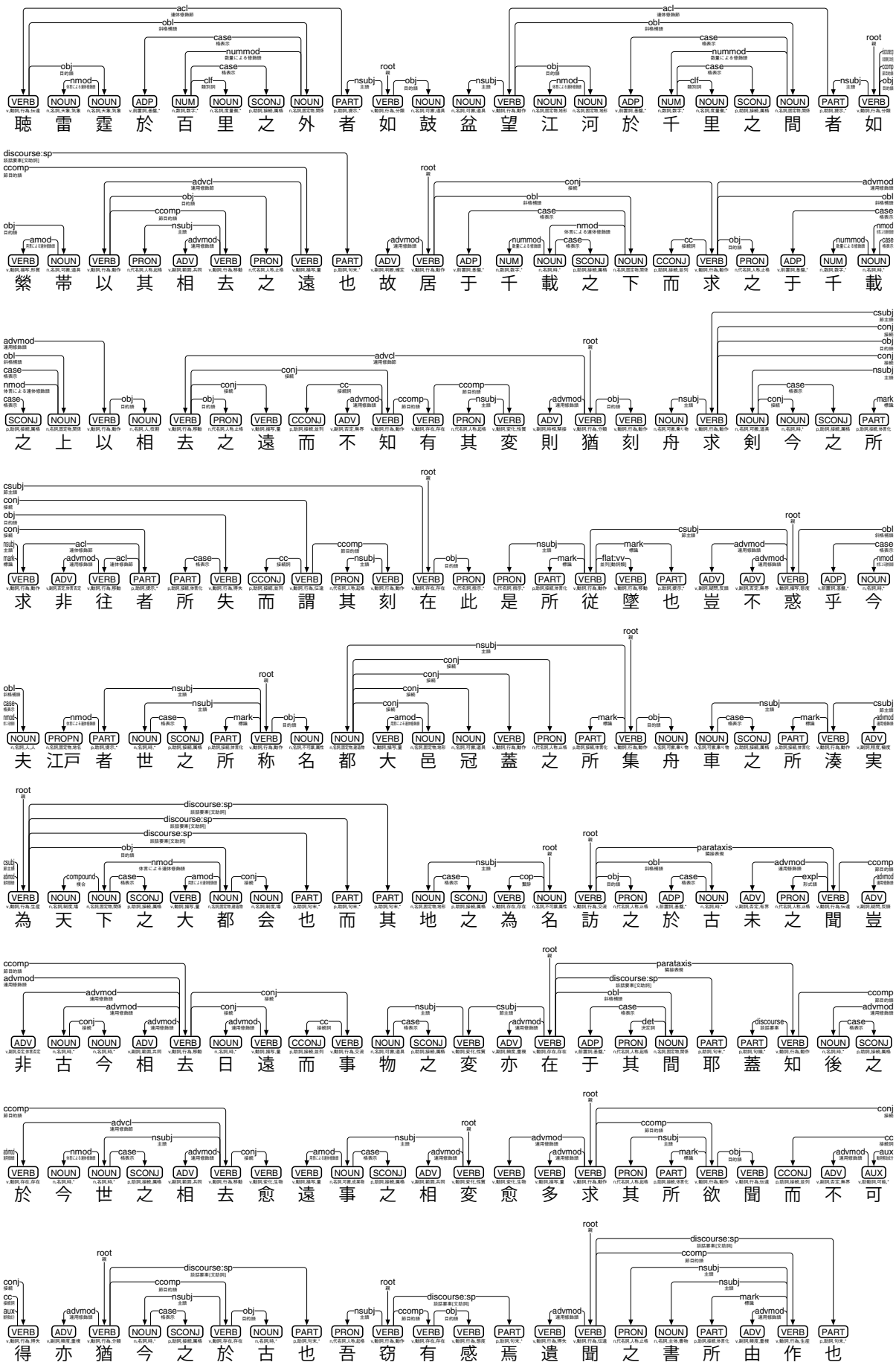


図 45: 手法②の処理結果 (2017 年)

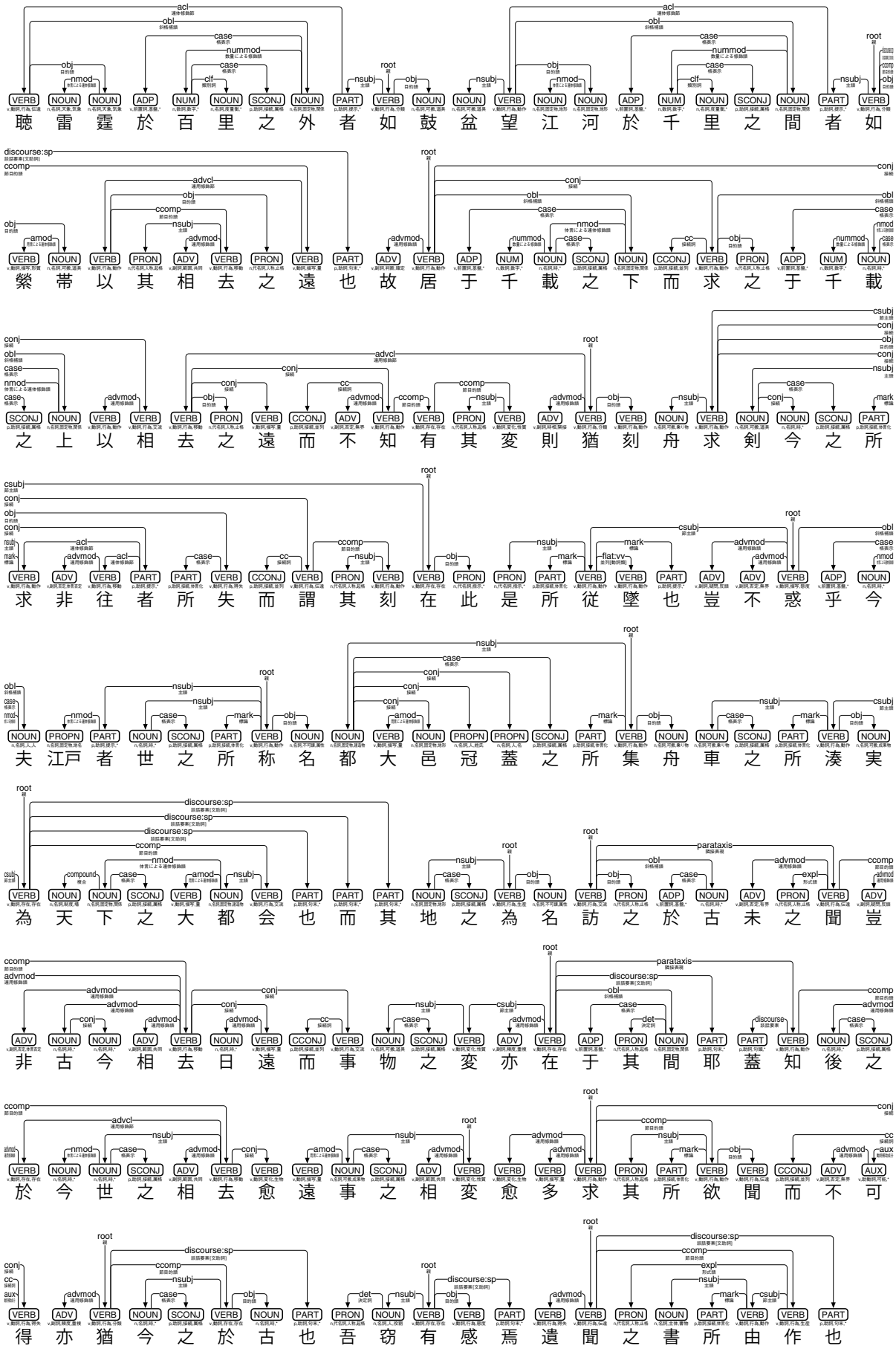


図 46: 手法③の処理結果 (2017 年)

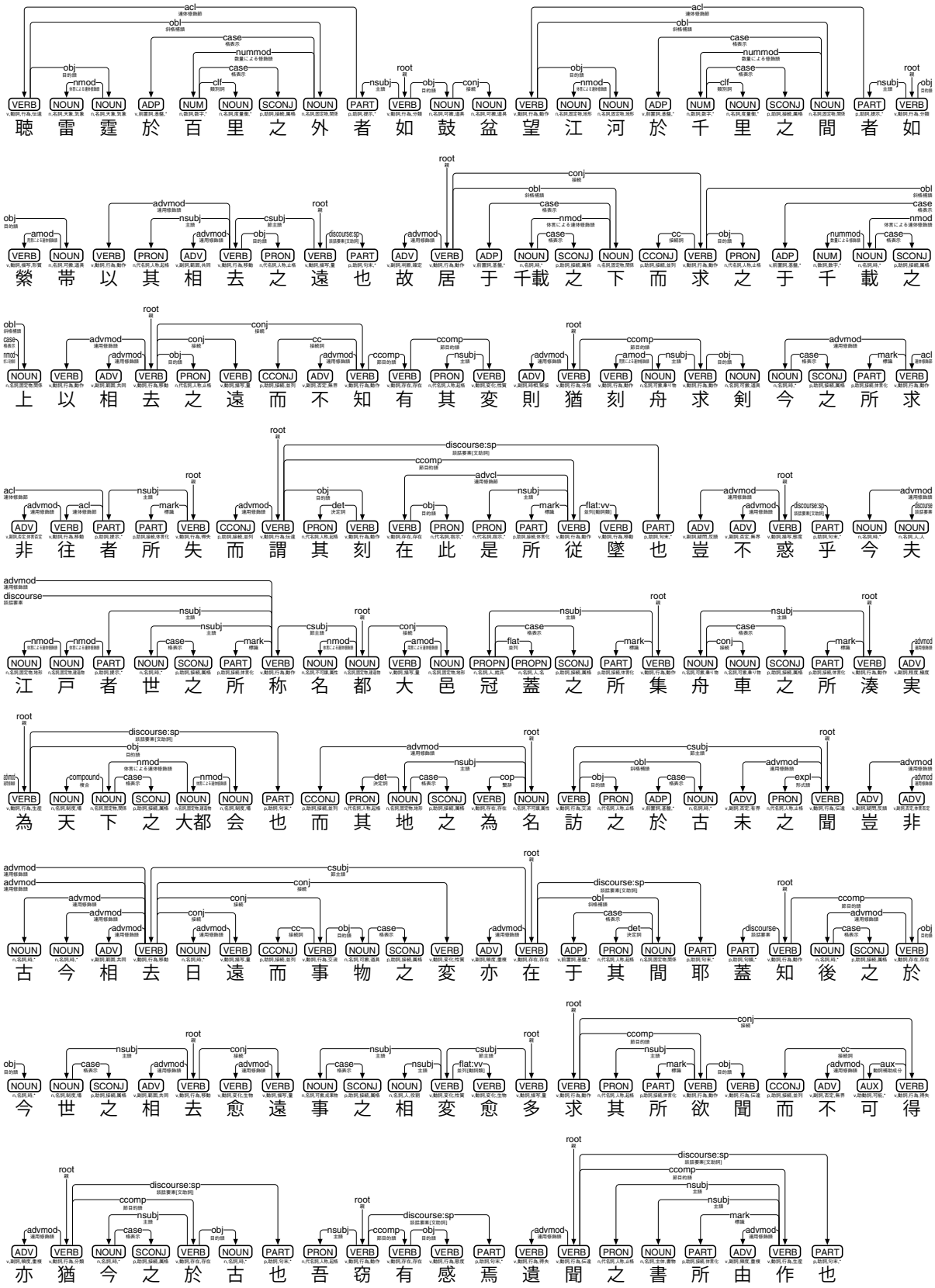


図 47: 手法1の処理結果 (2017 年)

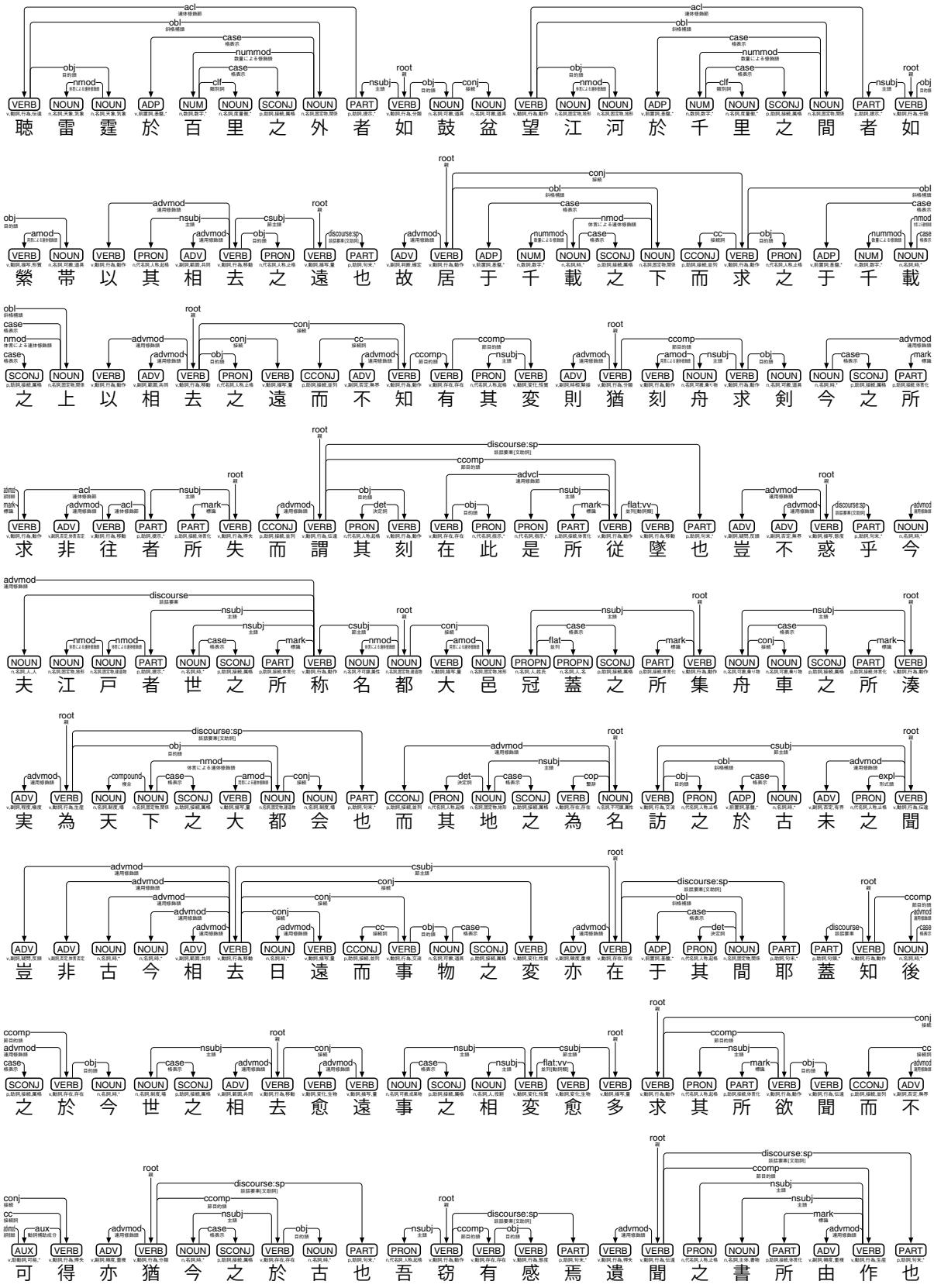


図 48: 手法2の処理結果 (2017 年)

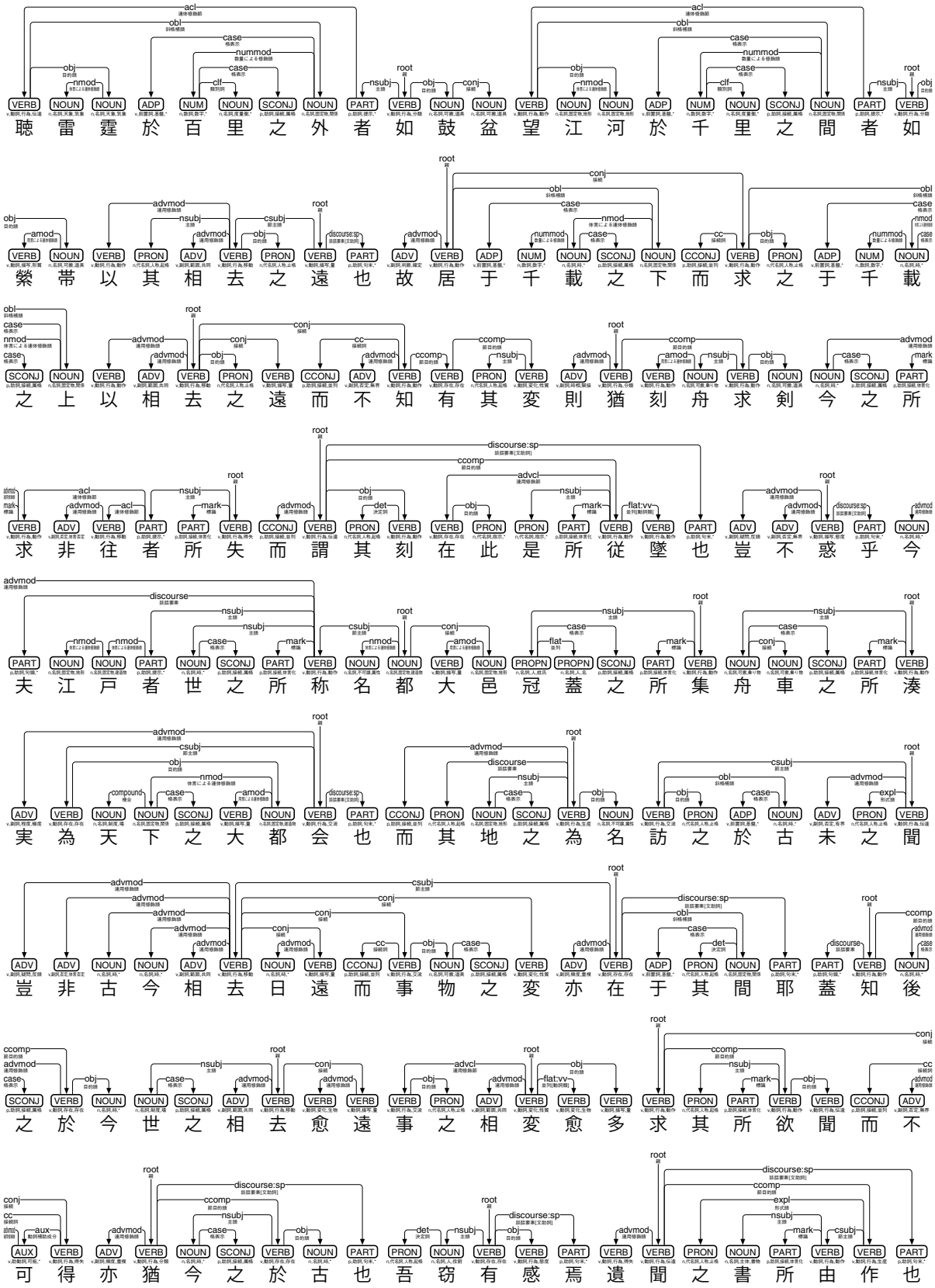


図 49: 手法3の処理結果 (2017 年)

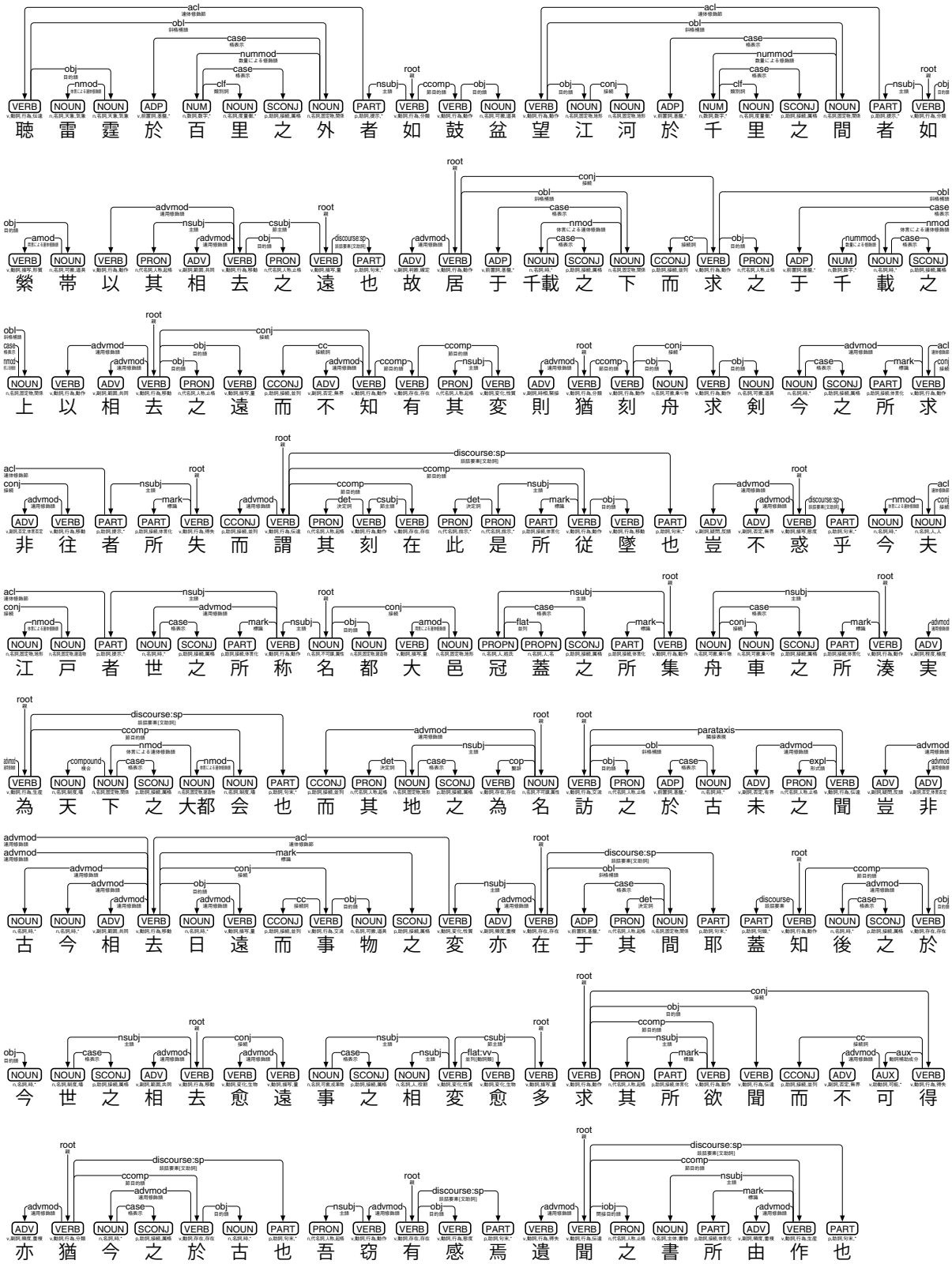


図 50: 手法❶の処理結果 (2017 年)

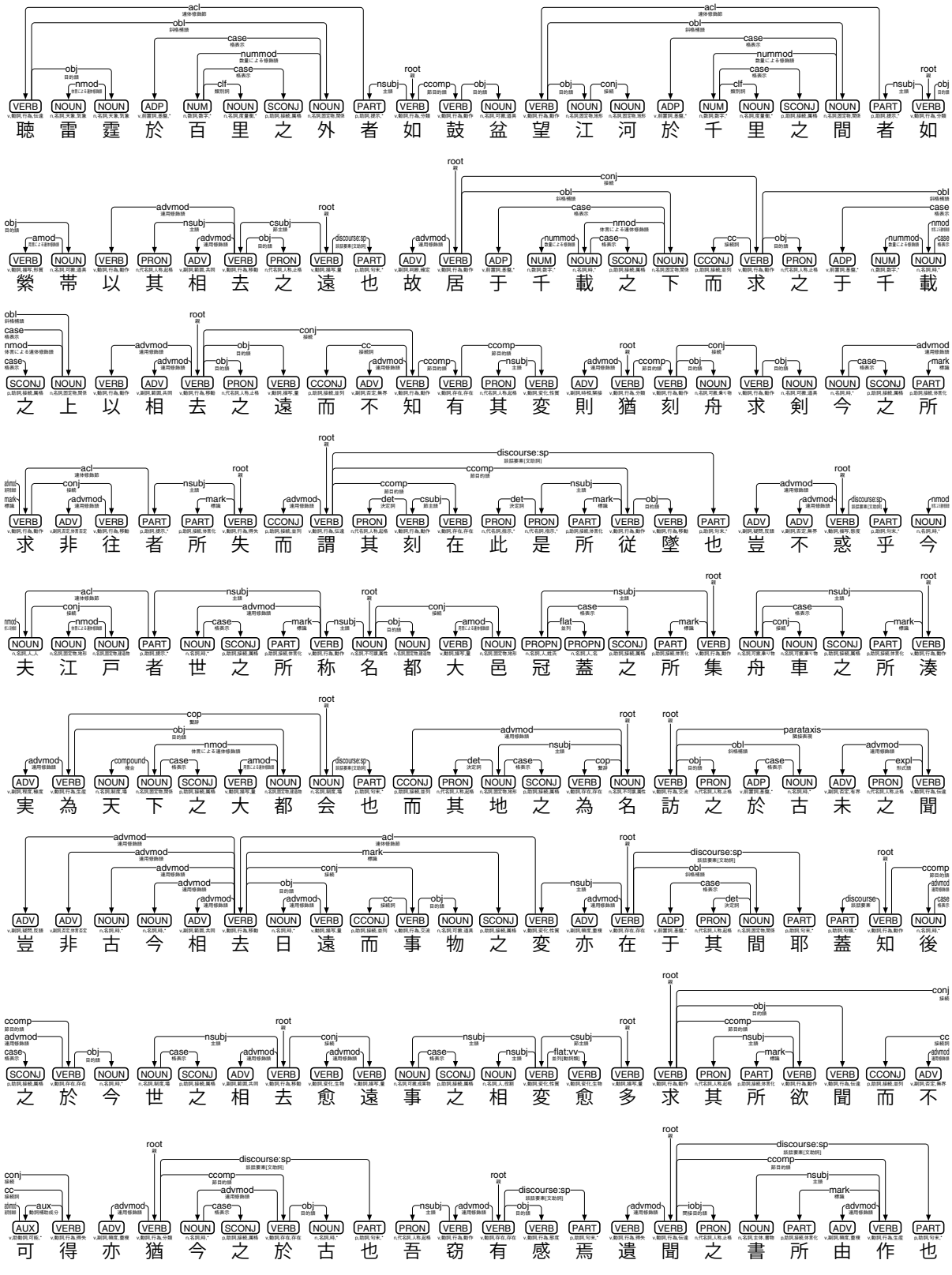


図 51: 手法❷の処理結果 (2017 年)

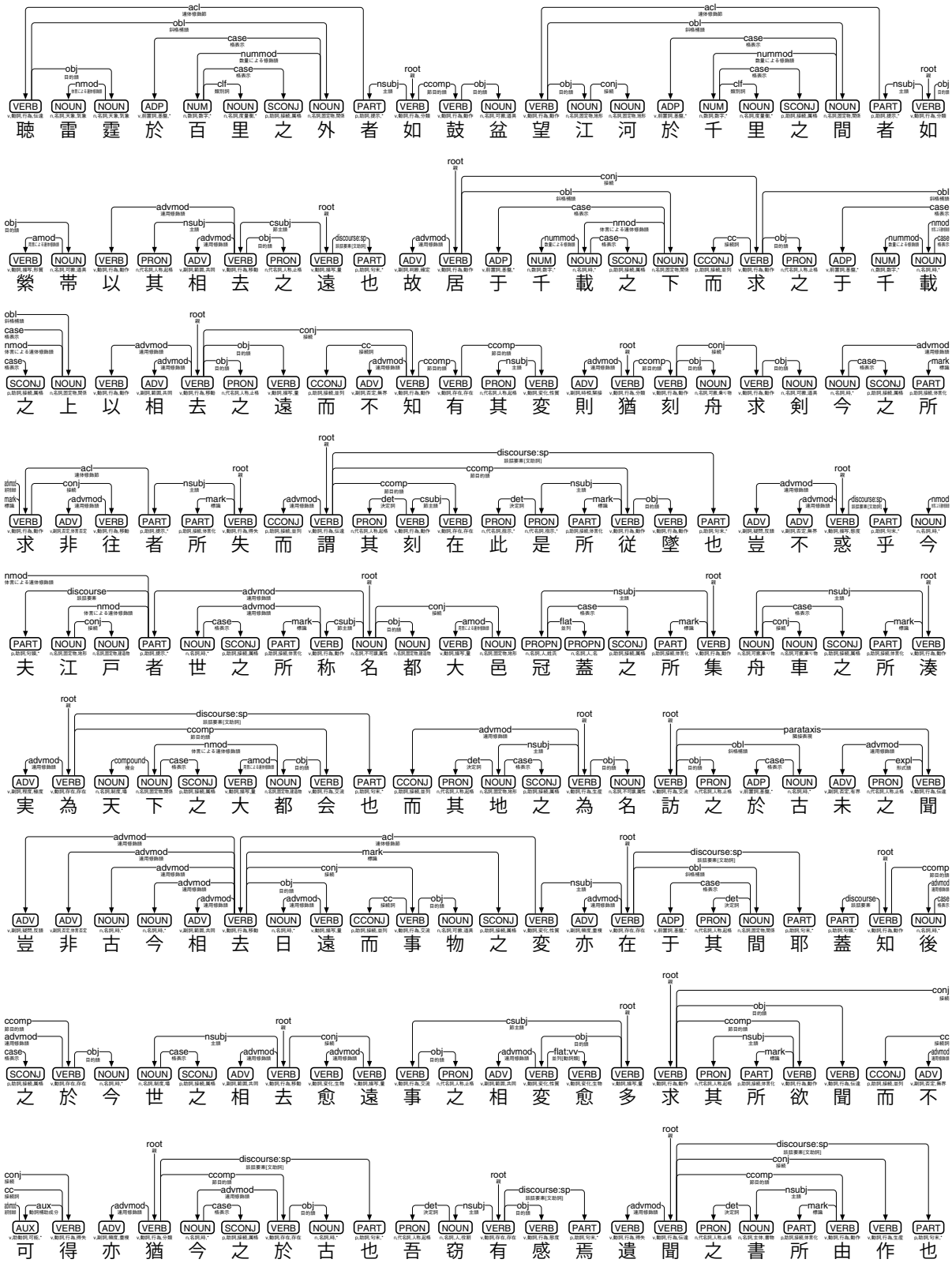


図 52: 手法❸の処理結果 (2017 年)

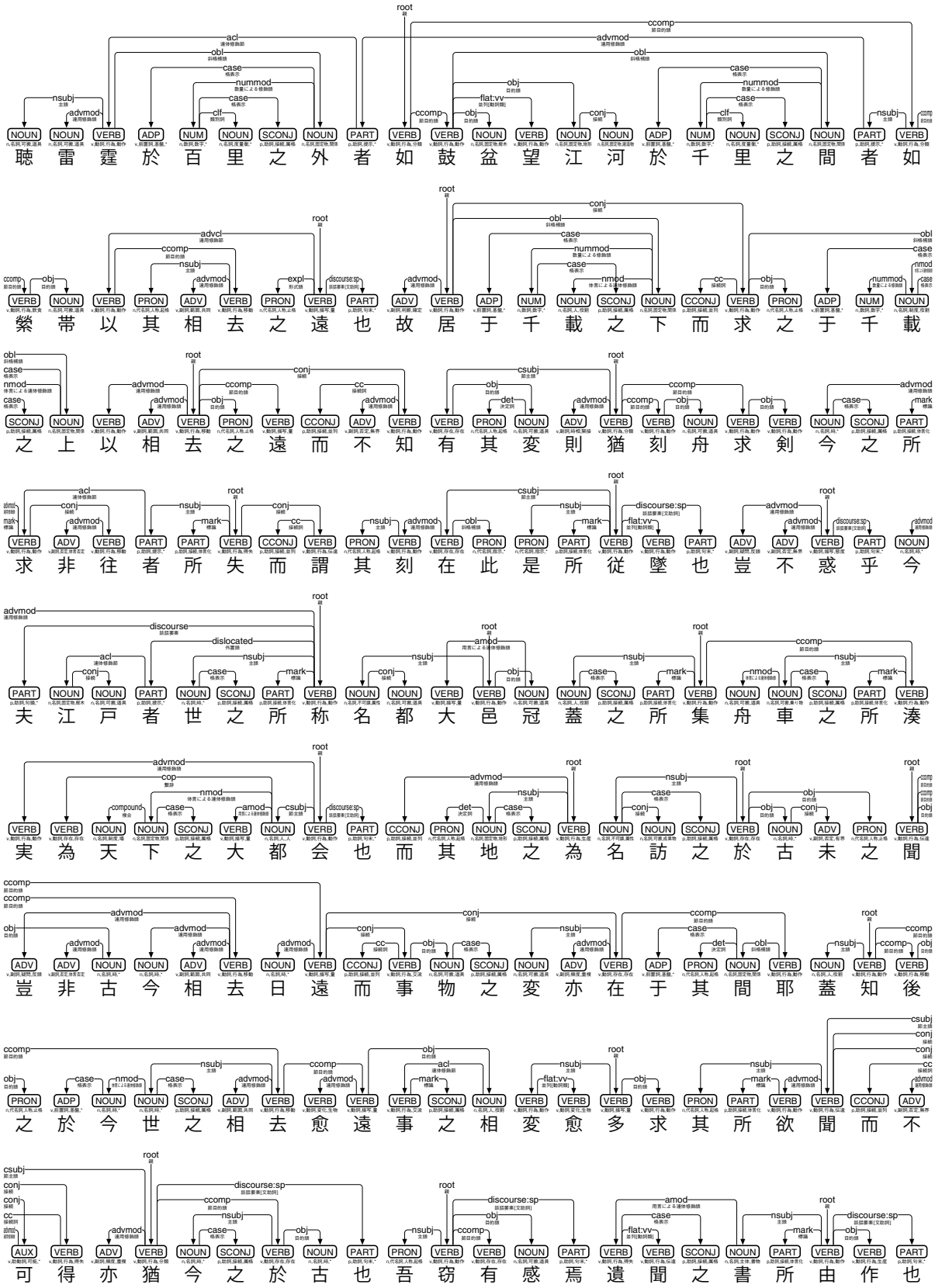


図 53: 手法Ⓐの処理結果 (2017 年)

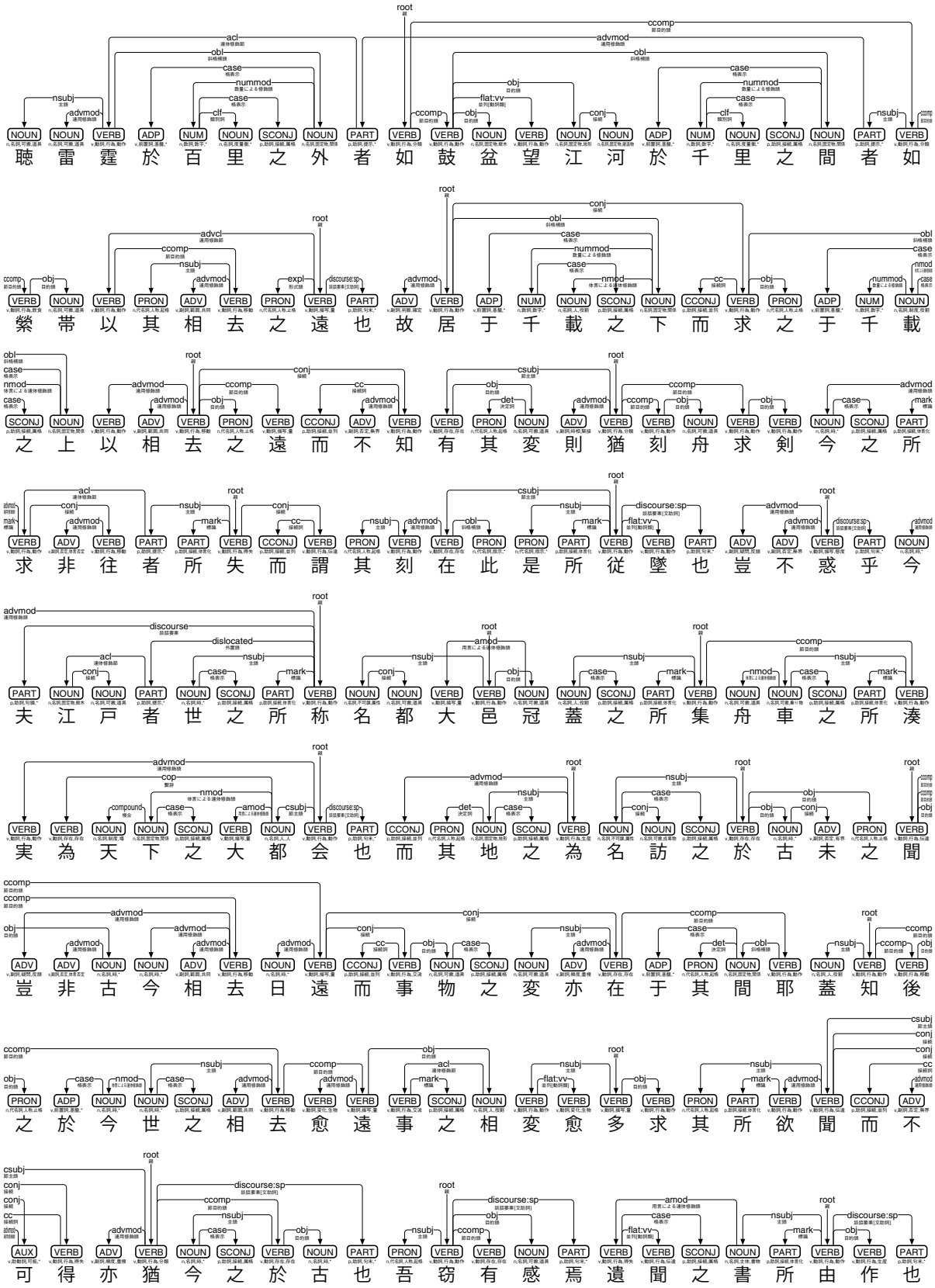


図 54: 手法Ⓑの処理結果 (2017 年)

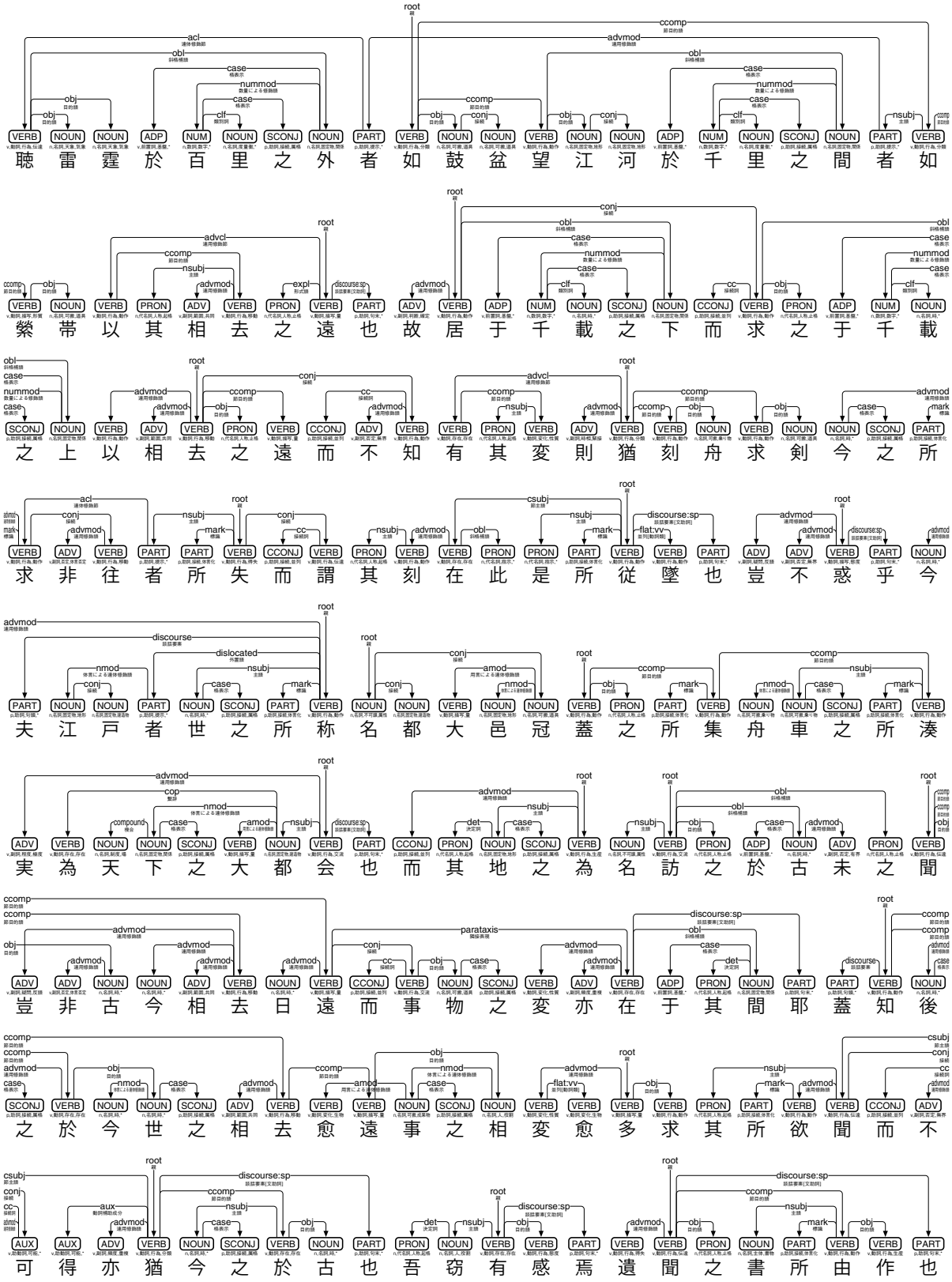
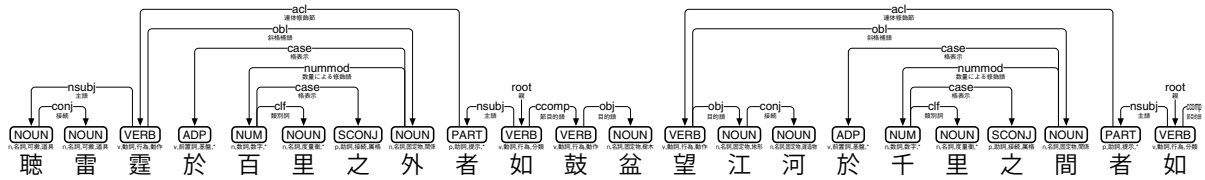
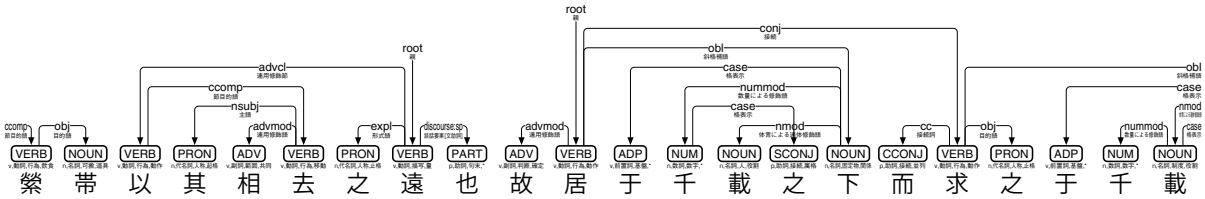
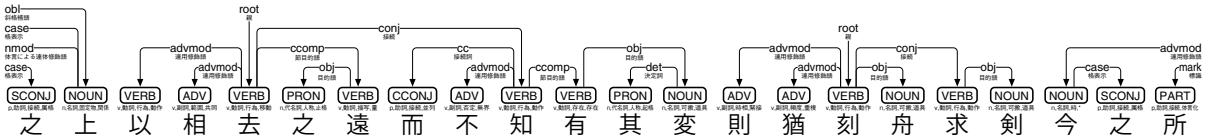
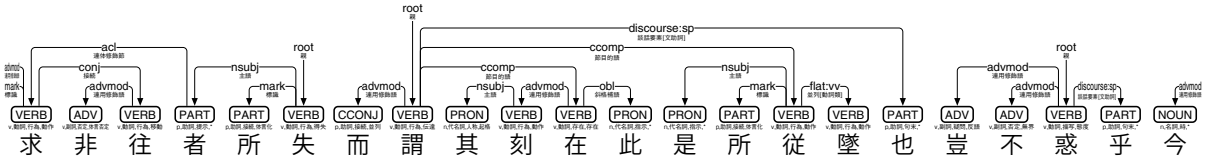


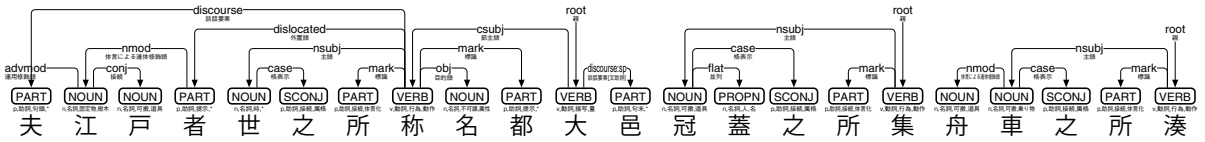
図 55: 手法Ⓒの処理結果 (2017 年)

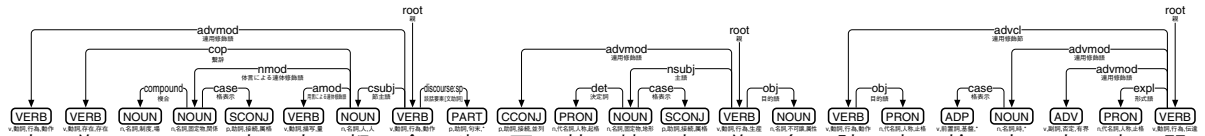


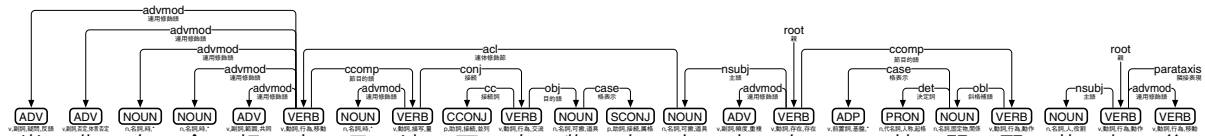


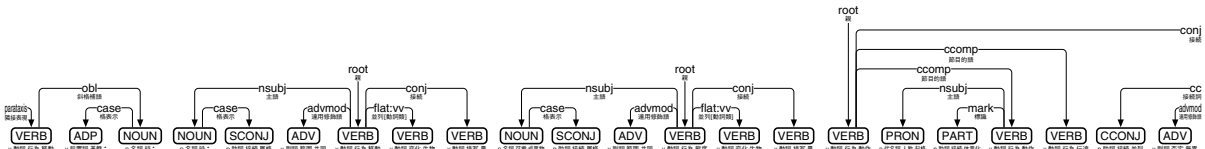












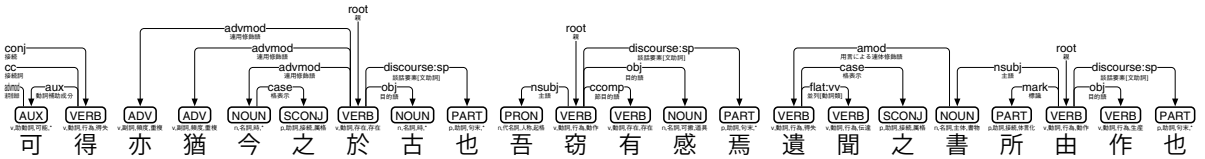


図 56: 手法Aの処理結果 (2017 年)

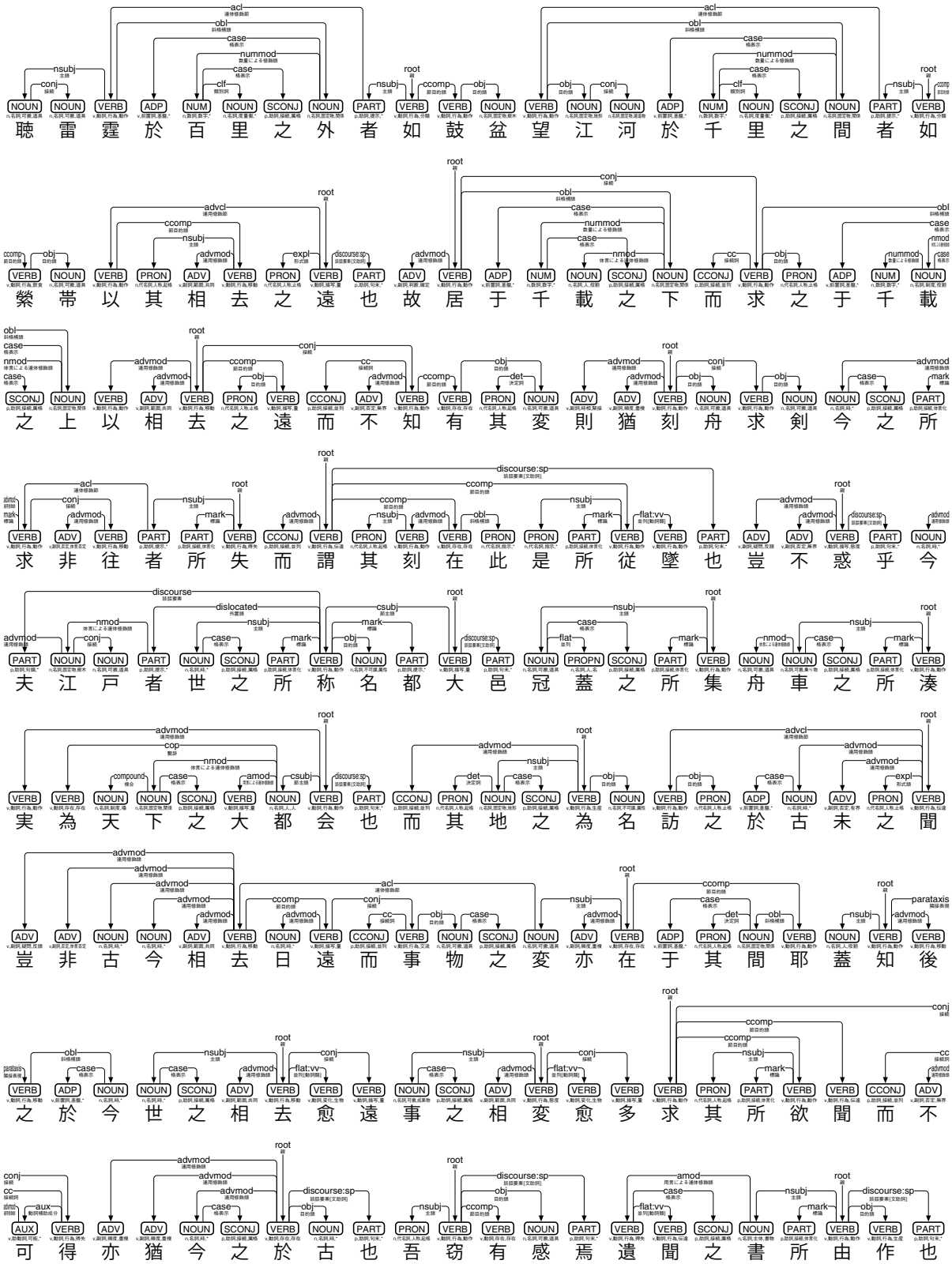


図 57: 手法Bの処理結果 (2017 年)

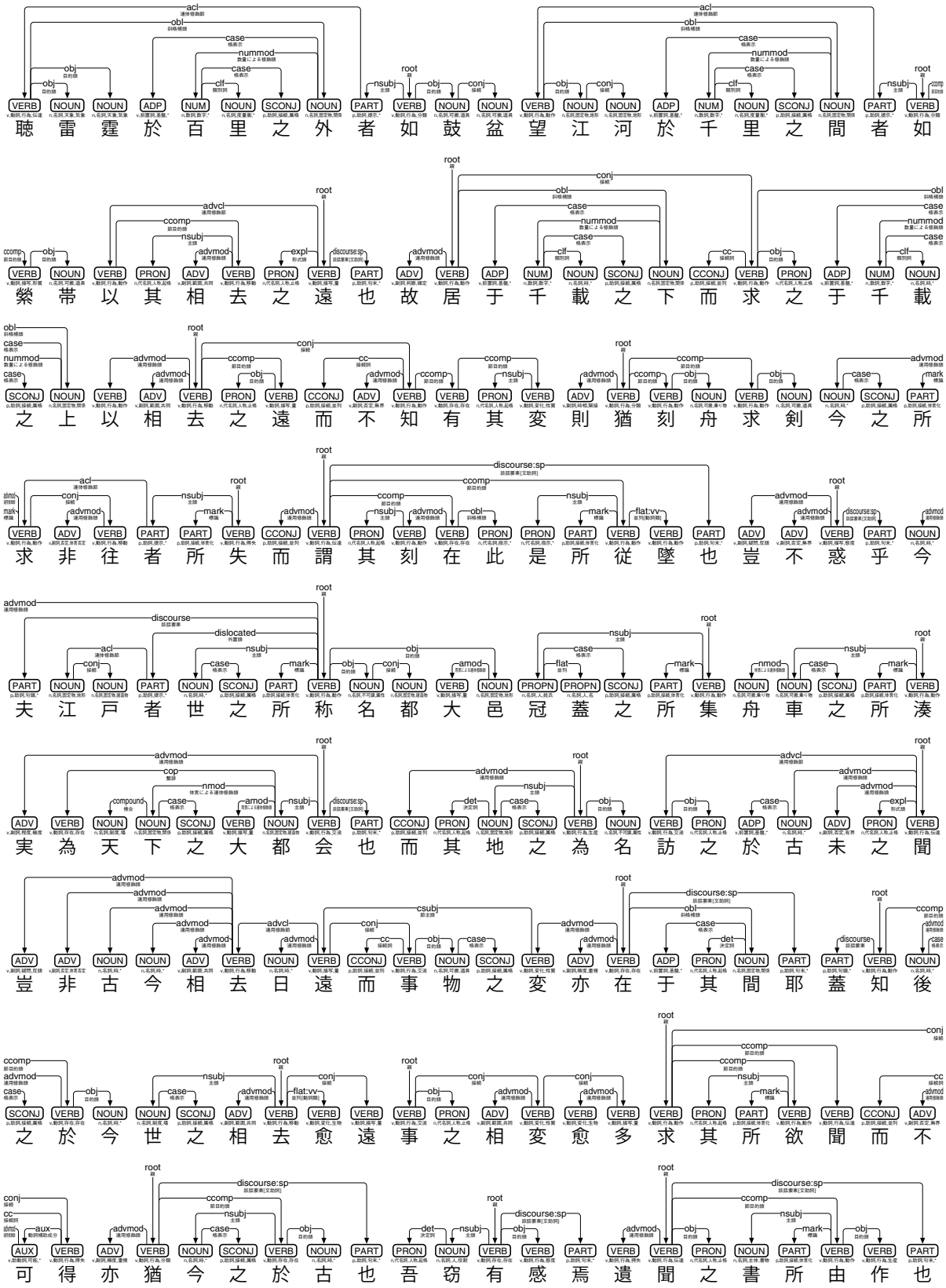


図 58: 手法Ⓒの処理結果 (2017 年)

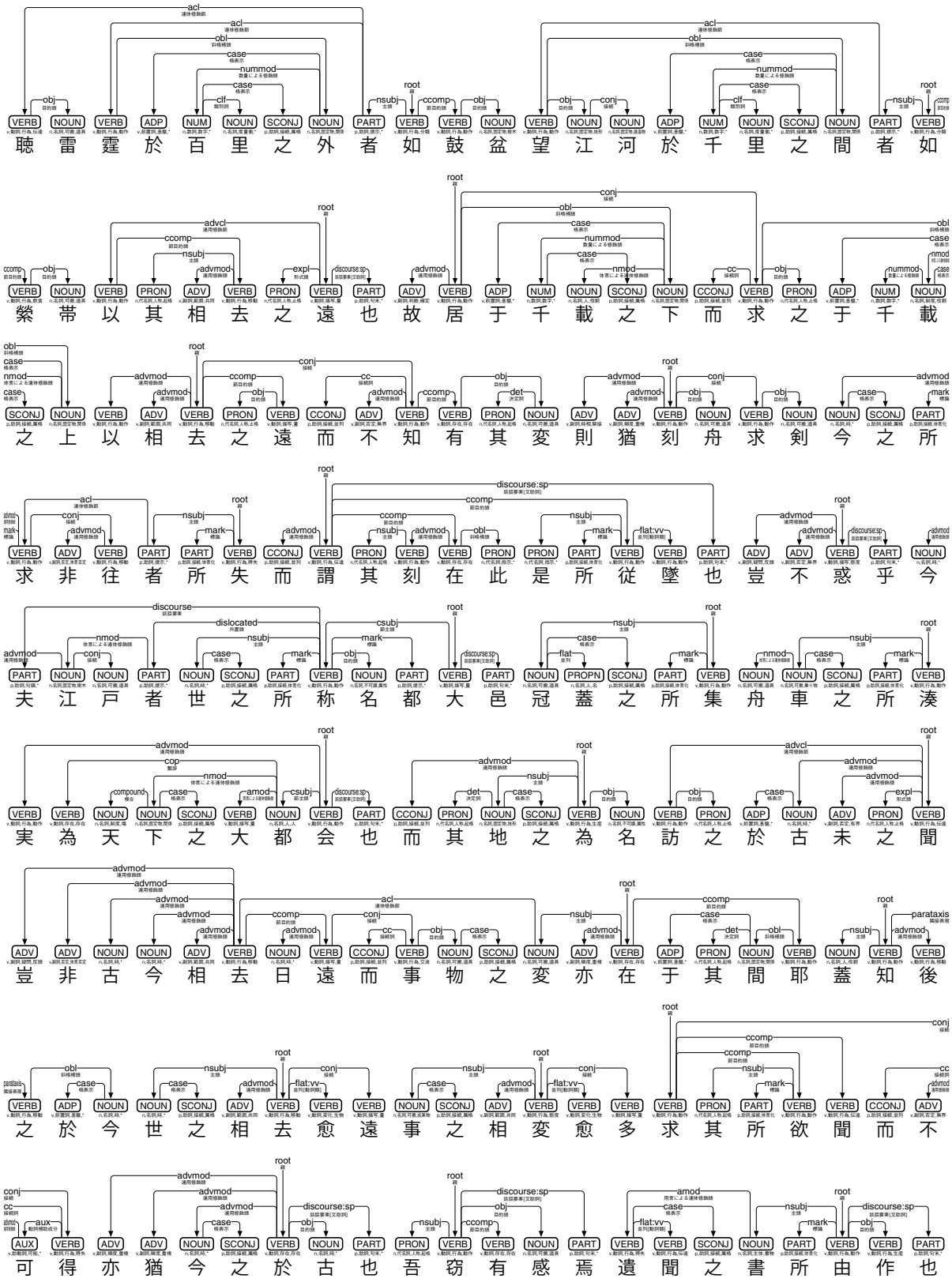


図 59: 手法Ⓐの処理結果 (2017 年)

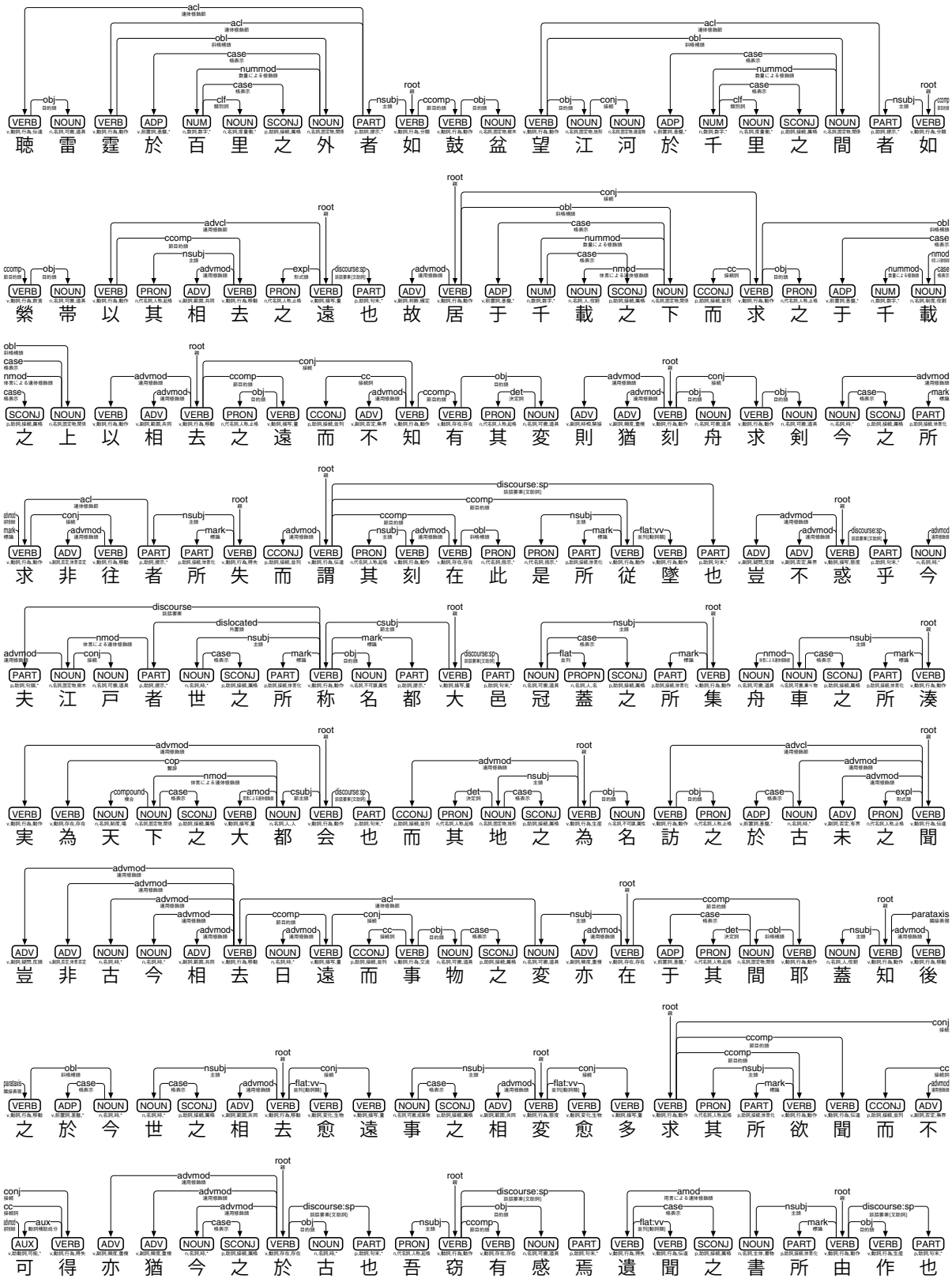


図 60: 手法Ⓑの処理結果 (2017 年)

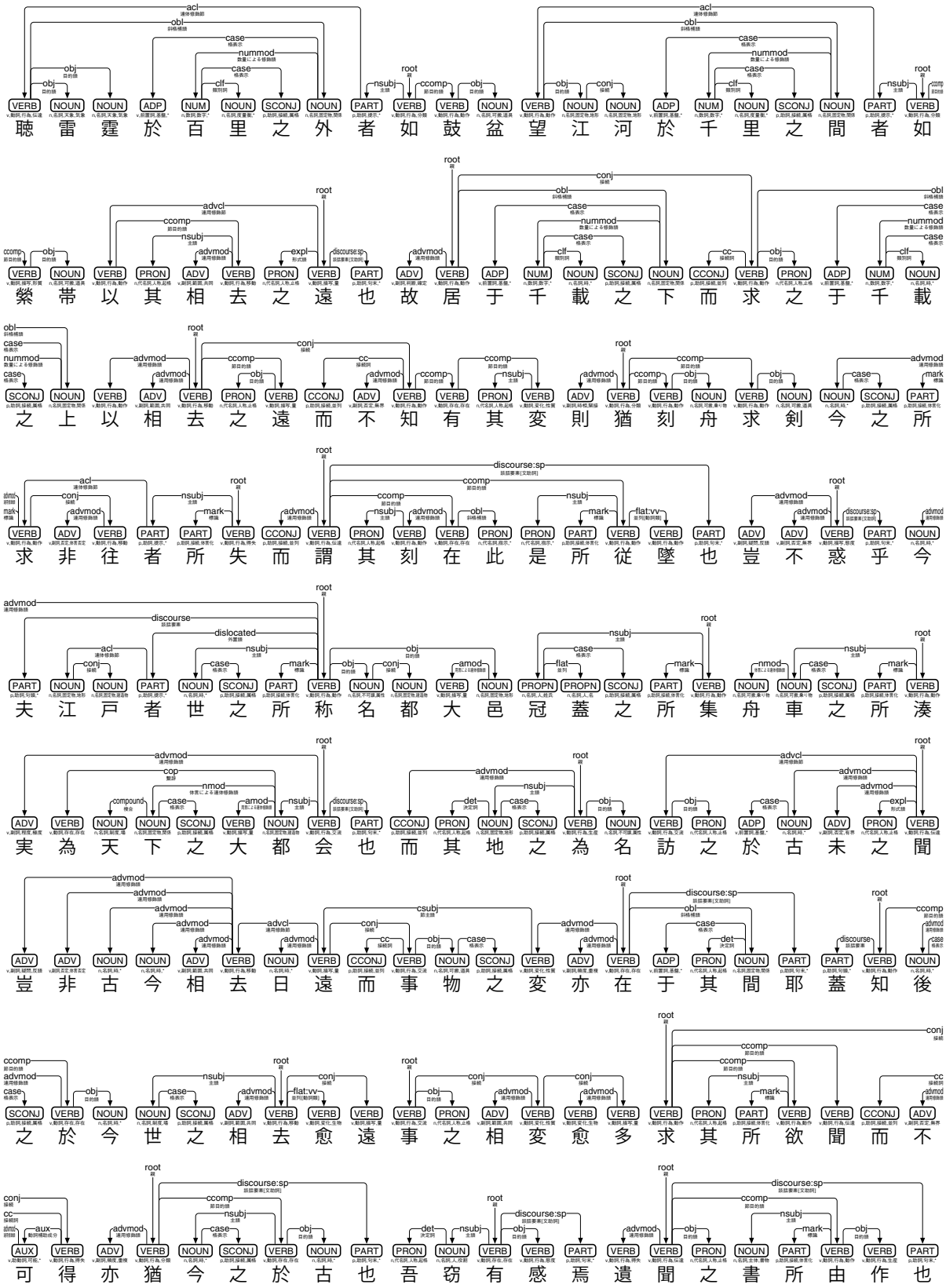


図 61: 手法Ⓒの処理結果 (2017 年)

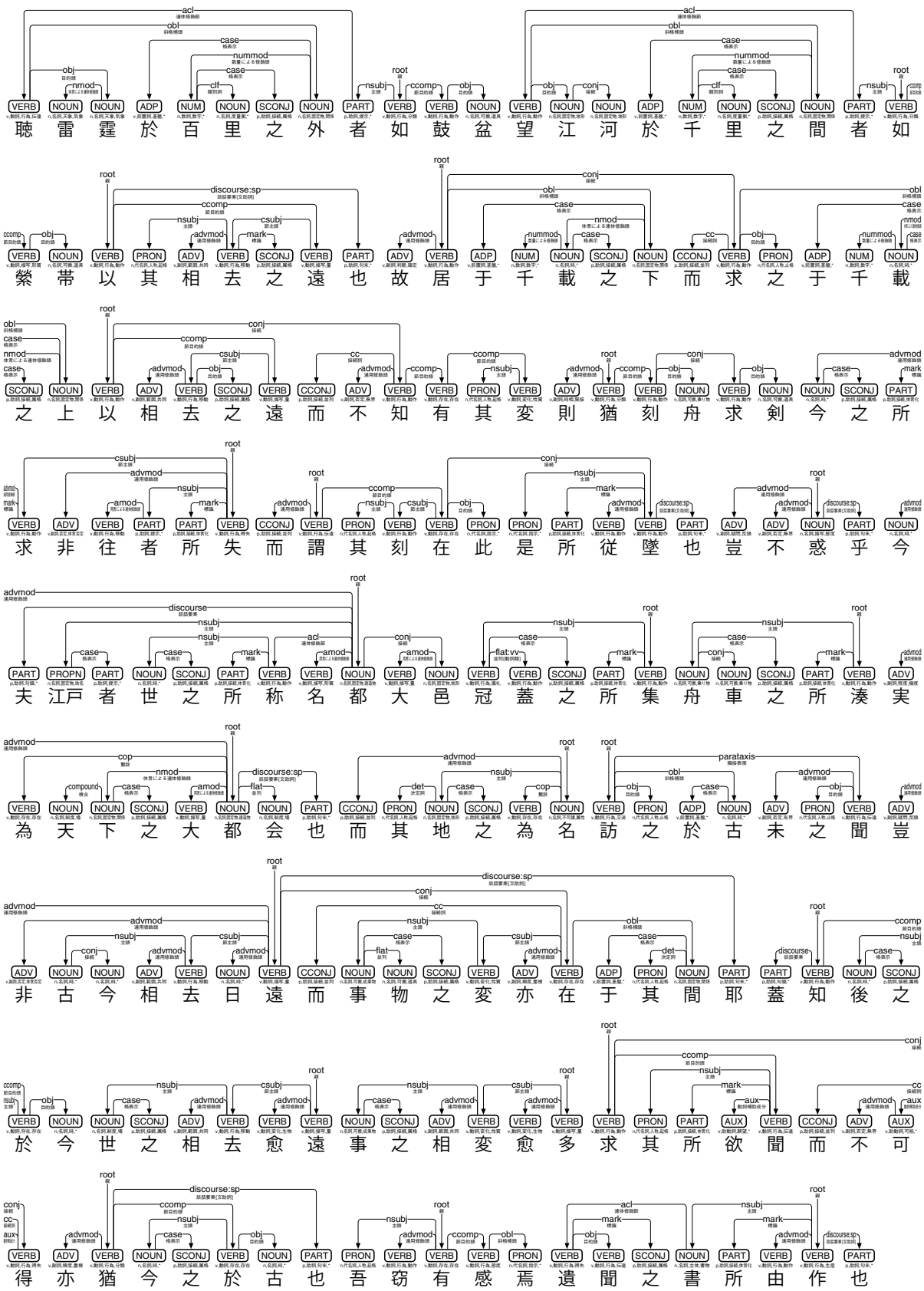


図 62: 手作業で作成した「正解」UD (2017 年)

荷宇生十月而喪其母。及有知、即時時念母不置、弥久
弥篤。哀其身不能一日事乎母也。哀母之言語動作亦未
能識也。

荷宇香(注1)河人。嘗南遊(2)而反、至乎銭唐。夢母来(注3)前、夢中即
知其為母也。既覚、乃(イ)嗷然(注4)以哭曰「此真吾母也。母胡為乎
使我至今日乃得見也。母、又何去我之速也。母、其可使我
継此而得見也」於是即夢所見為之図。此図吾不之見也。

今之図吾見之、則其夢母之境而已。

余因語之曰「夫人精誠所感、無幽(注5)明死生之隔、此理之
可信不誣(注6)者。況子之於親、其喘息(注7)呼吸相通、本無有間之
者乎。」

図 63: 大学入試センター試験『国語』(2016 年 1 月 16 日) 第 4 問本文

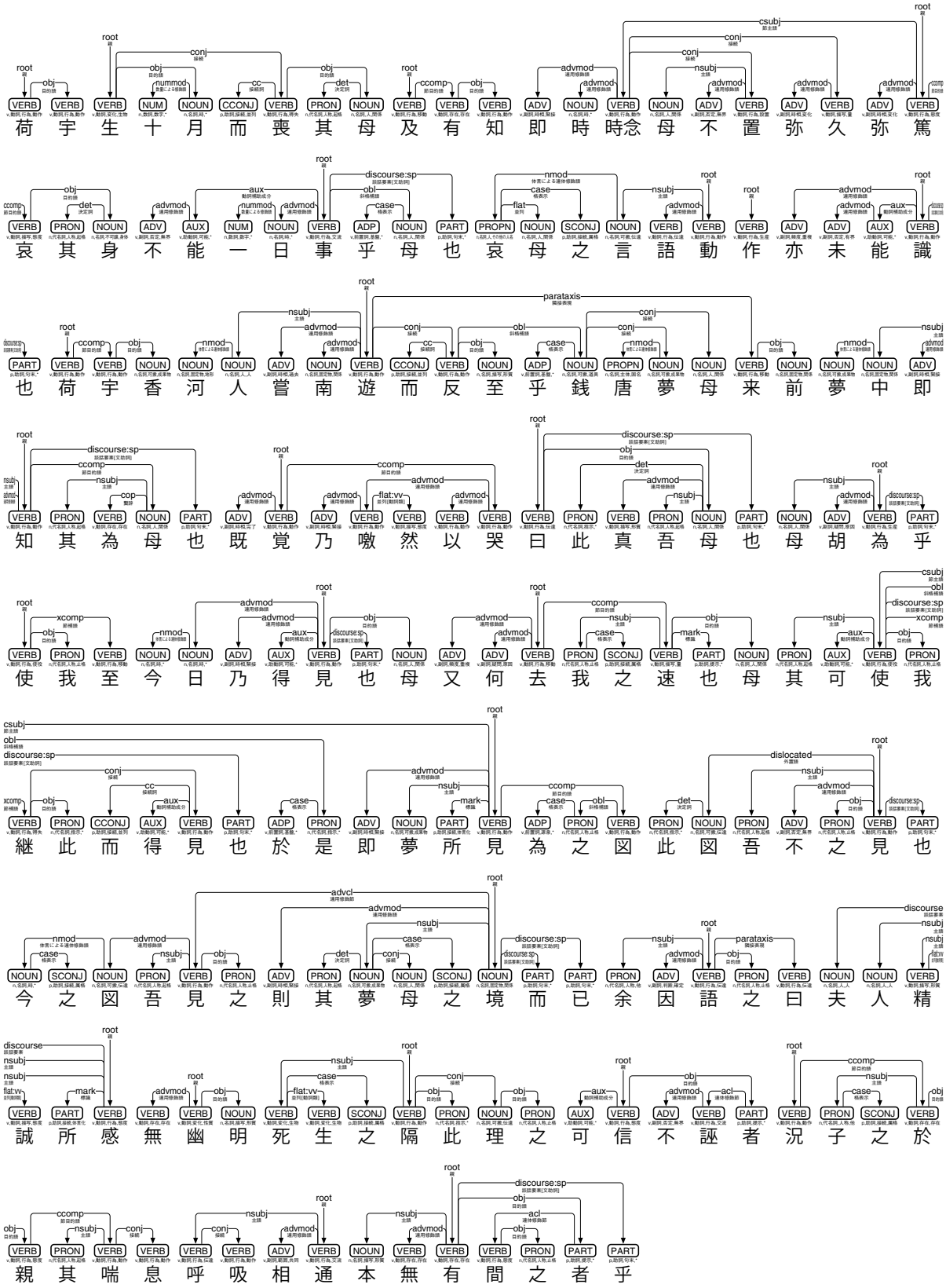


図 64: 手法①の処理結果 (2016 年)

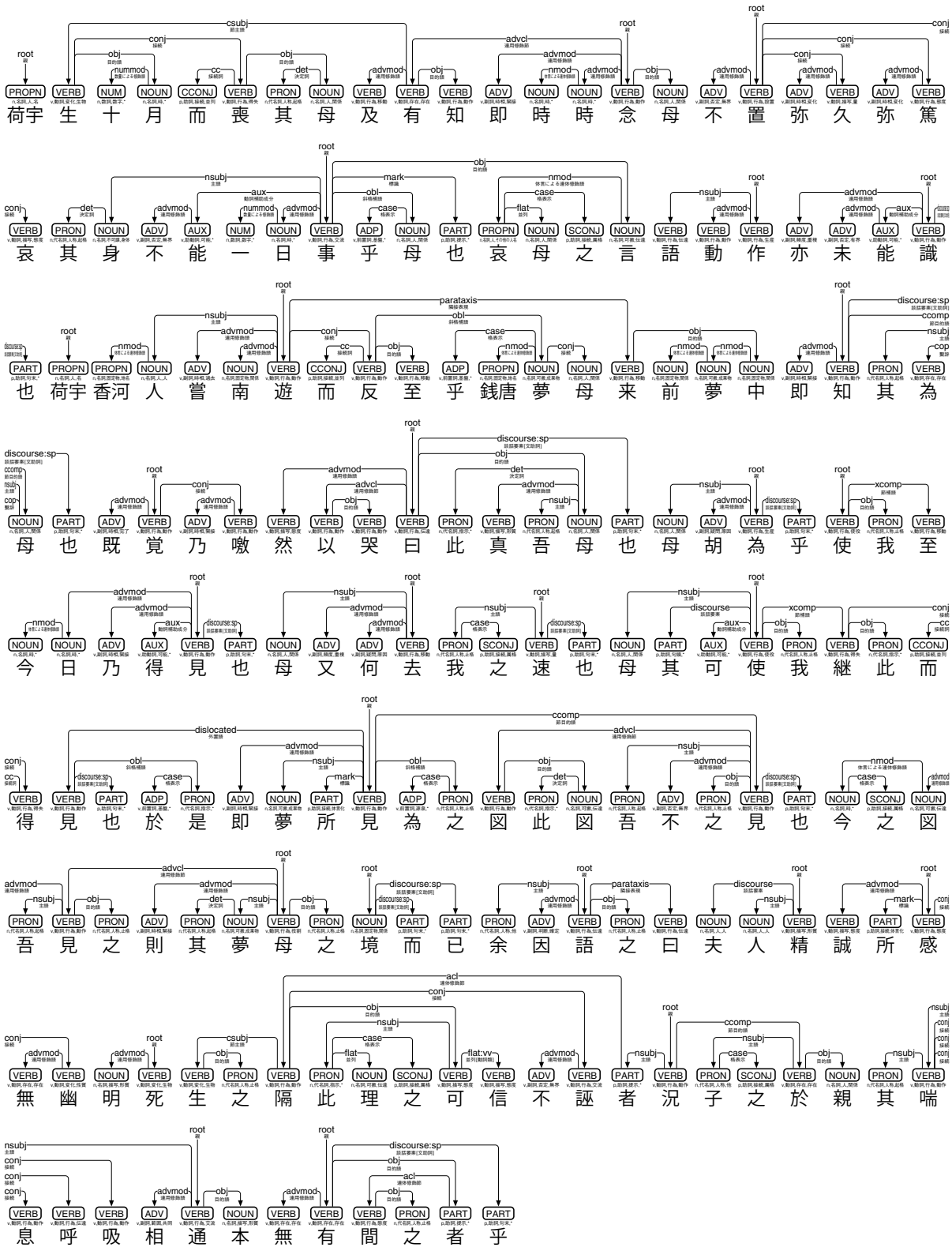


図 65: 手法②の処理結果 (2016 年)

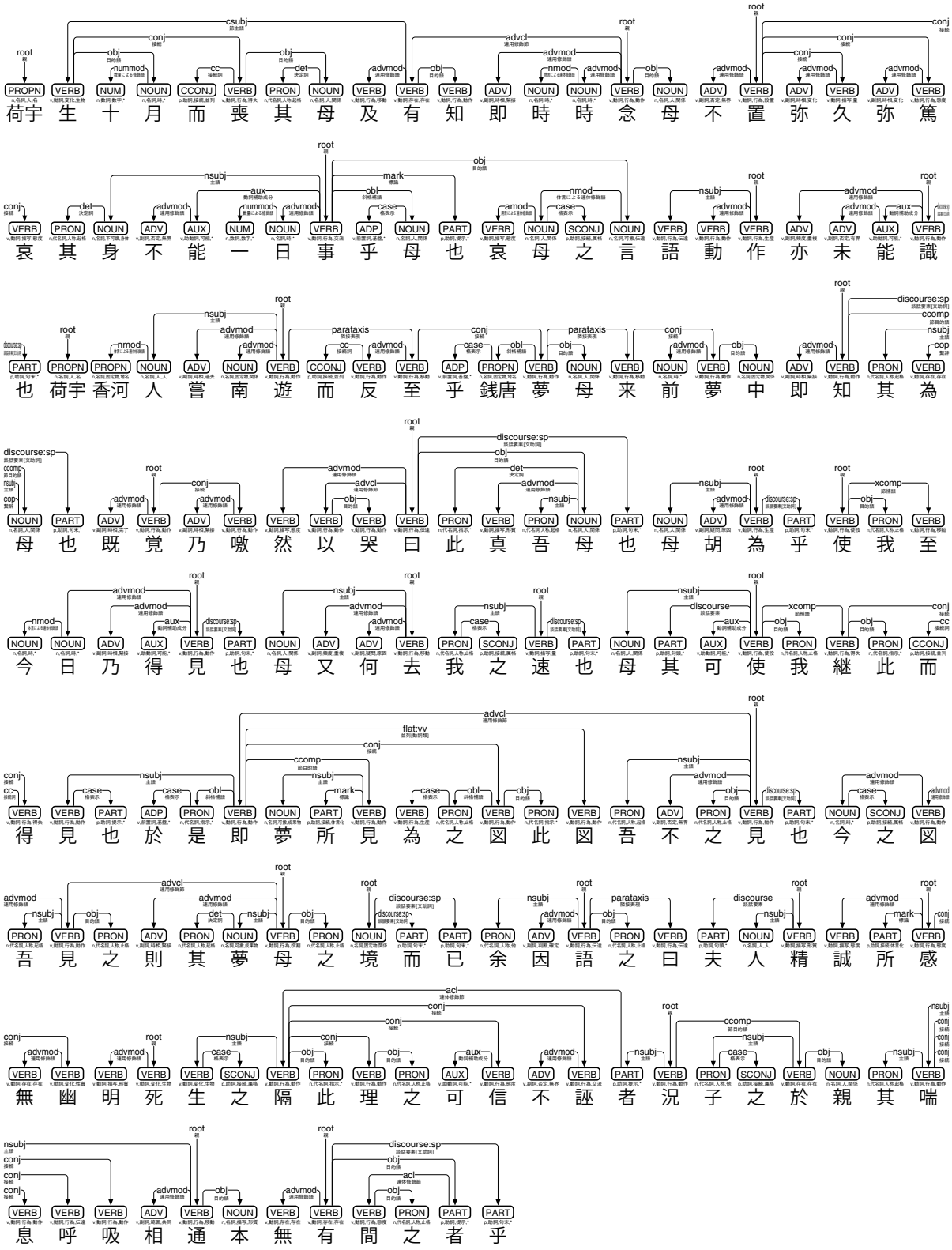


図 66: 手法③の処理結果 (2016 年)

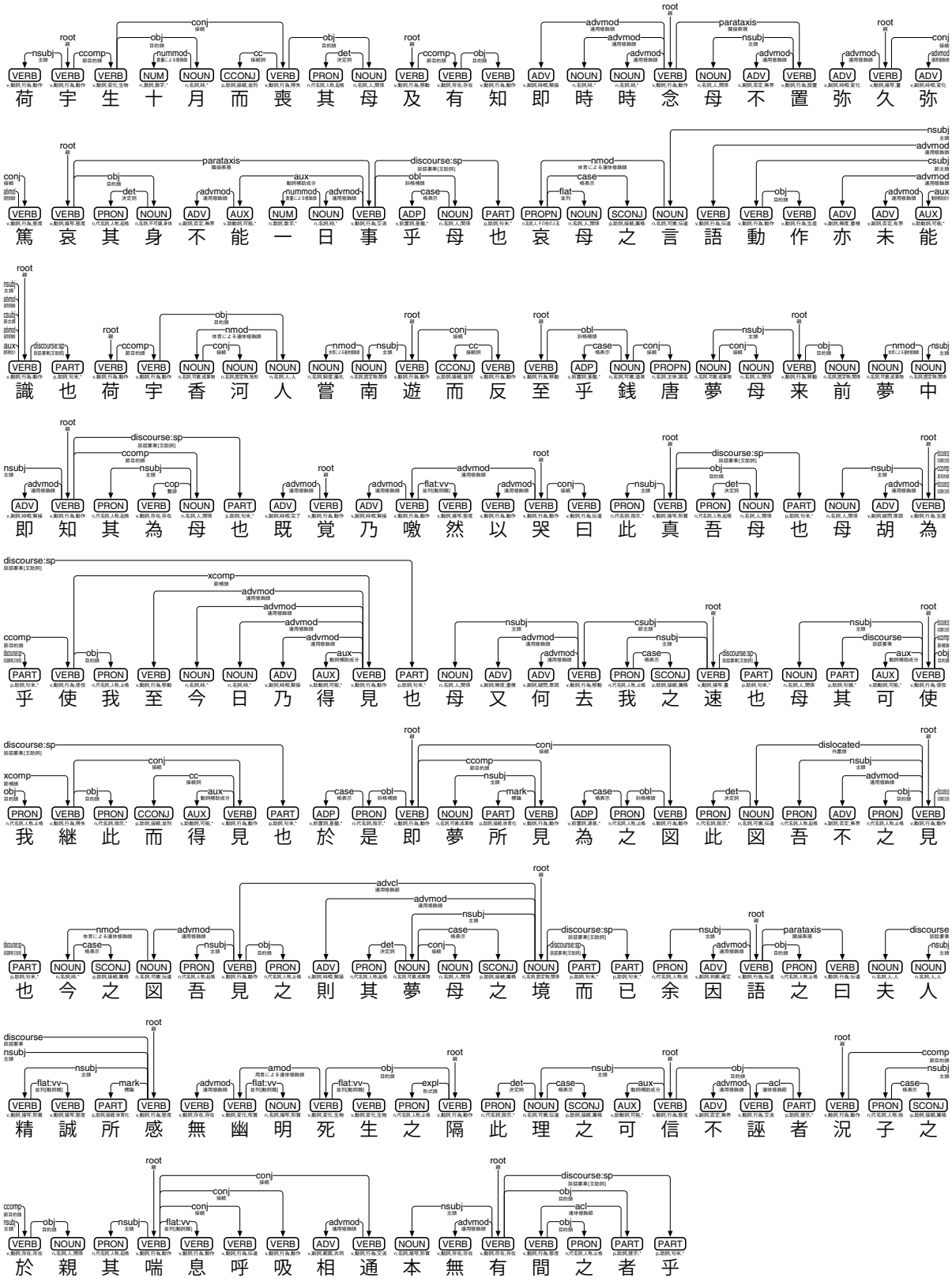
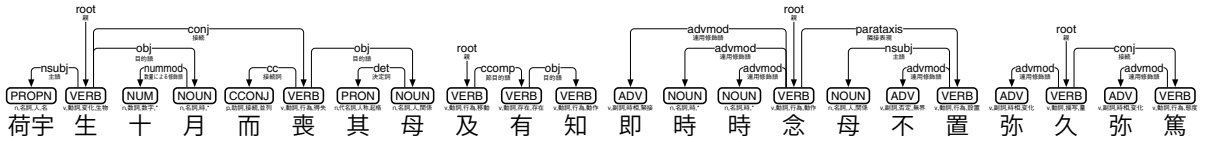
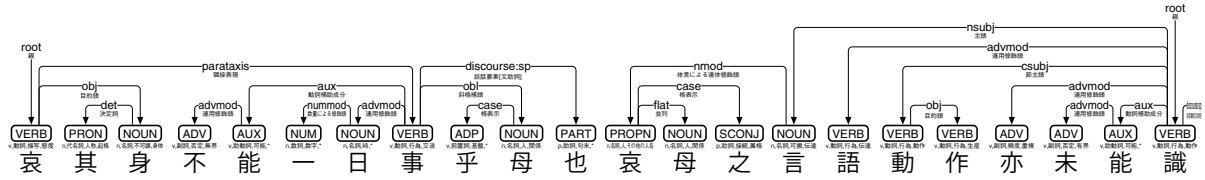
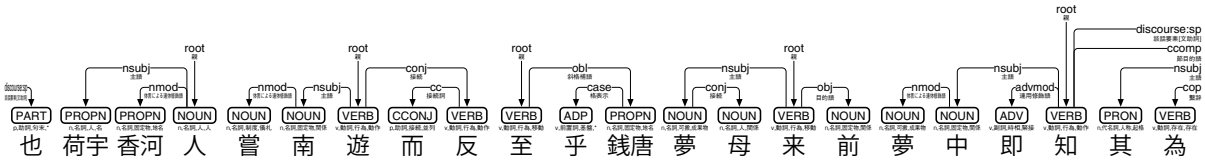
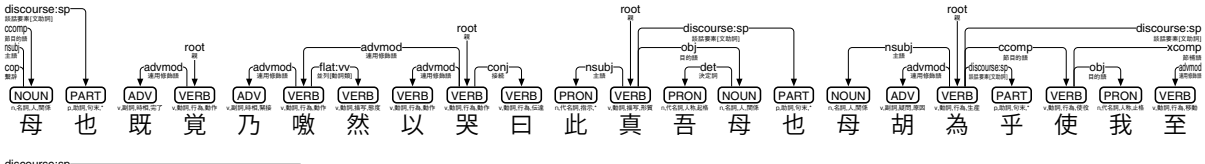


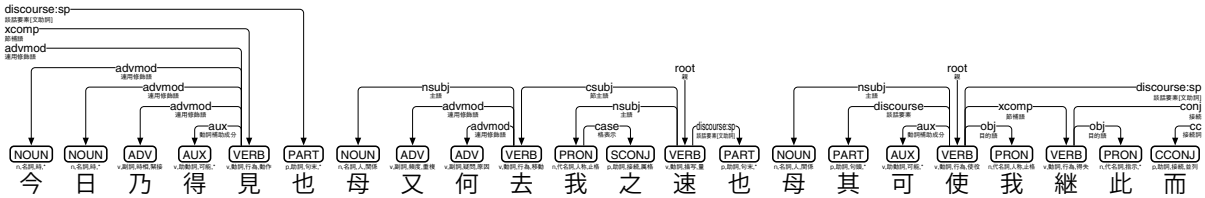
図 67: 手法1の処理結果 (2016 年)

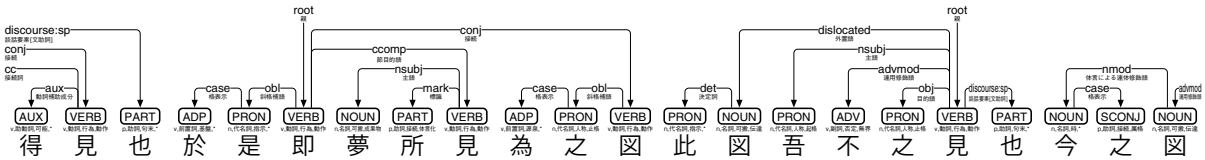


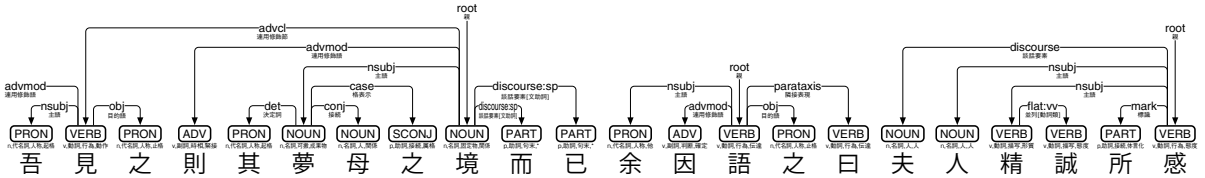


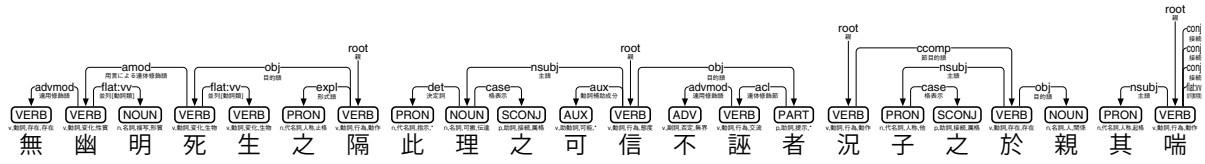












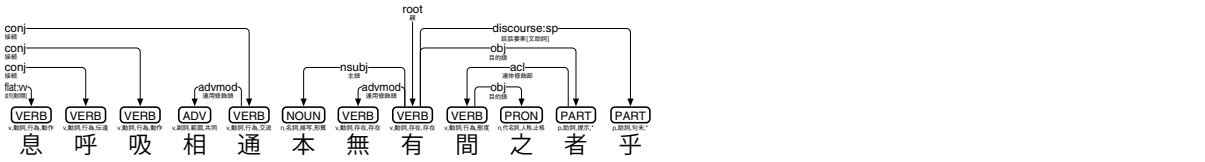


図 68: 手法2の処理結果 (2016 年)

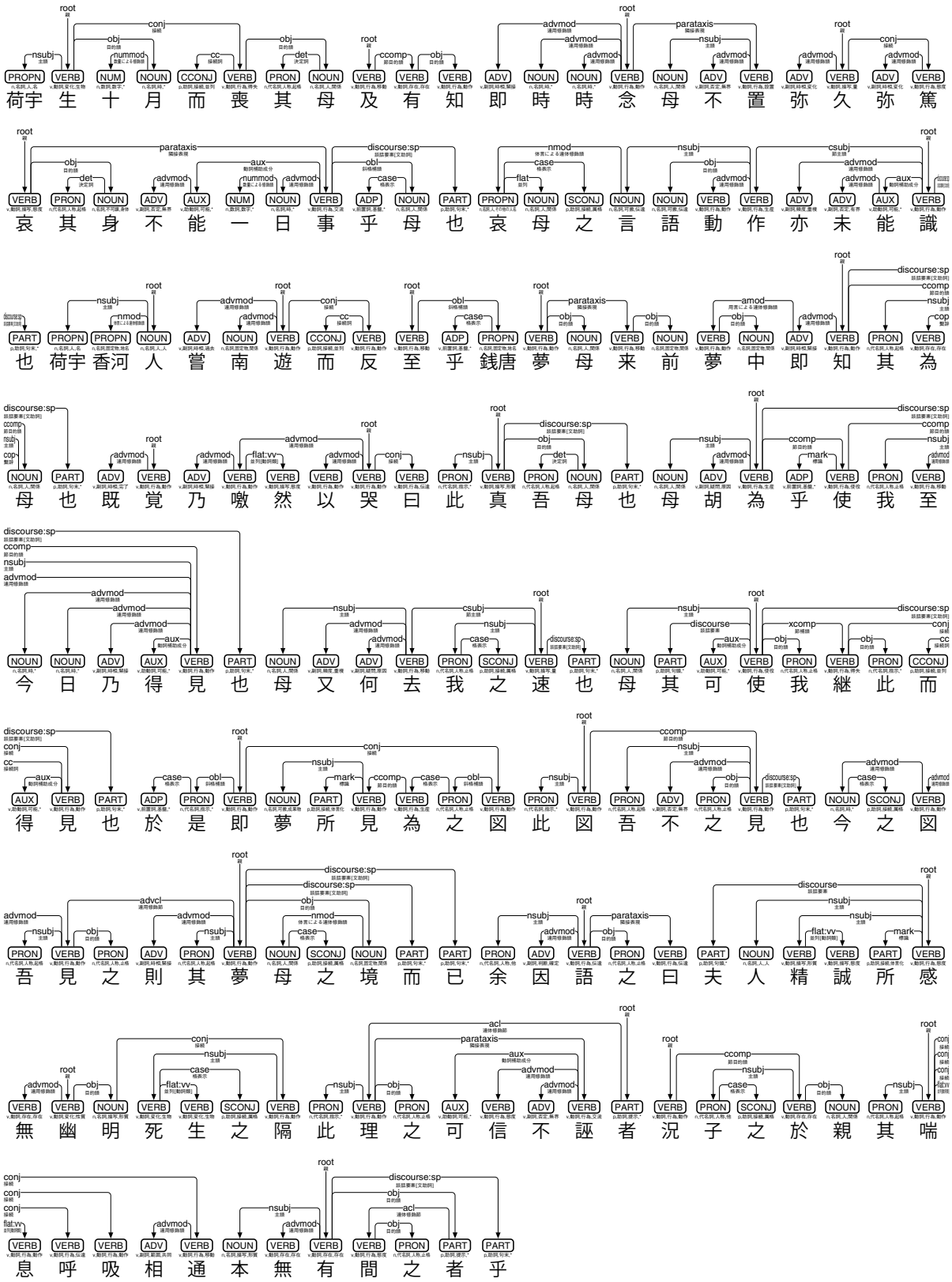


図 69: 手法3の処理結果 (2016 年)

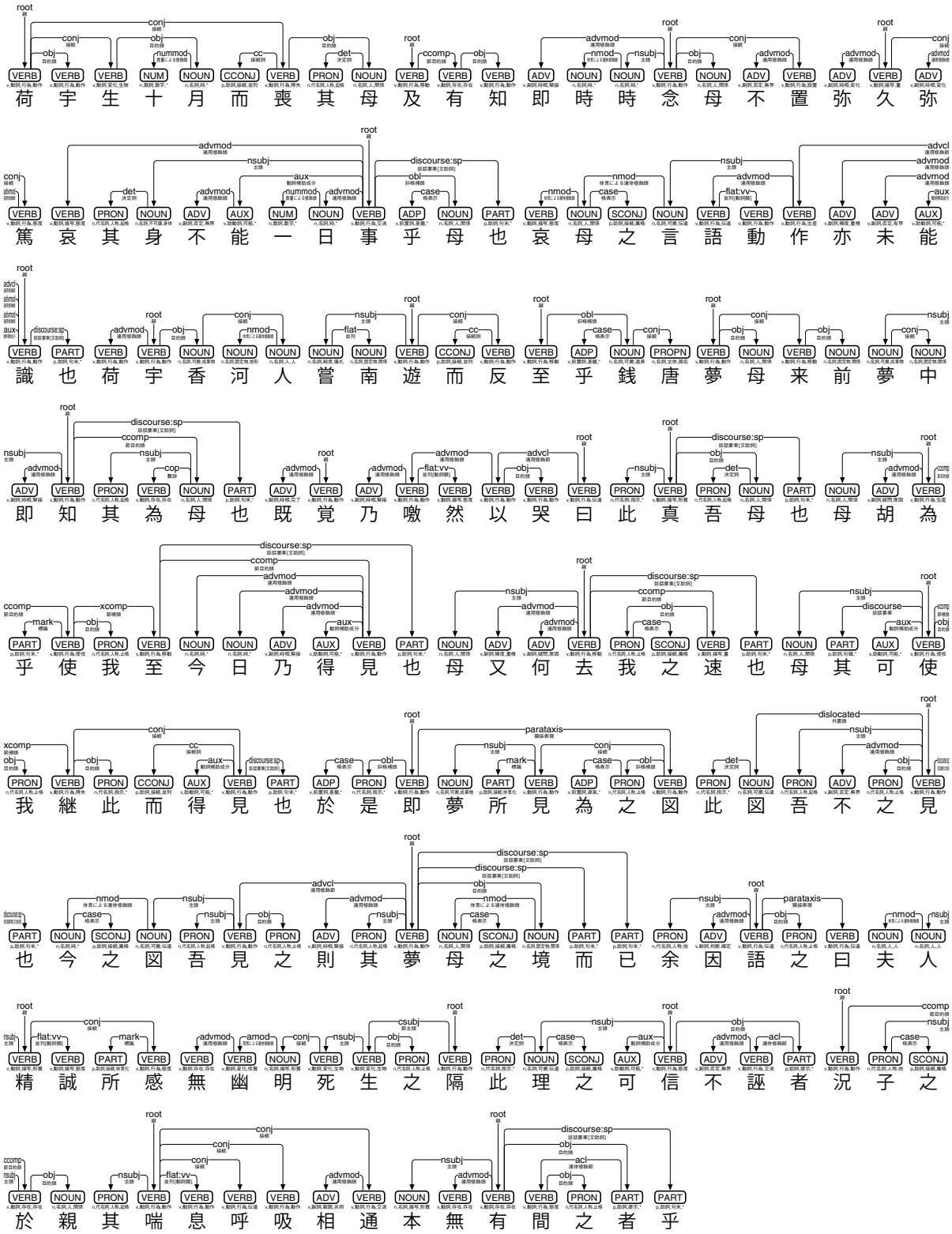


図 70: 手法❶の処理結果 (2016 年)

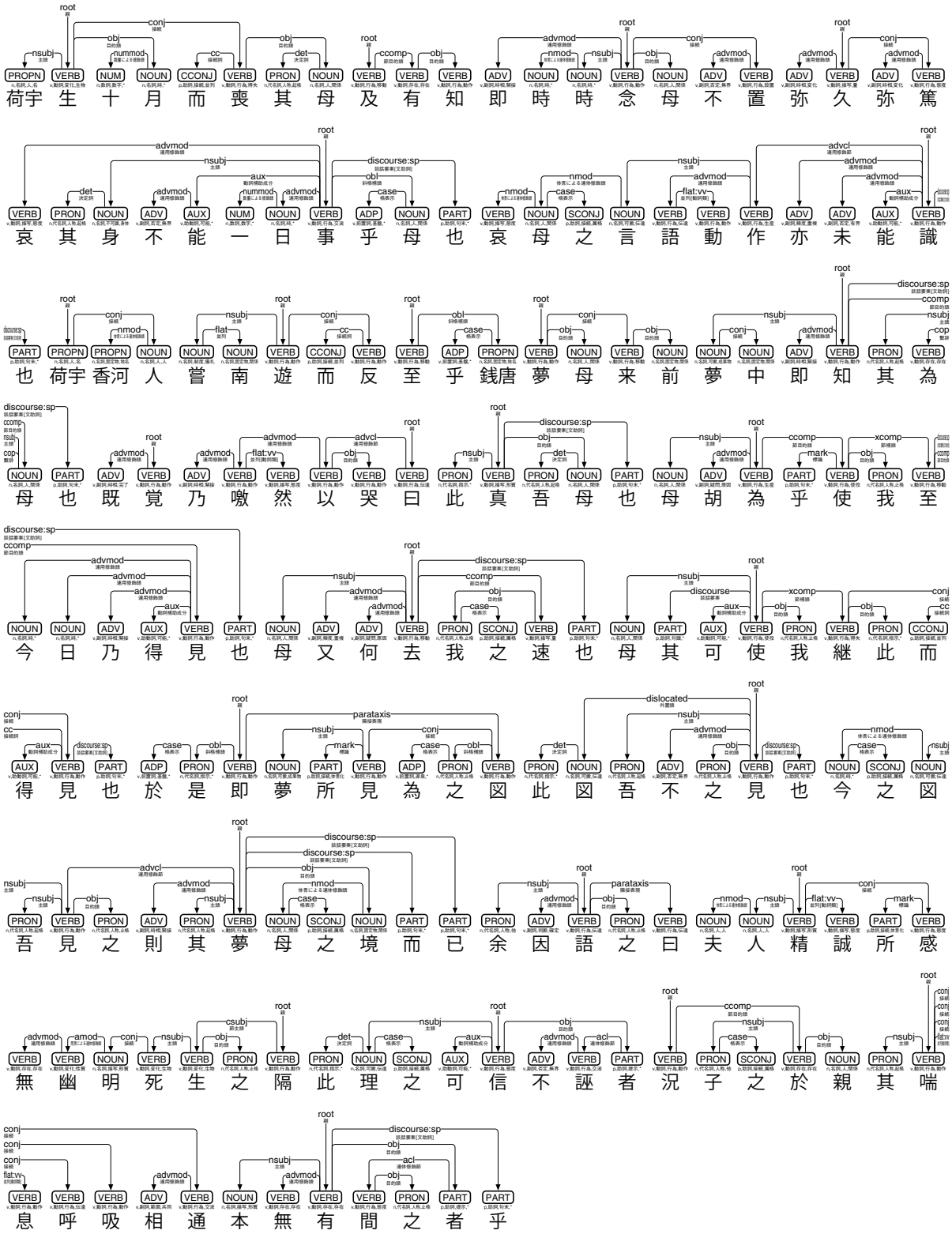


図 71: 手法❷の処理結果 (2016 年)

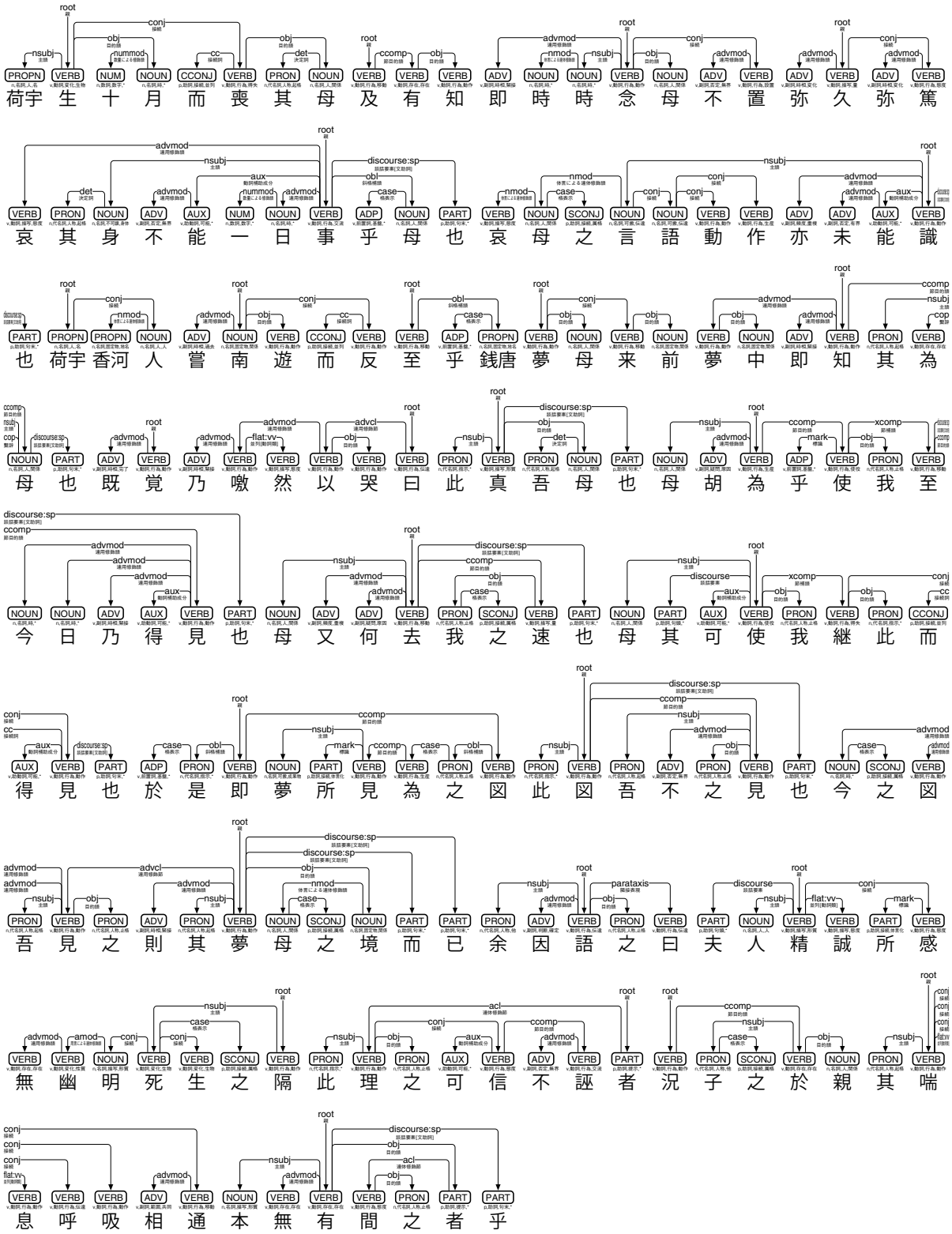


図 72: 手法❸の処理結果 (2016 年)

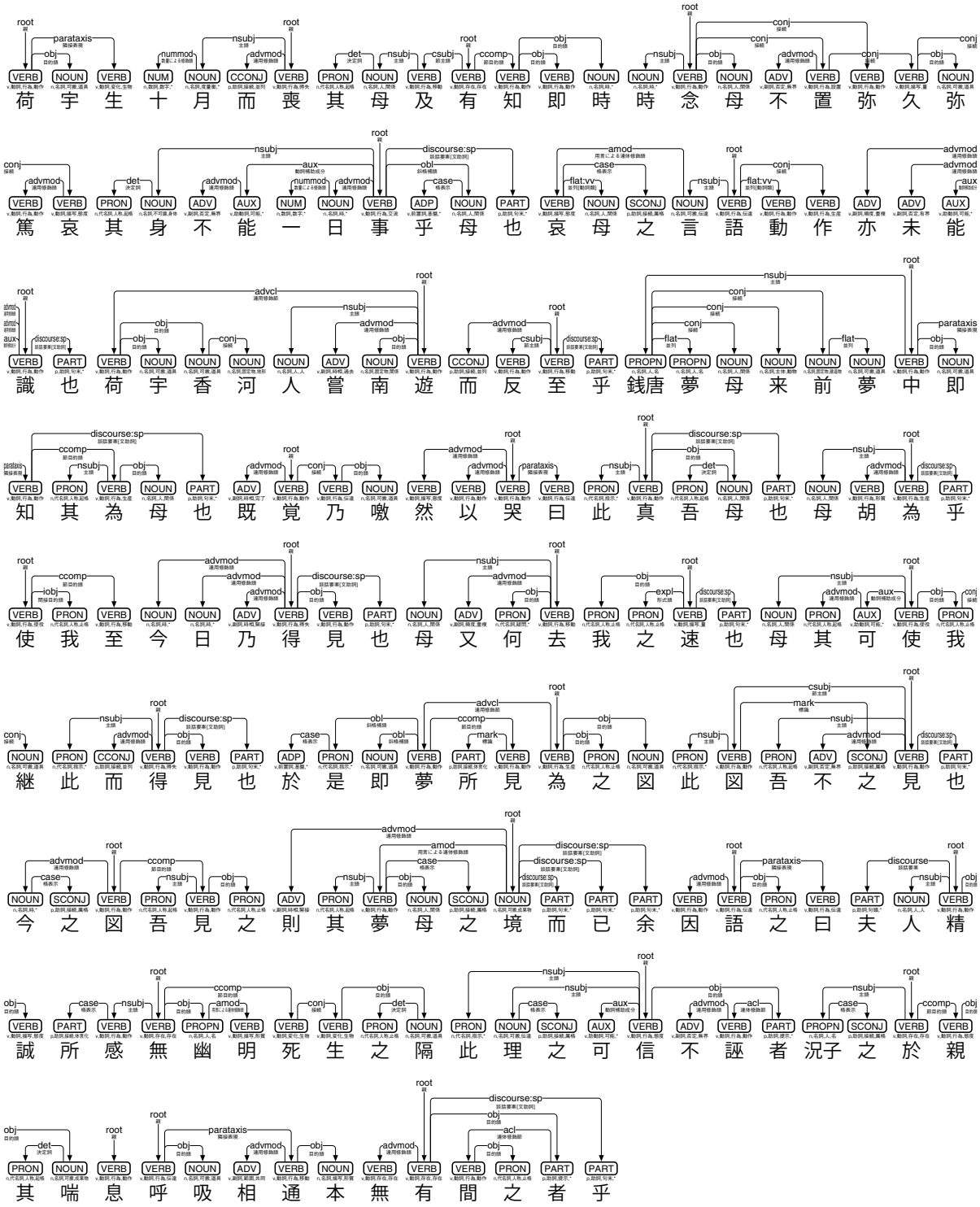


図 73: 手法Ⓐの処理結果 (2016 年)

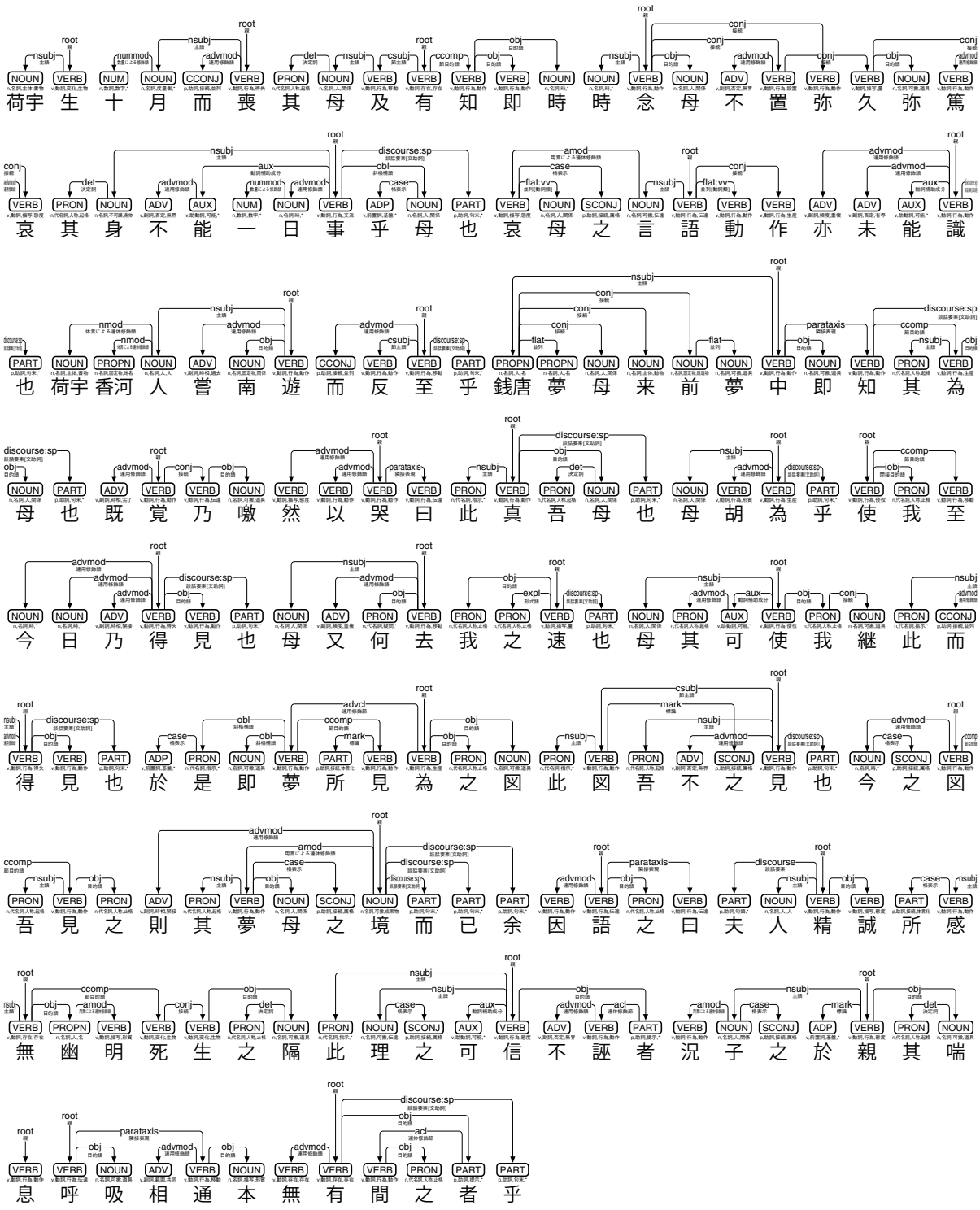


図 74: 手法Ⓑの処理結果 (2016 年)

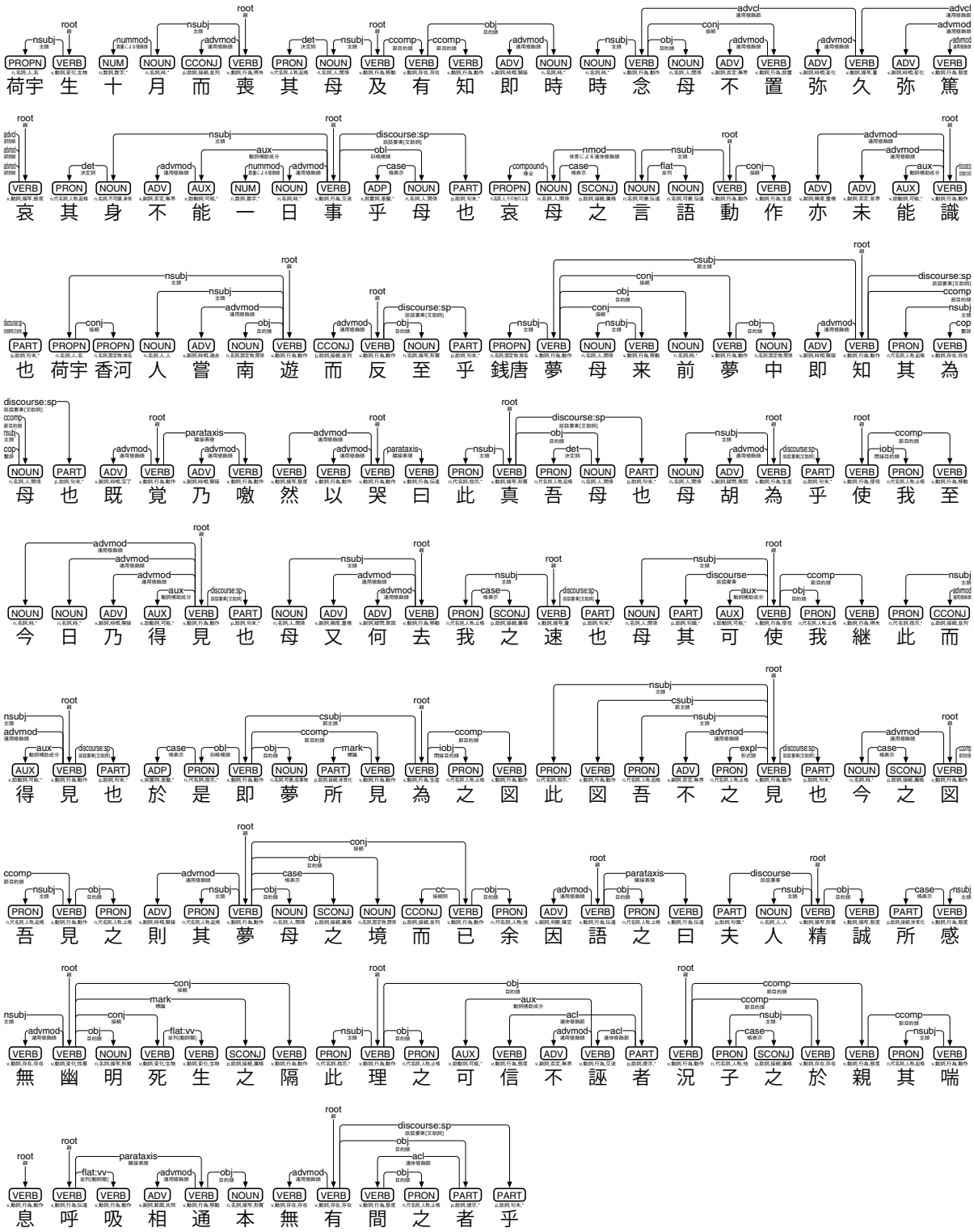


図 75: 手法Ⓒの処理結果 (2016 年)

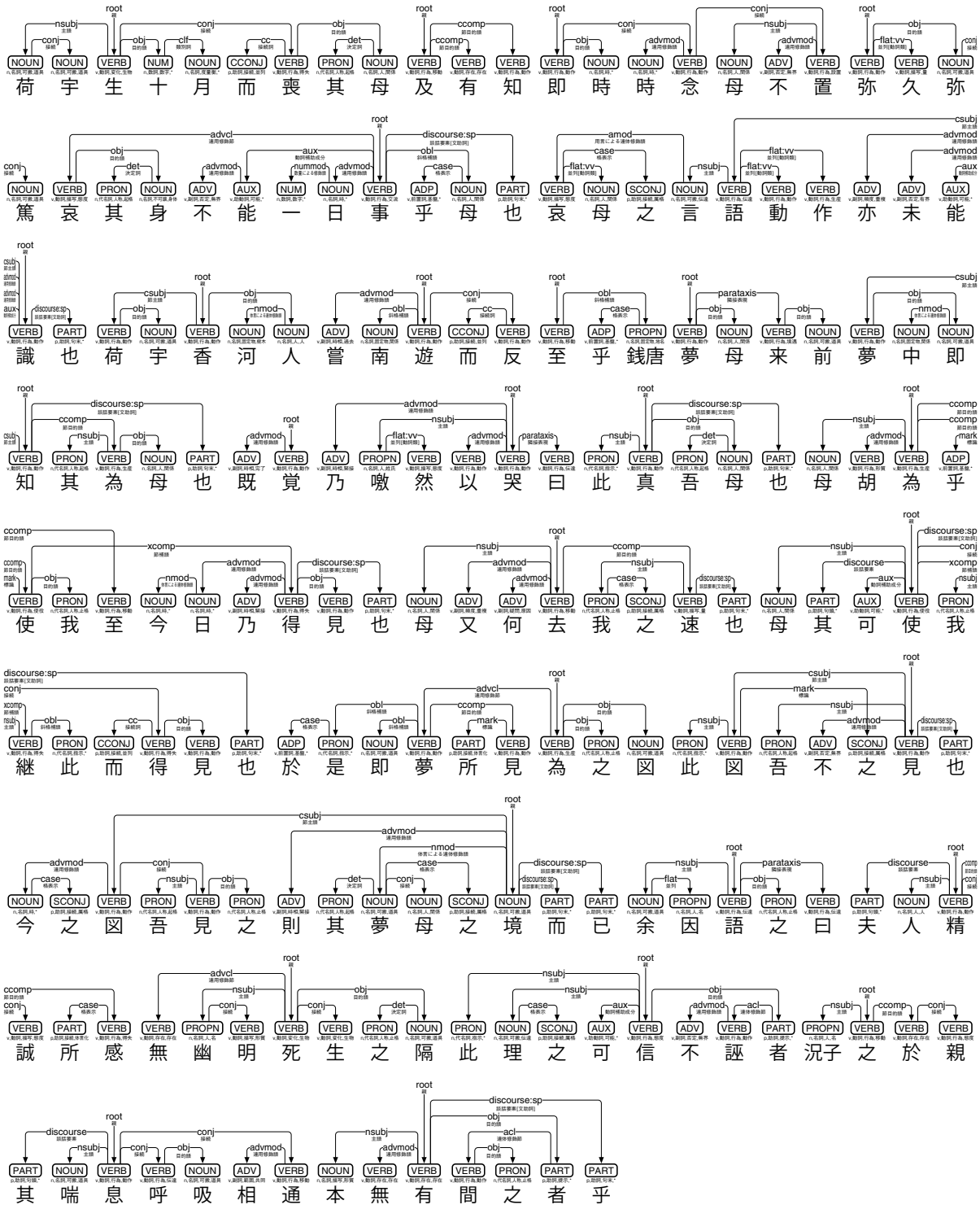


図 76: 手法Aの処理結果 (2016 年)

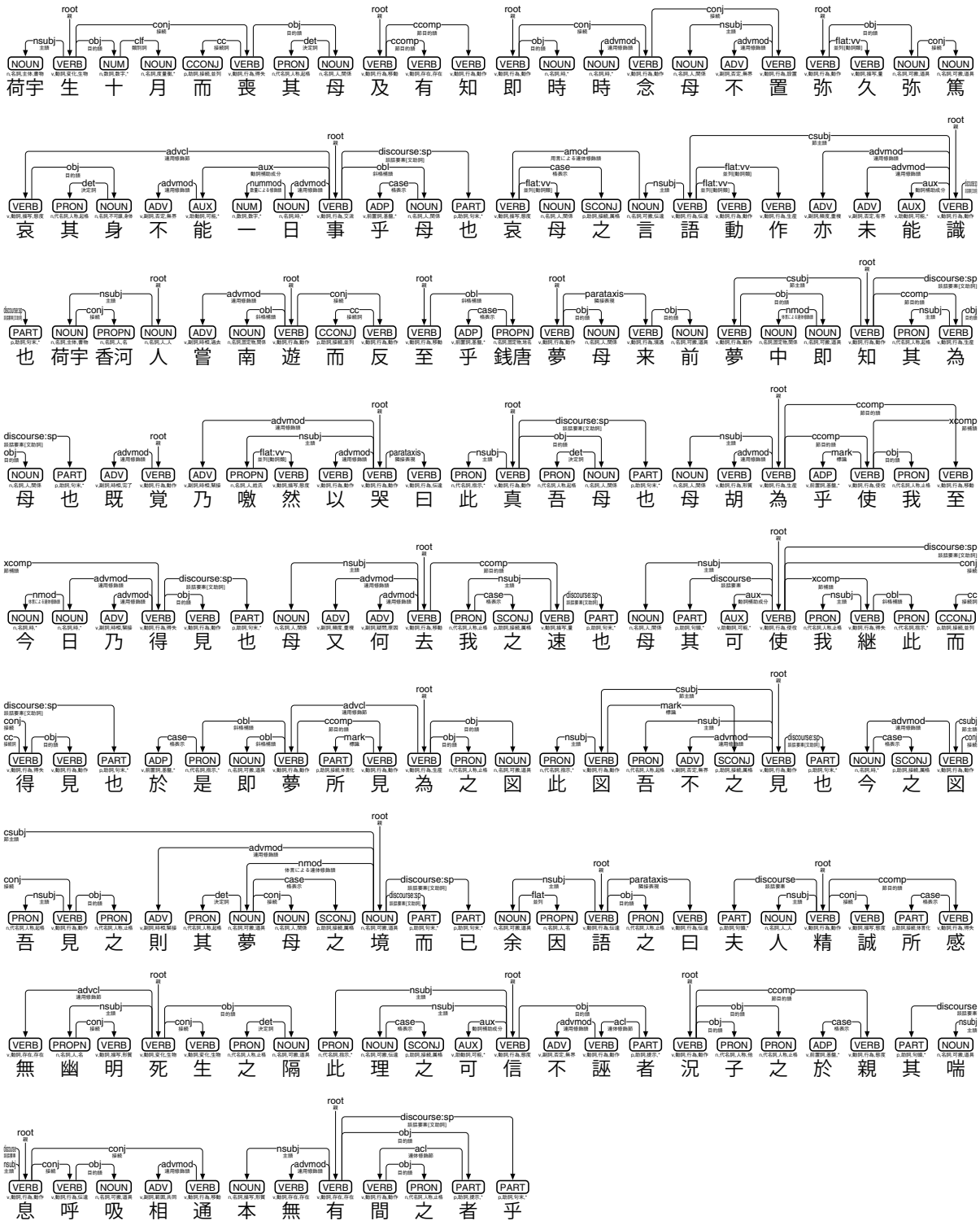


図 77: 手法Bの処理結果 (2016 年)

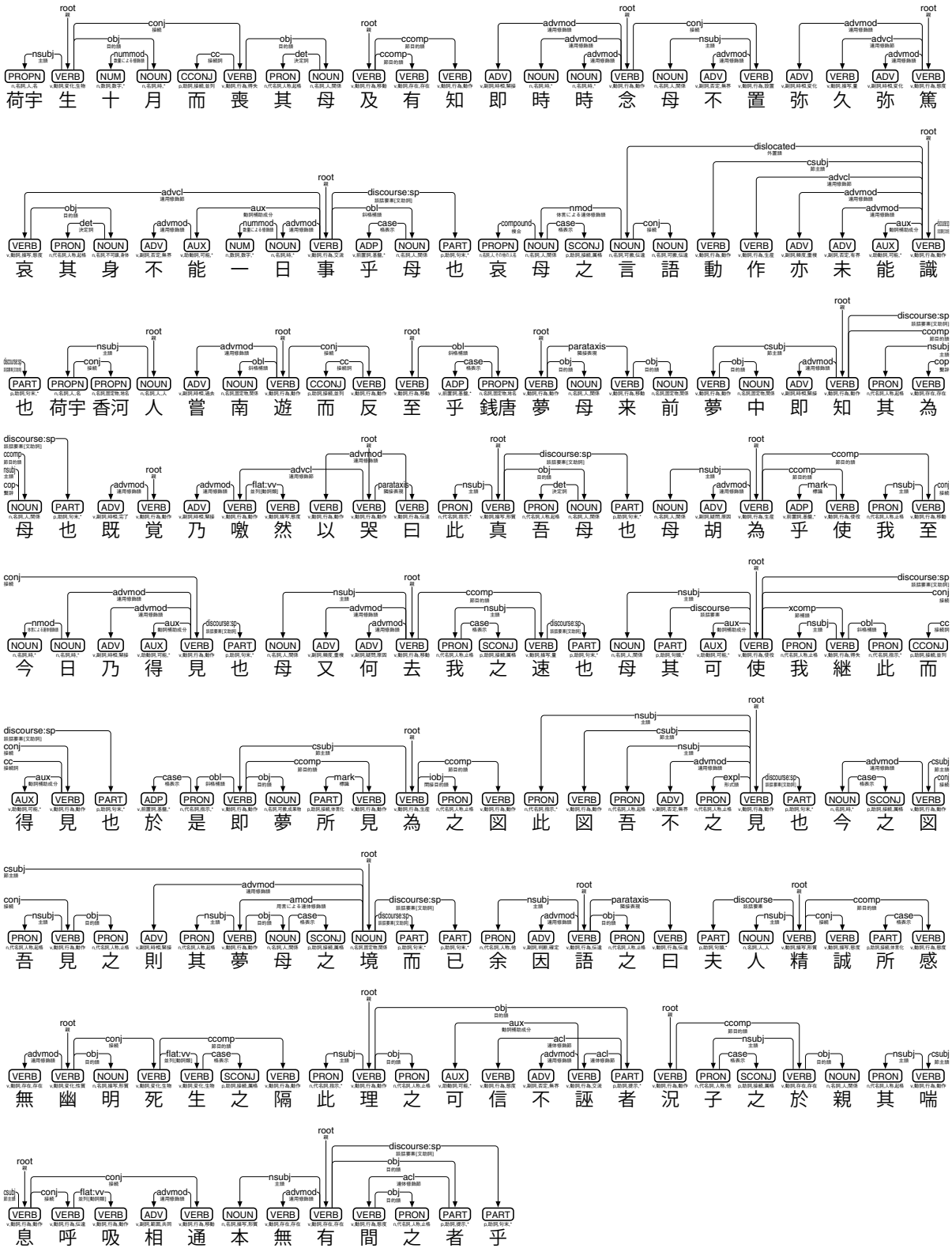


図 78: 手法Ⓒの処理結果 (2016 年)

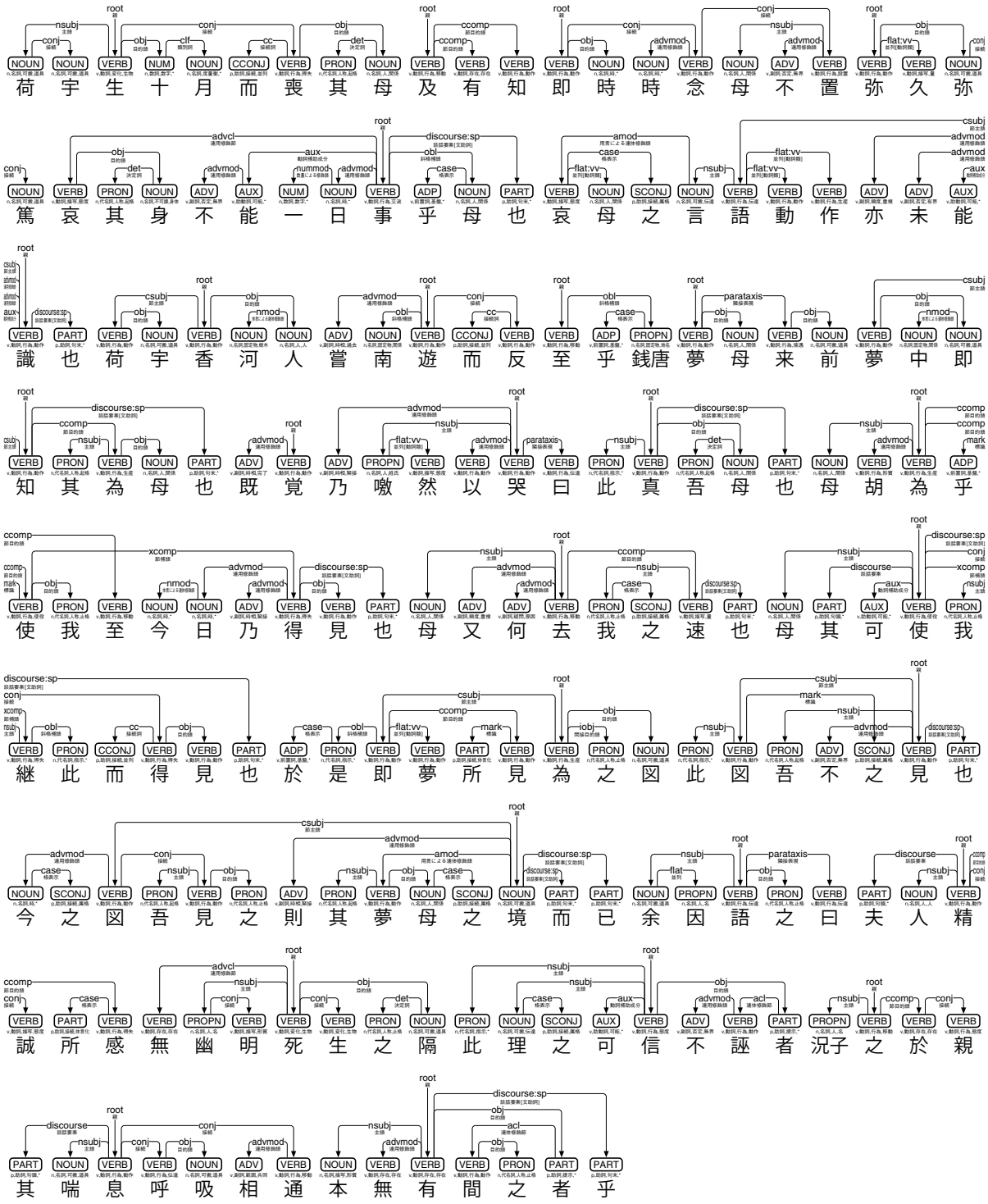


図 79: 手法Ⓐの処理結果 (2016 年)

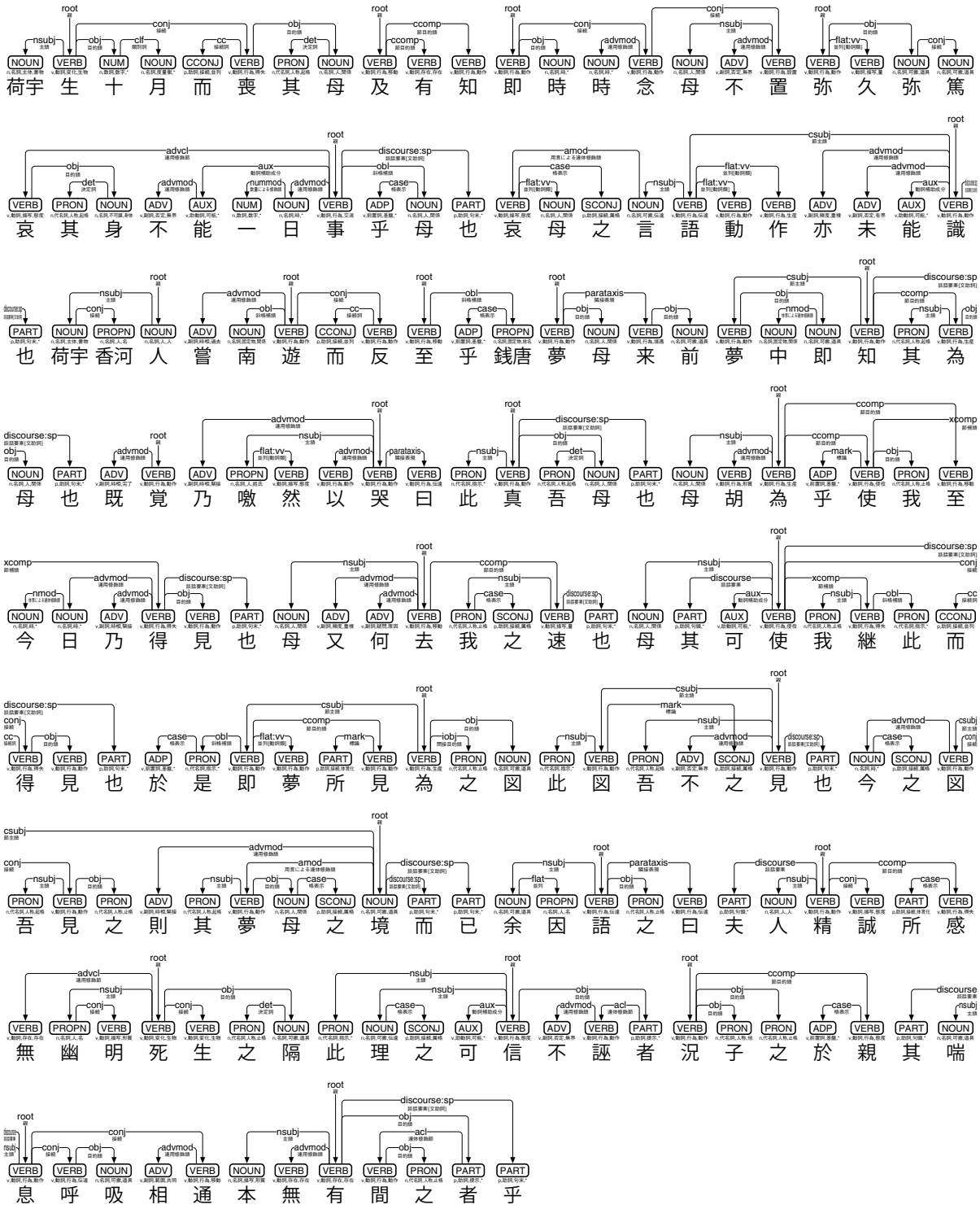


図 80: 手法Ⓑの処理結果 (2016 年)

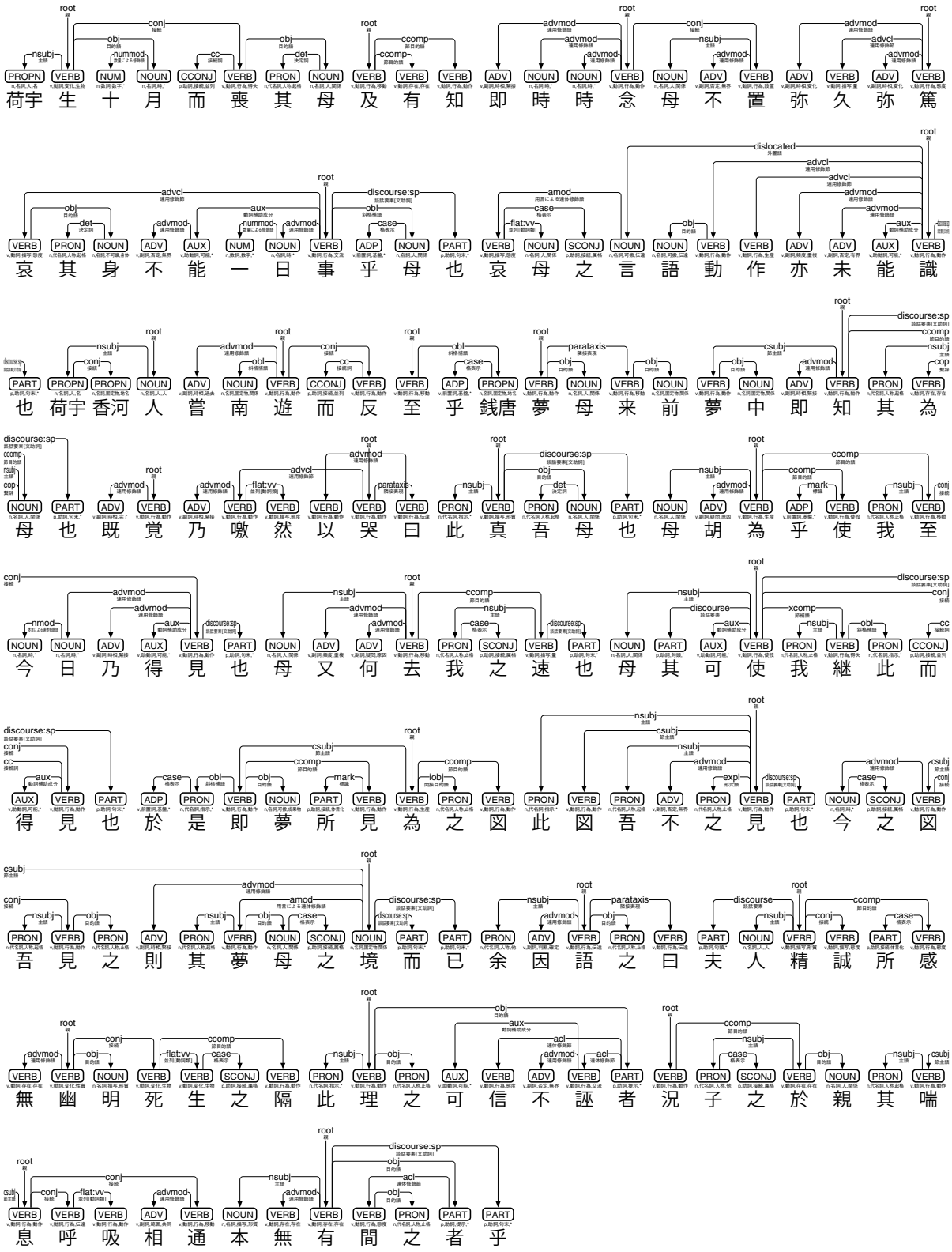


図 81: 手法Ⓒの処理結果 (2016 年)

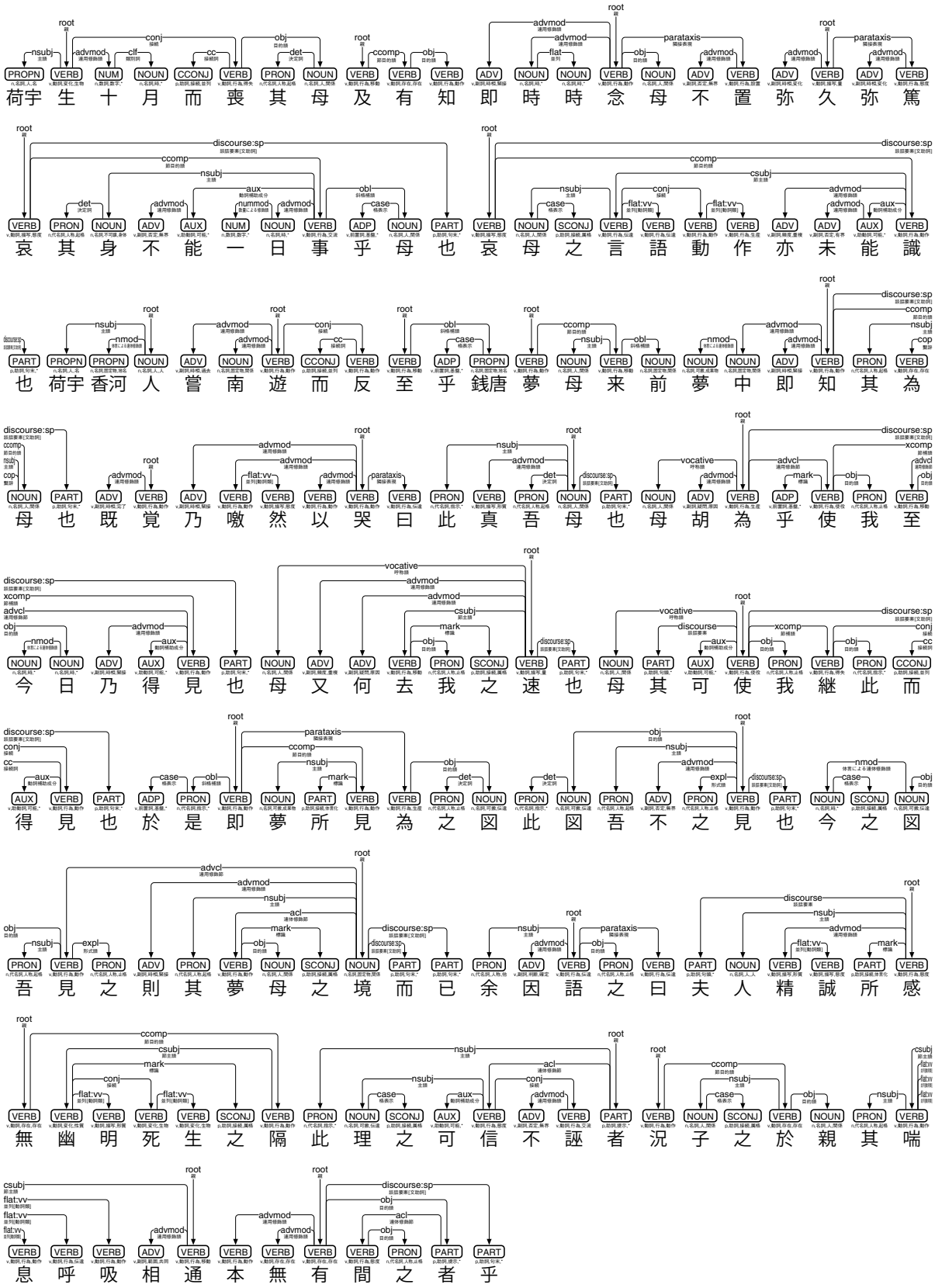


図 82: 手作業で作成した「正解」UD (2016 年)

家ニ蓄ヘシ一老狸奴(注1)ニ。(ア)将レ誕レ子ヲ(a)矣。一女童誤リテ触レ之ニ、而シテ堕ス日夕鳴ヲ(注2)然タリ。会タマタマ有リ下餽ニ両小狸奴ヲ一者。其ノ始メ、蓋シ漠然(注3)トシテ不ル相ヒ能ヒ一也。老狸奴ナル者、従ヒテ而撫ブレ之ヲ、傍ハウ(注4)徨クワウ焉タリ、躑テキ躅チヨク焉タリ。臥スレバ則チ擁シレ之ヲ、行ケバ則チ翊タスクレ之ヲ。舐メテ二其ノ氄ジヨウ(注5)一而譲ルニ之ニ食ヲ一。両小狸奴ナル者、亦マタ久シクシテ而相ヒ忘ルル也。稍ヤウヤク即ツキレ之ニ、遂ニ(1)承ニ其ノ乳ヲ一焉。(イ)自レ是レ欣キン(注6)然ゼントシテ以テ為スニ良マコトニ己之母ナリト。老狸奴ナル者、亦タ居(注7)然トシテ以テ為スニ良ニ己ガ出ダスト(b)一也。A吁ああ、亦タ異ナル哉カナ。

昔、漢ノ明(注8)徳馬后ニ無シレ子。顕(注9)宗取リニ他ノ人ジン子シヲ、命ジテ養ハシメテレ之ヲ曰ハク、B「人子何必ズ親生。但ダ恨ムニ愛之不ルレ至ラ(c)ト耳。」后遂ニ尽クシテレ心ヲ撫シ育シ、而シテ章帝モ亦タ恩(注10)性天至タリ。母子ノ慈孝、始終無シ(注11)ニ纖セン芥カイ之間一。狸奴之事、(2)適有リレ契カナフ(d)焉。然シカラバ則チ世之為ニ人親与レ子、而有ニ不慈不孝者一、豈独愧ニ于古人一。亦タ愧ハヅルニ此ノ異類ニ(e)一已。C

図 83: 大学入試センター試験『国語』(2015 年 1 月 17 日) 第 4 問本文

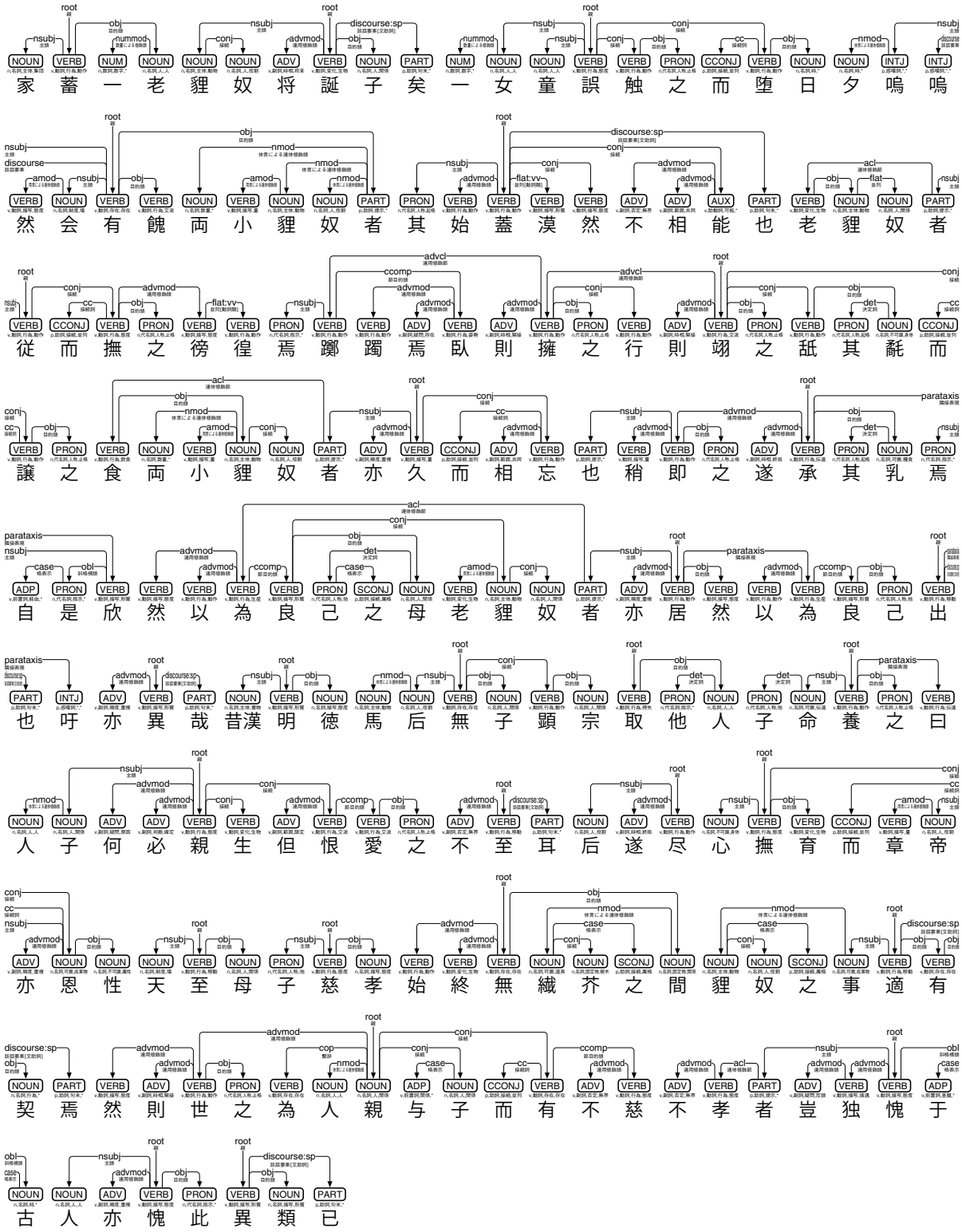


図 84: 手法①の処理結果 (2015 年)

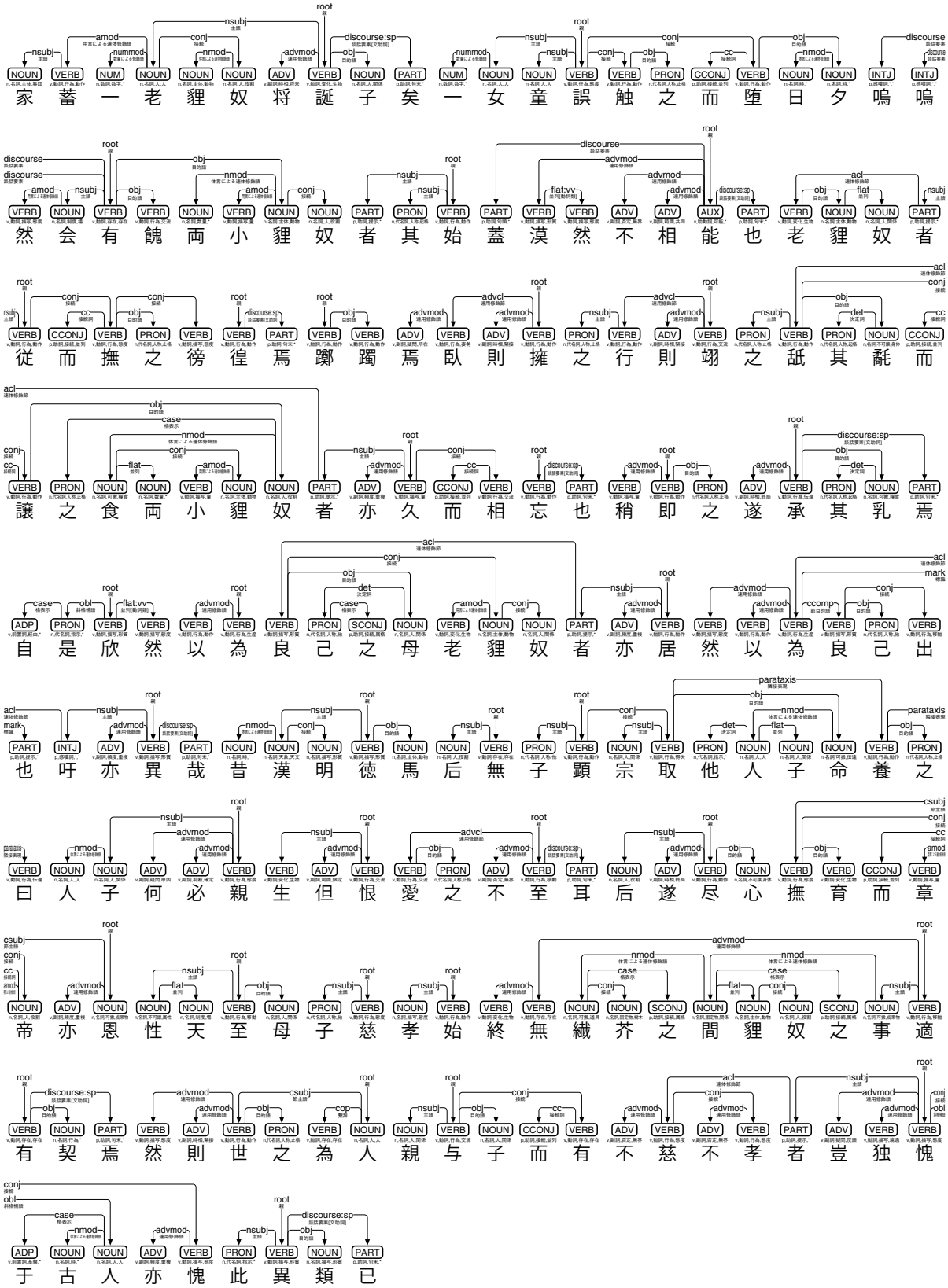


図 85: 手法②の処理結果 (2015 年)

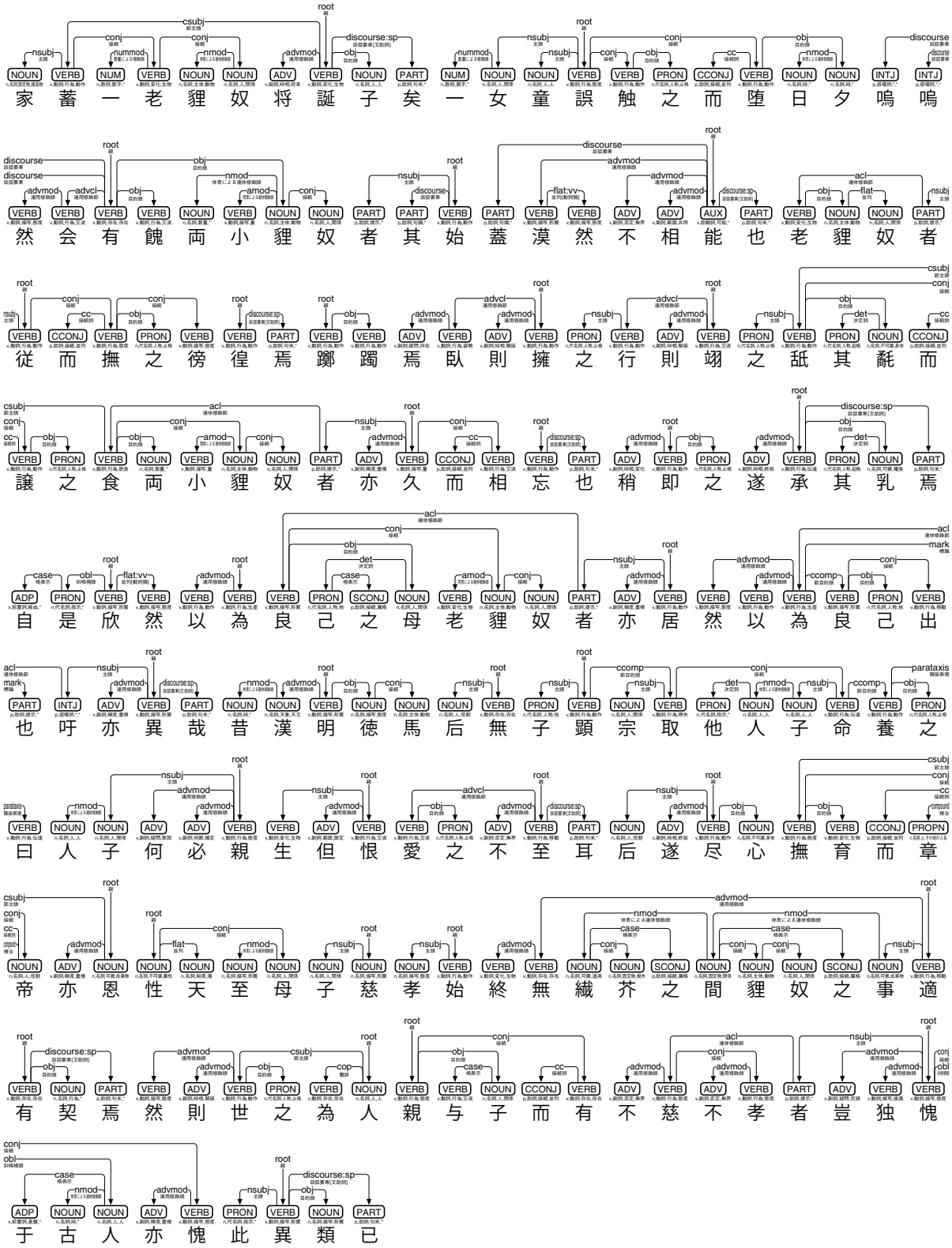


図 86: 手法③の処理結果 (2015 年)

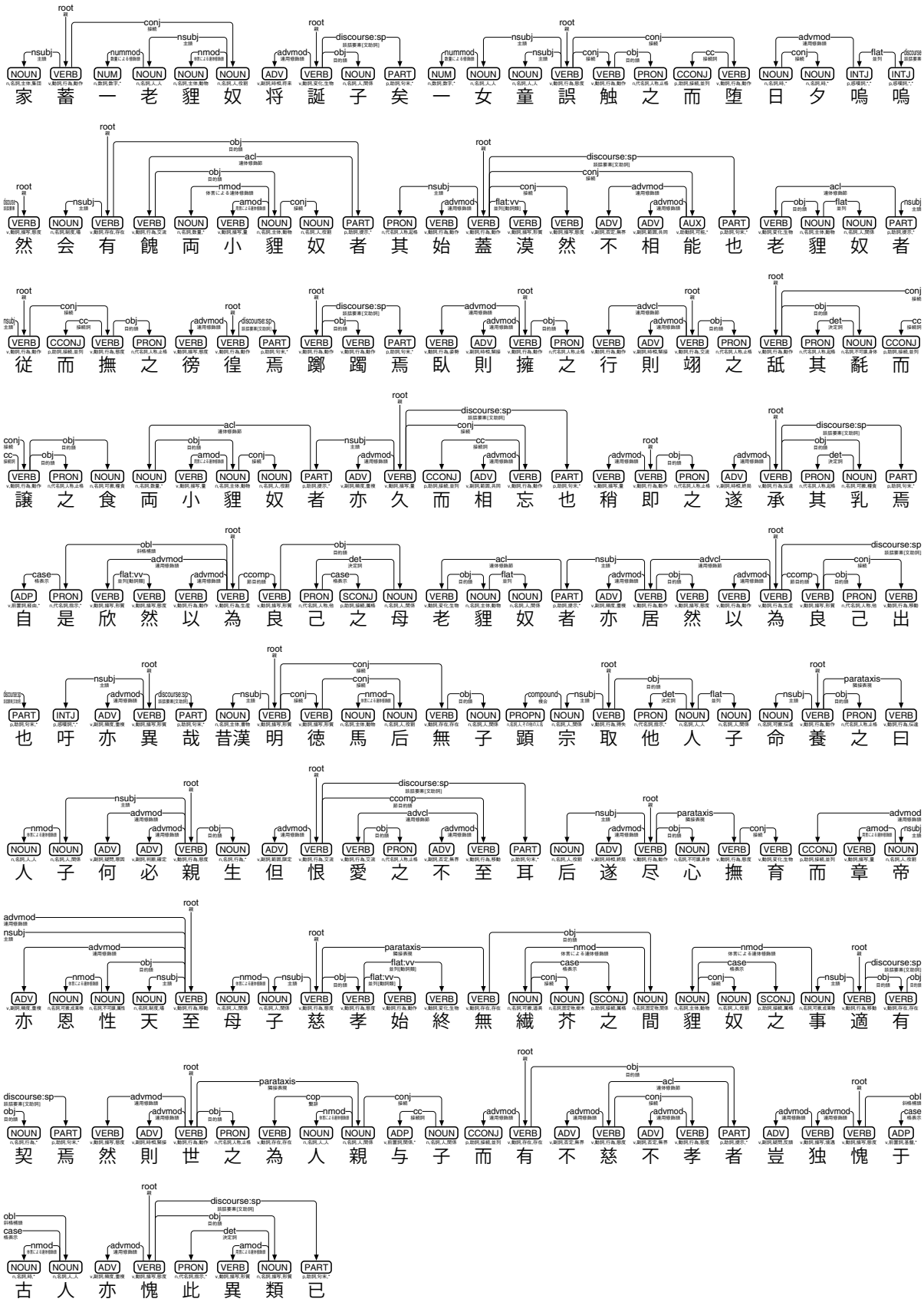


図 87: 手法1の処理結果 (2015 年)

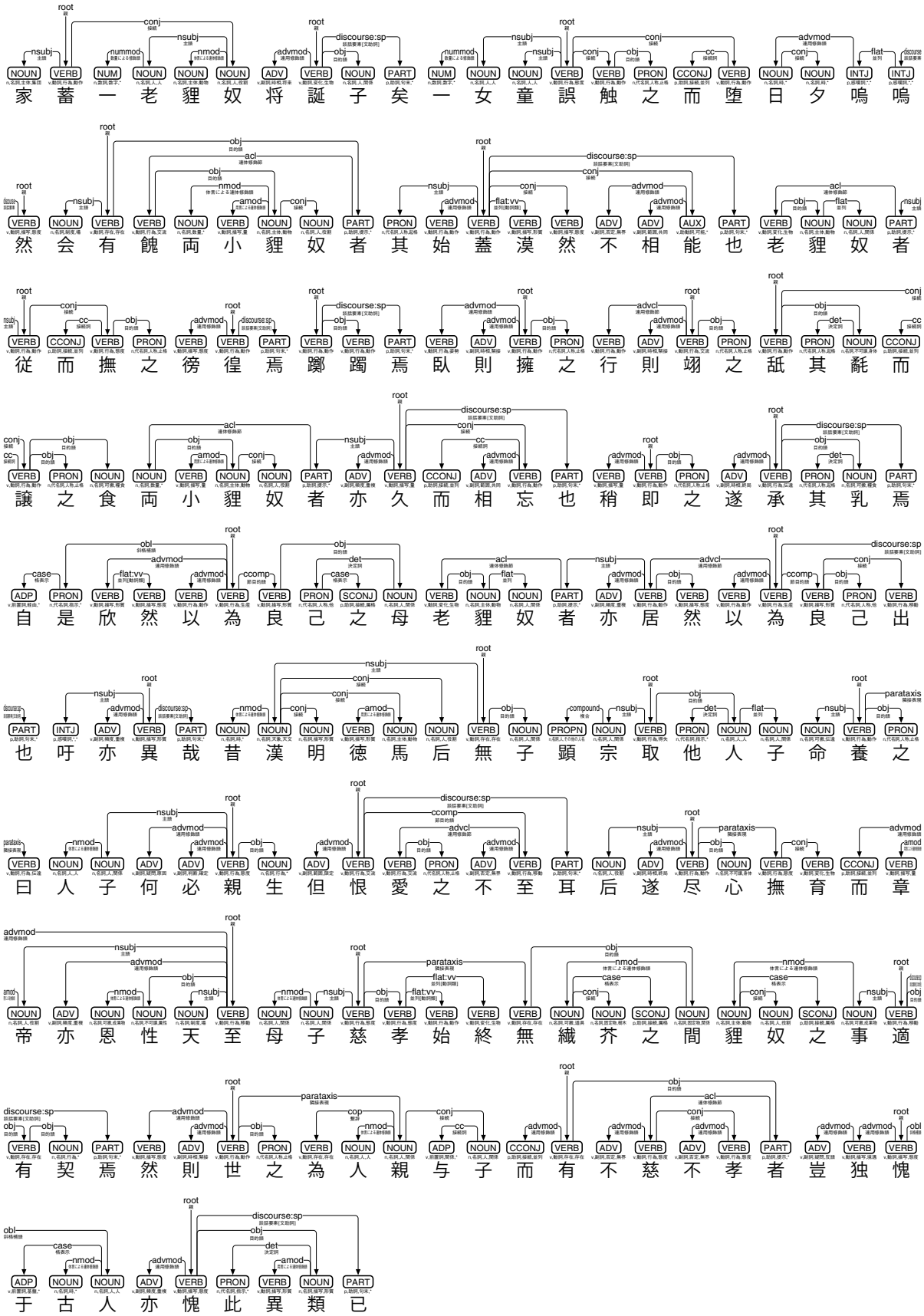


図 88: 手法2の処理結果 (2015 年)

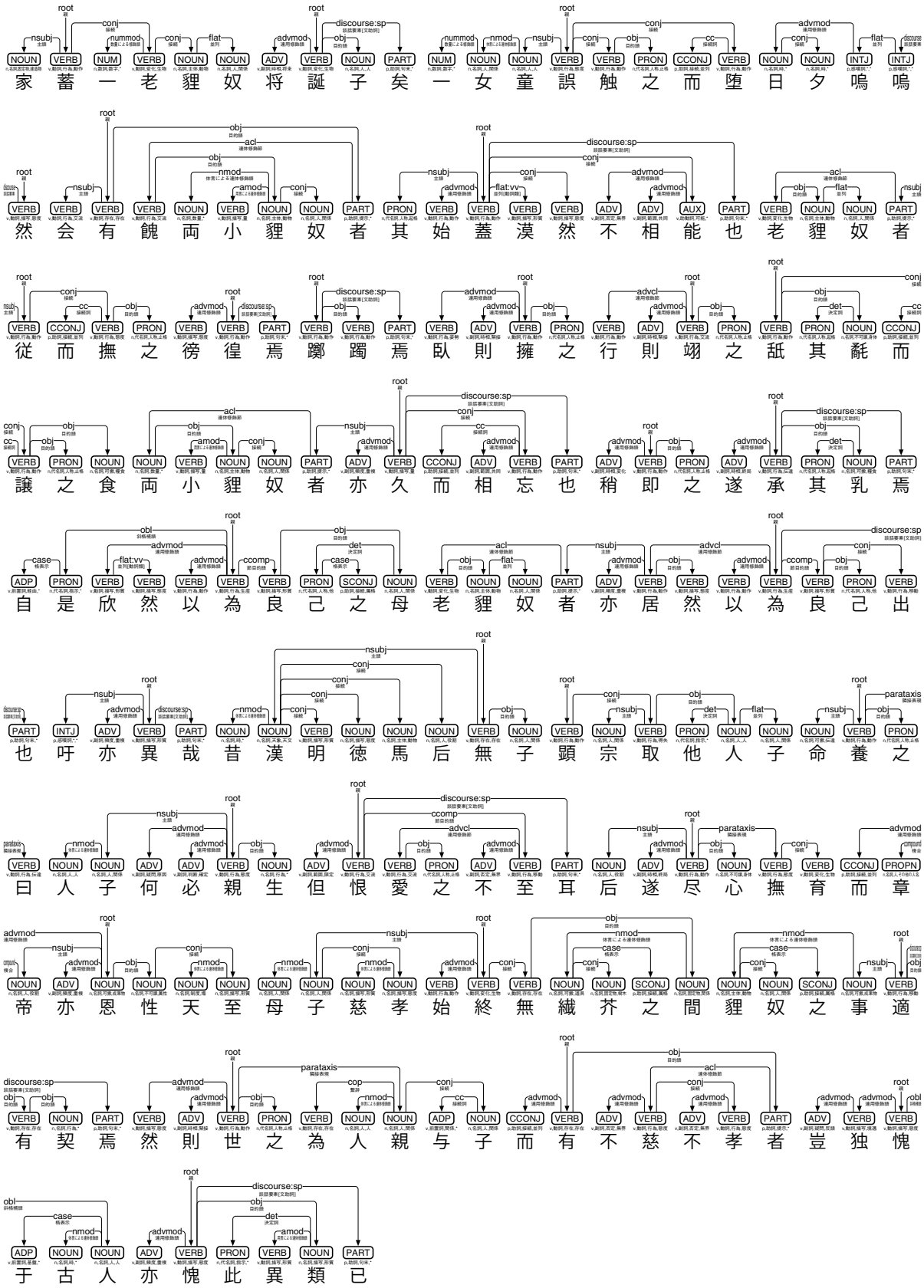


図 89: 手法3の処理結果 (2015 年)

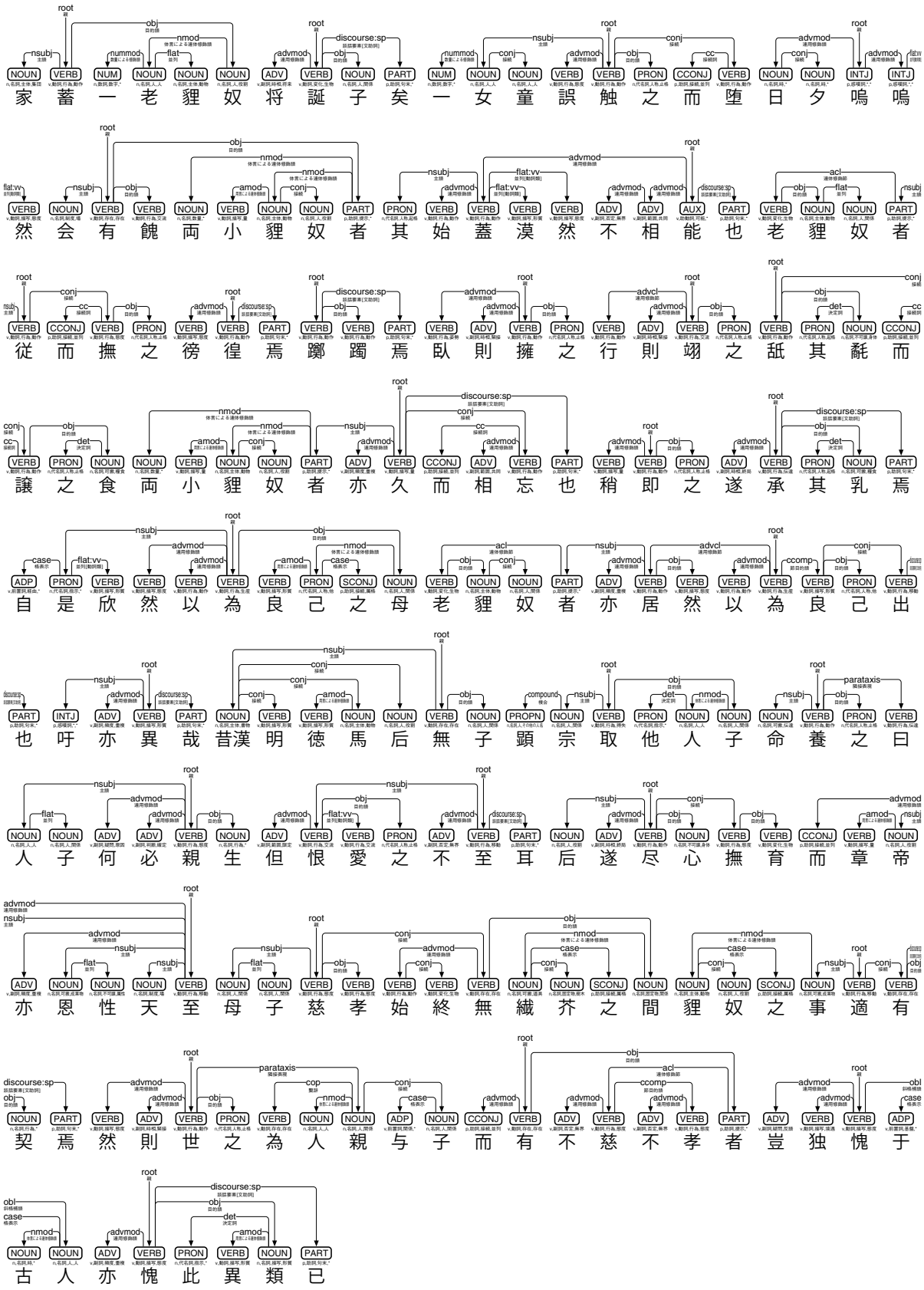


図 90: 手法❶の処理結果 (2015 年)

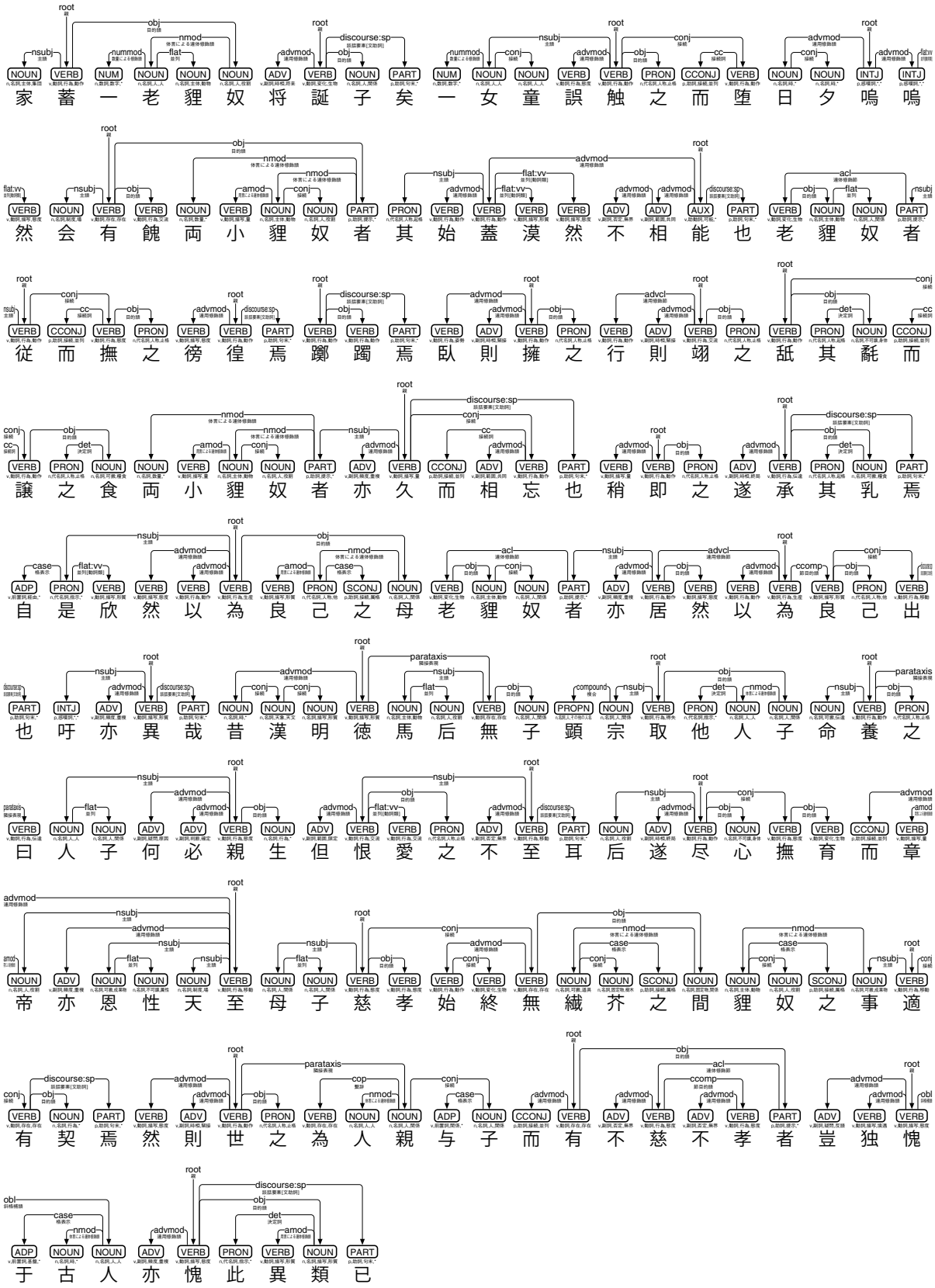


図 91: 手法❷の処理結果 (2015 年)

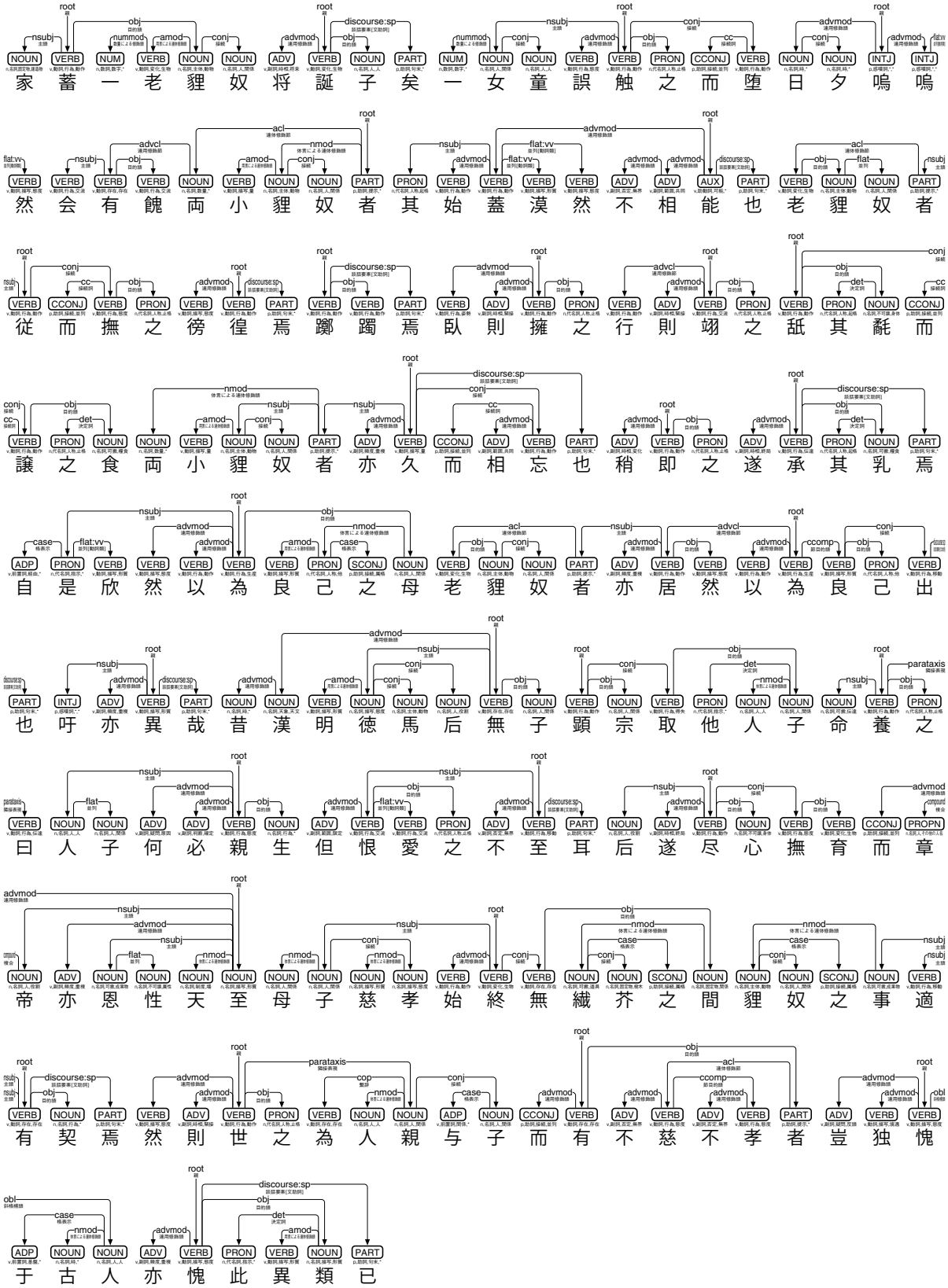


図 92: 手法❸の処理結果 (2015 年)

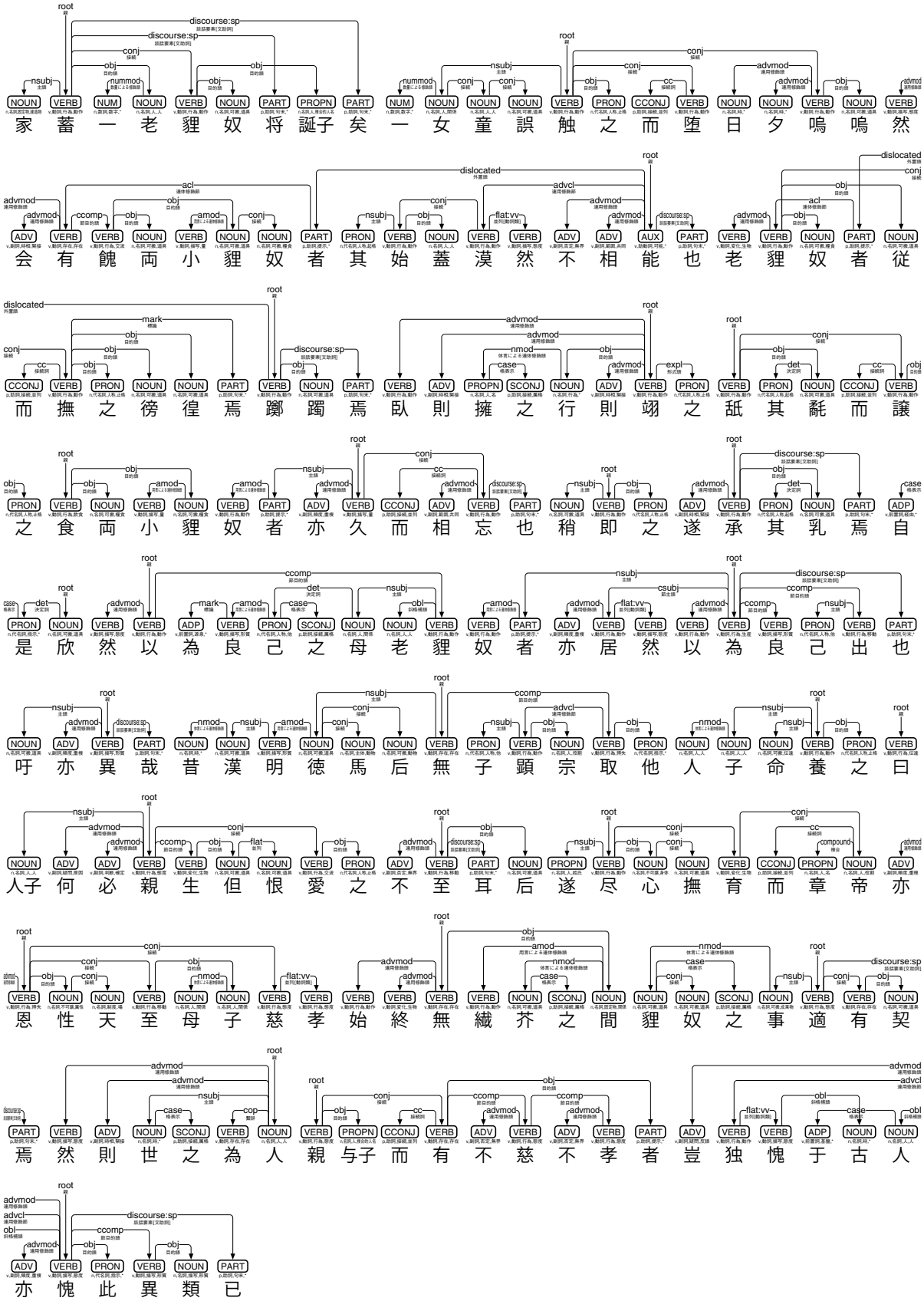


図 93: 手法Ⓐの処理結果 (2015 年)

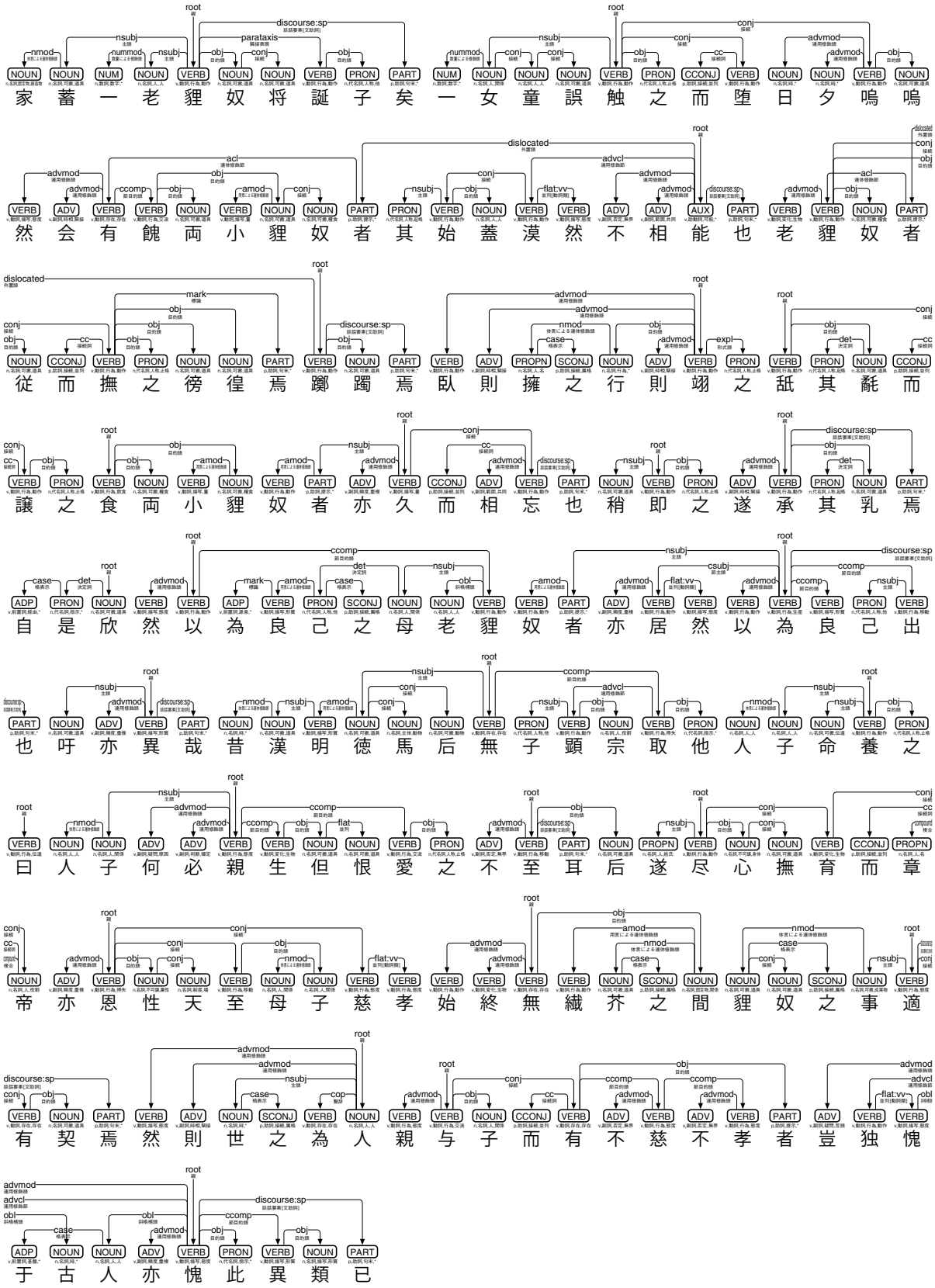


図 94: 手法Ⓑの処理結果 (2015 年)

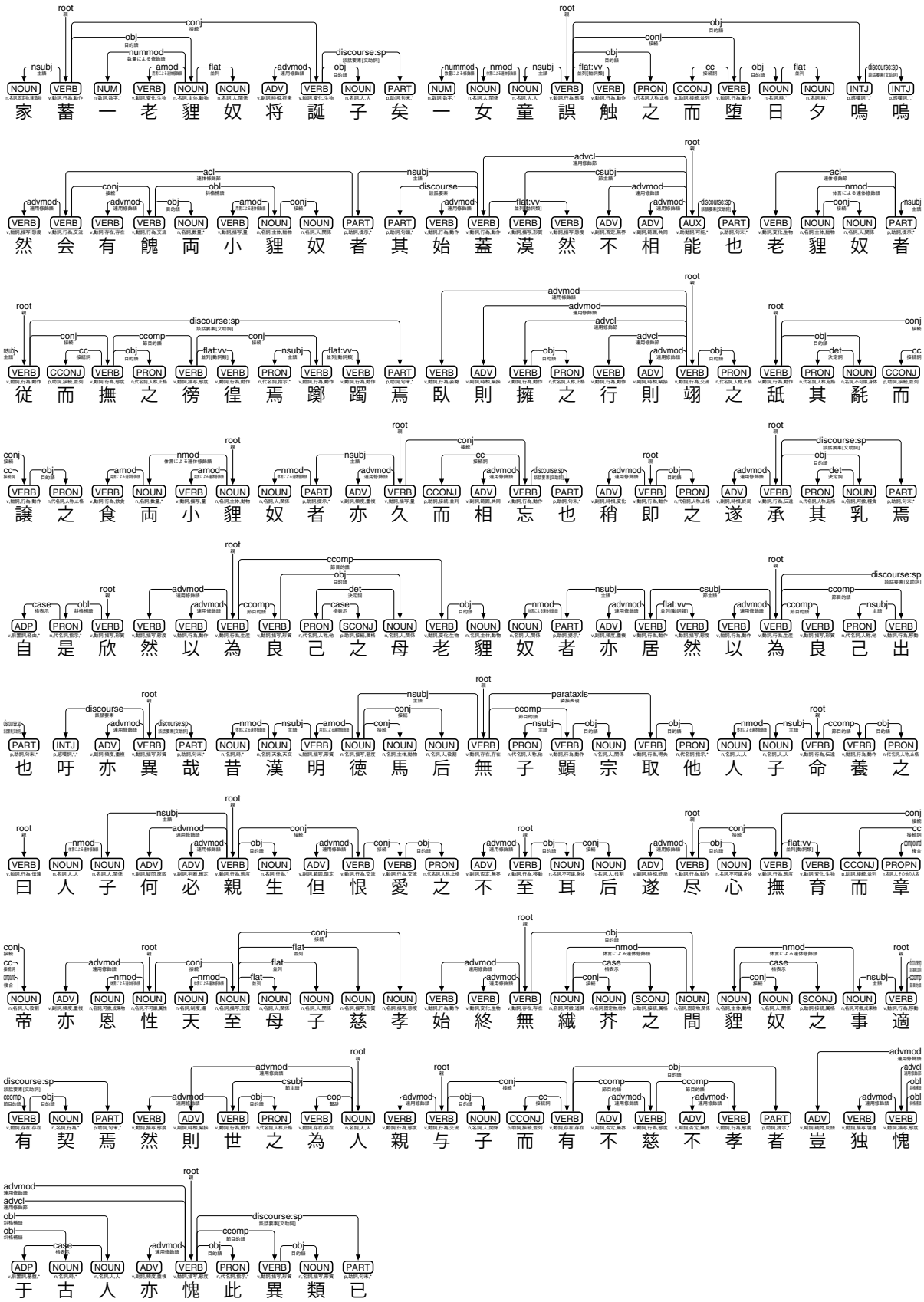


図 95: 手法Ⓒの処理結果 (2015 年)

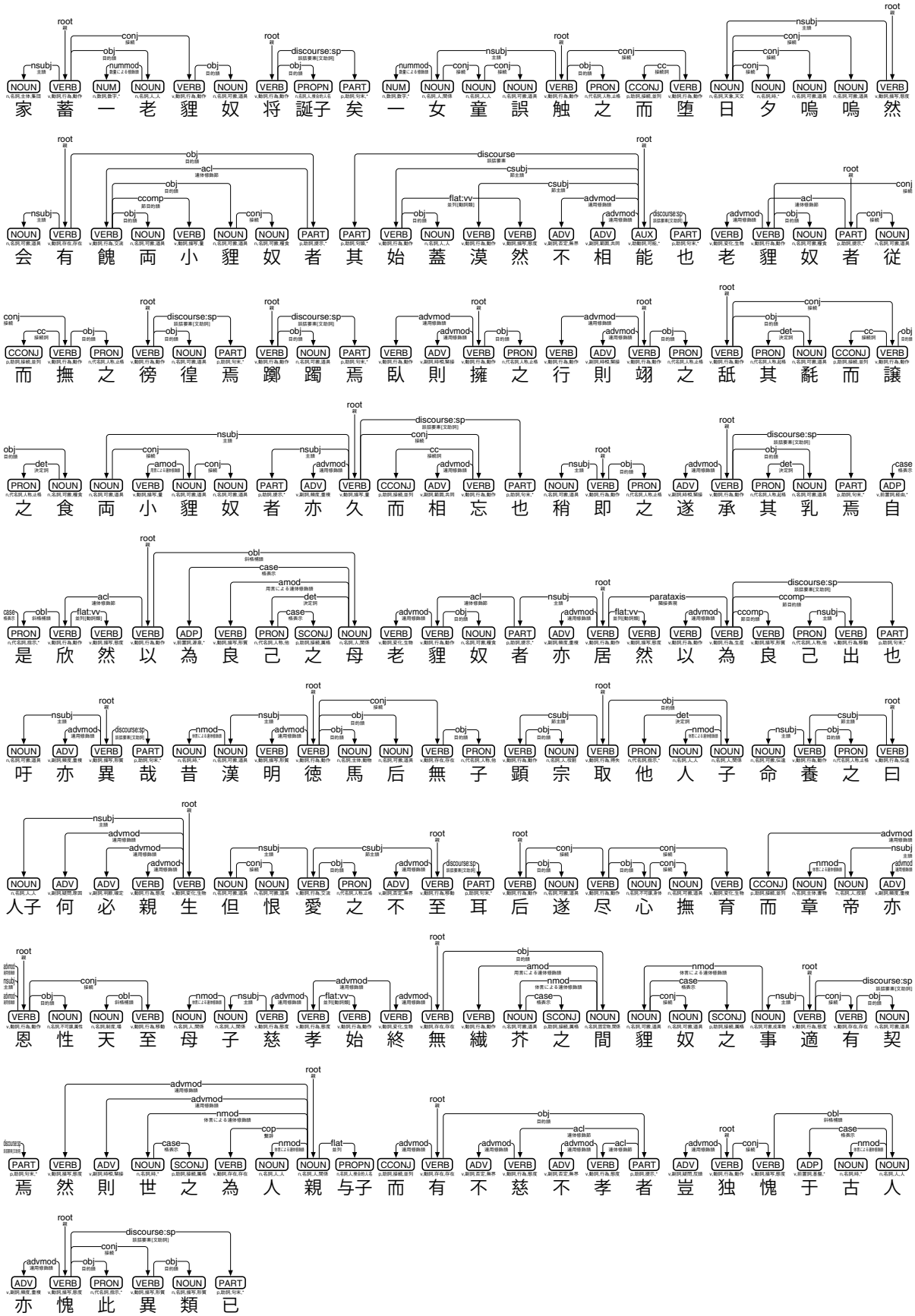


図 96: 手法Aの処理結果 (2015 年)

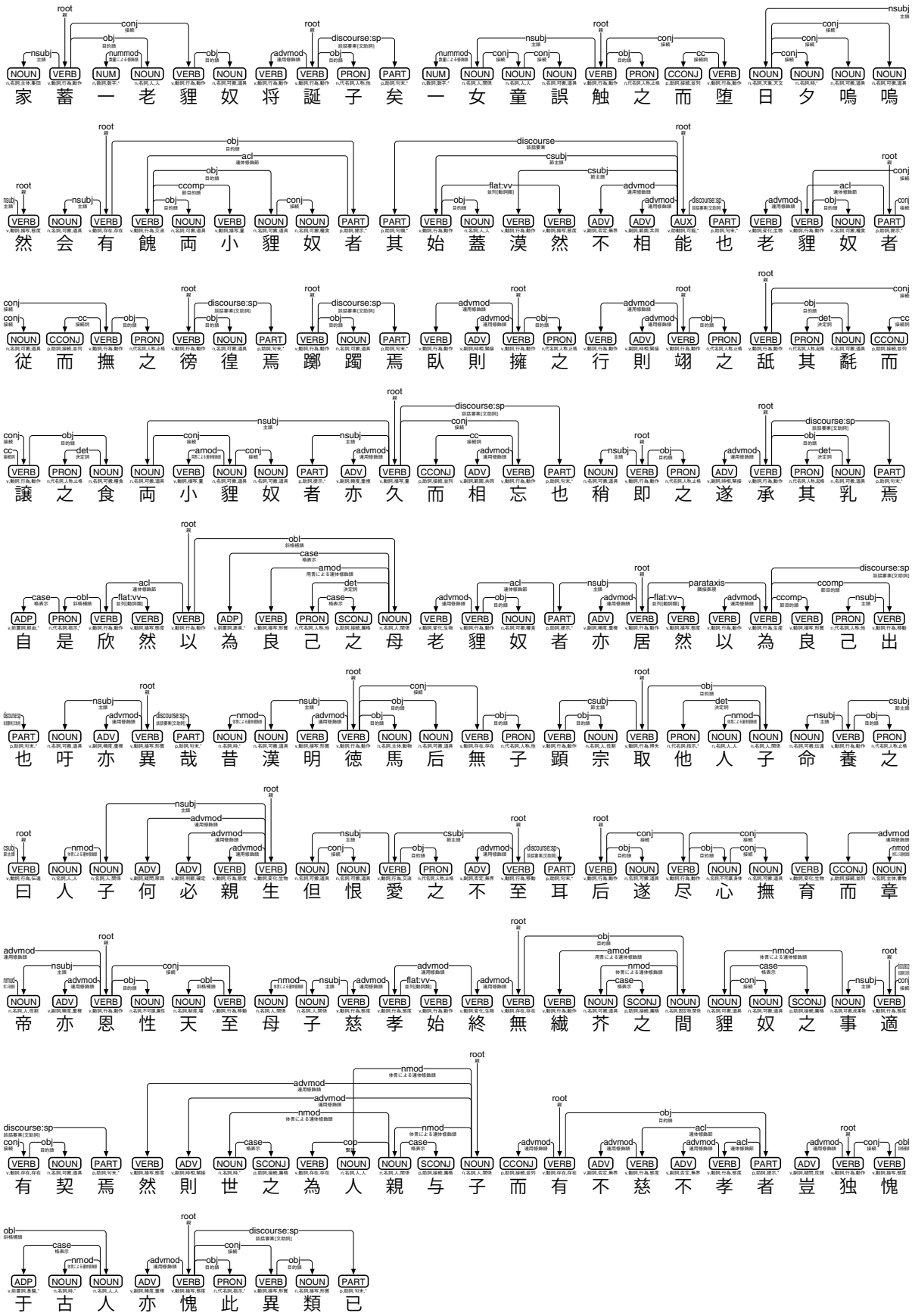


図 97: 手法Bの処理結果 (2015 年)

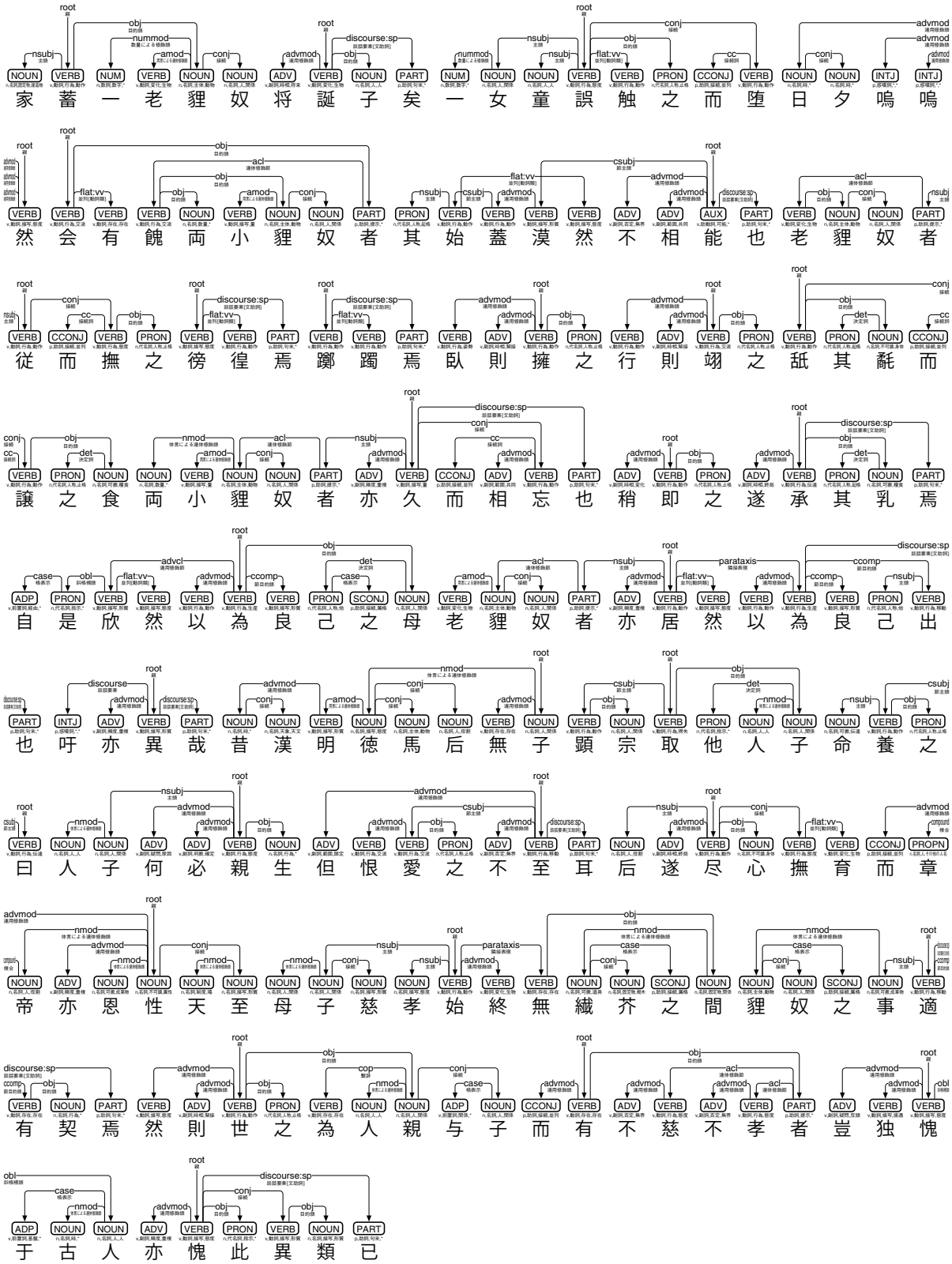


図 98: 手法Ⓒの処理結果 (2015 年)

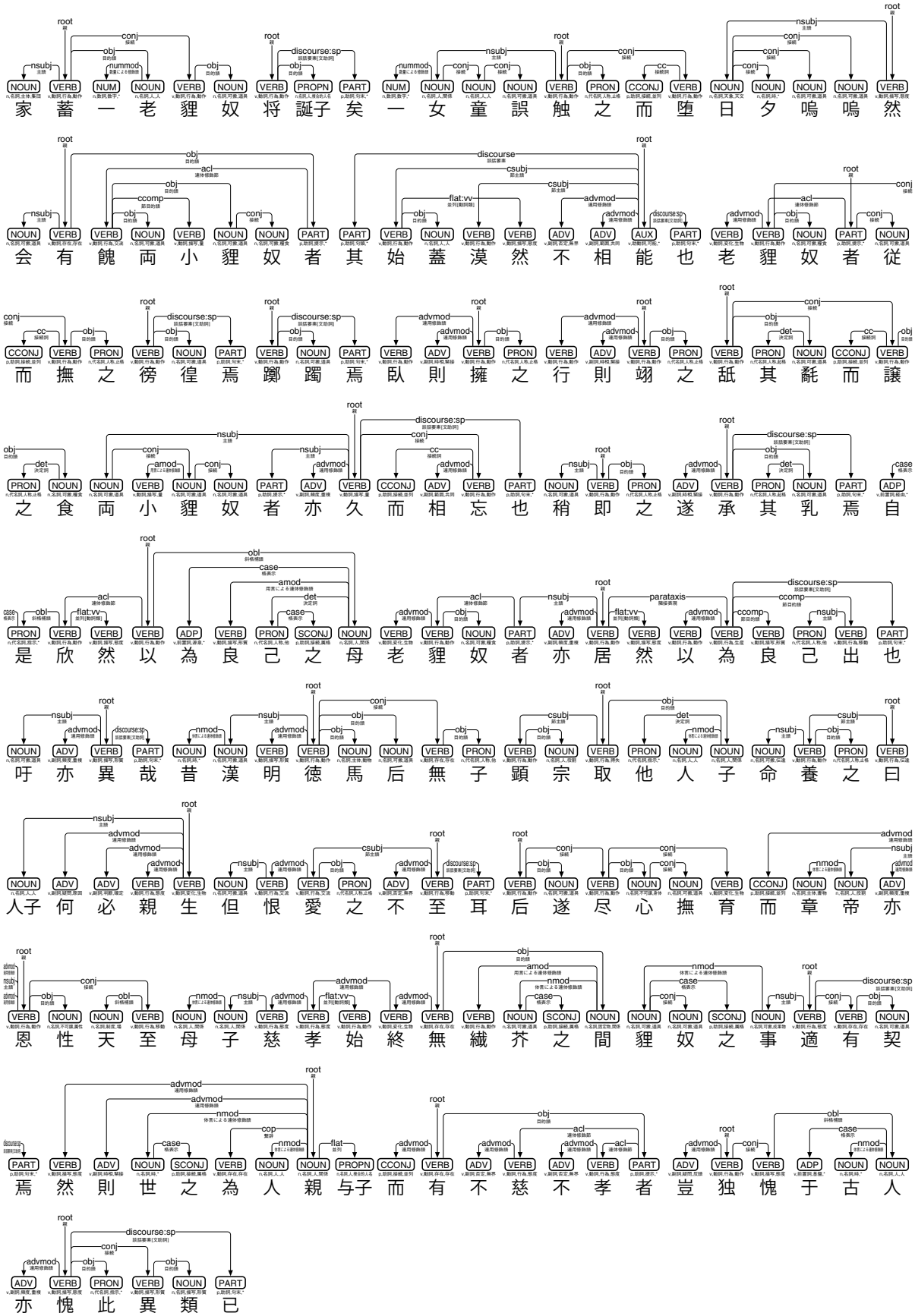


図 99: 手法Ⓐの処理結果 (2015 年)

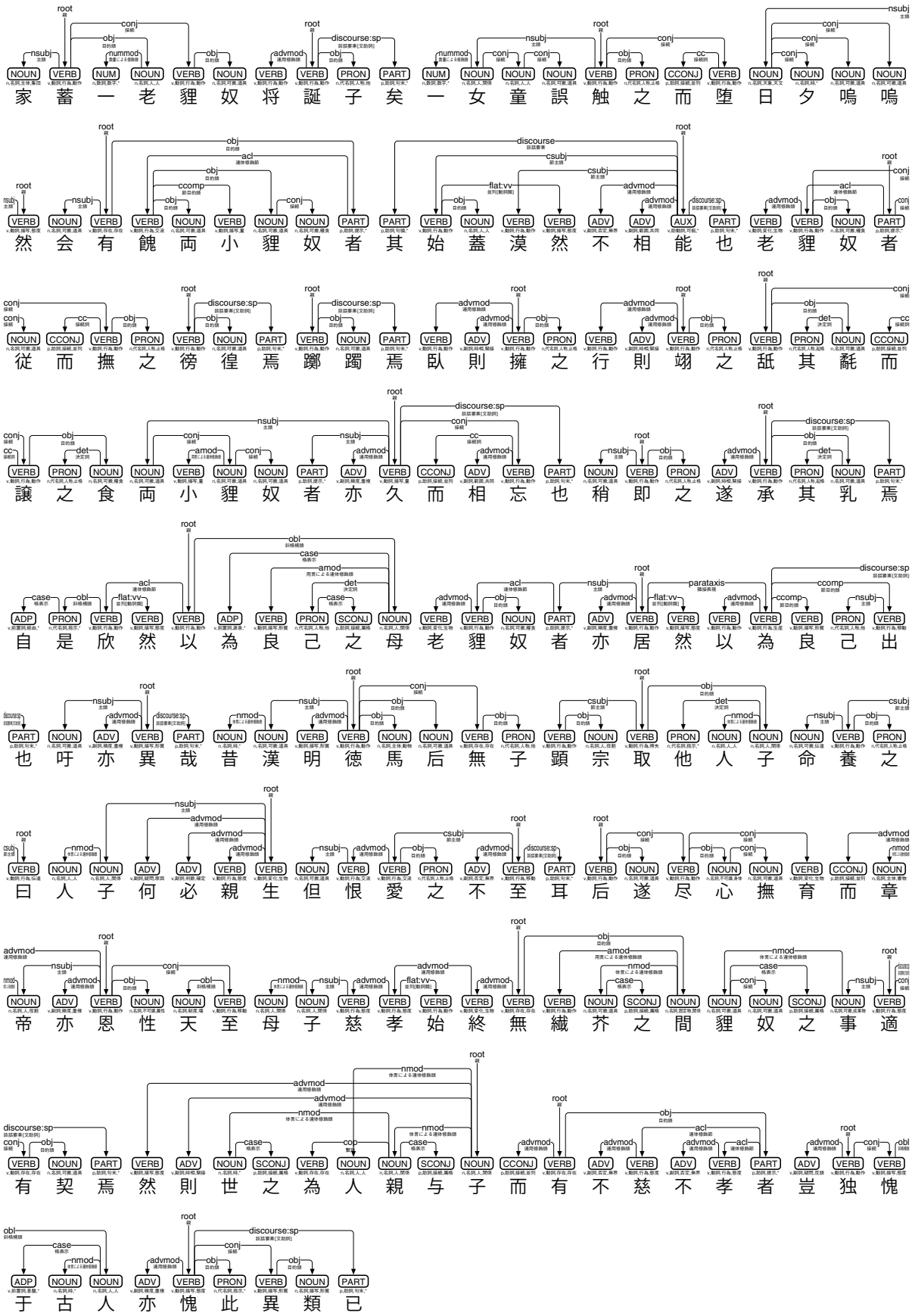


図 100: 手法Ⓑの処理結果 (2015 年)

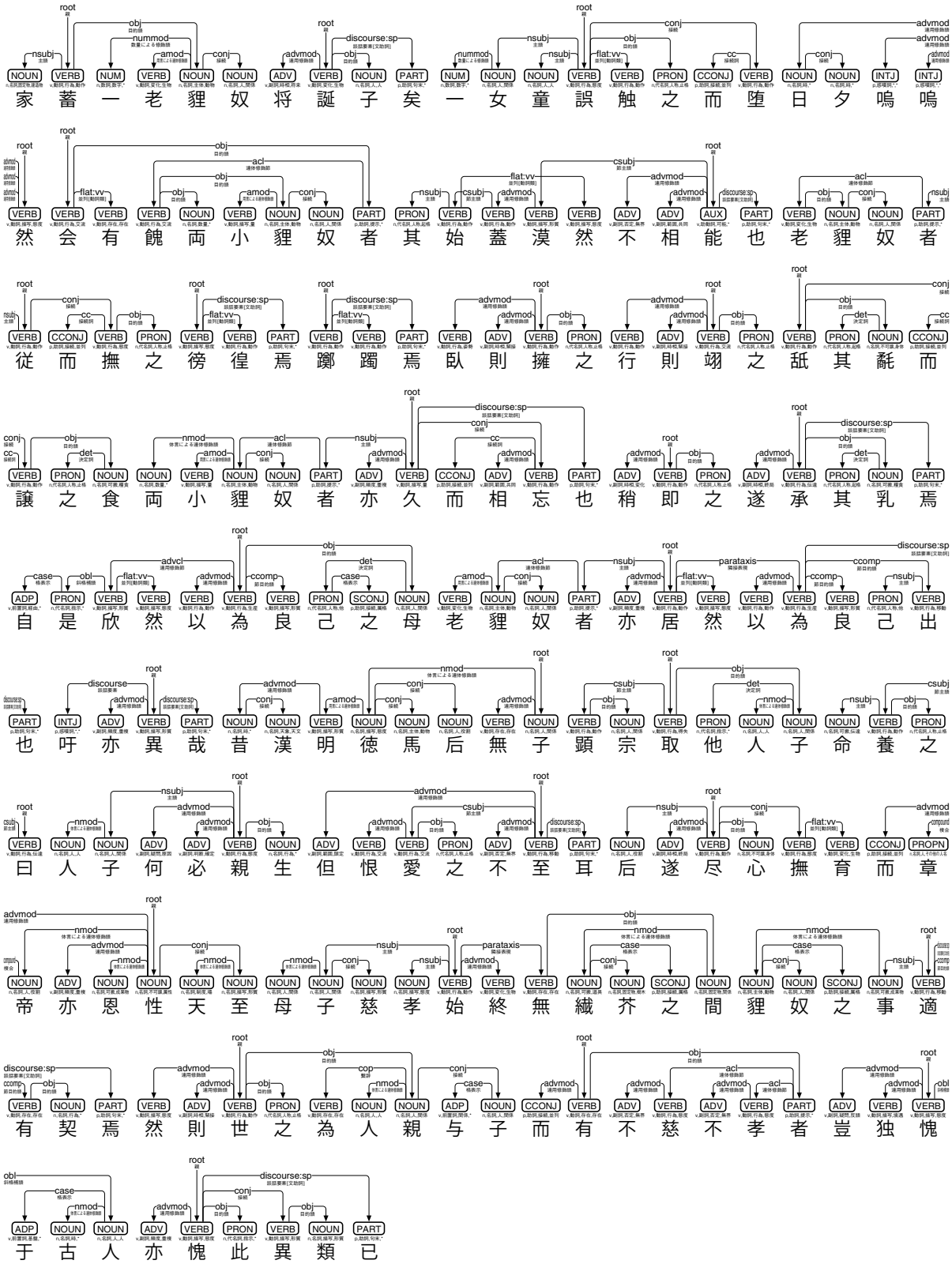


図 101: 手法Ⓒの処理結果 (2015 年)

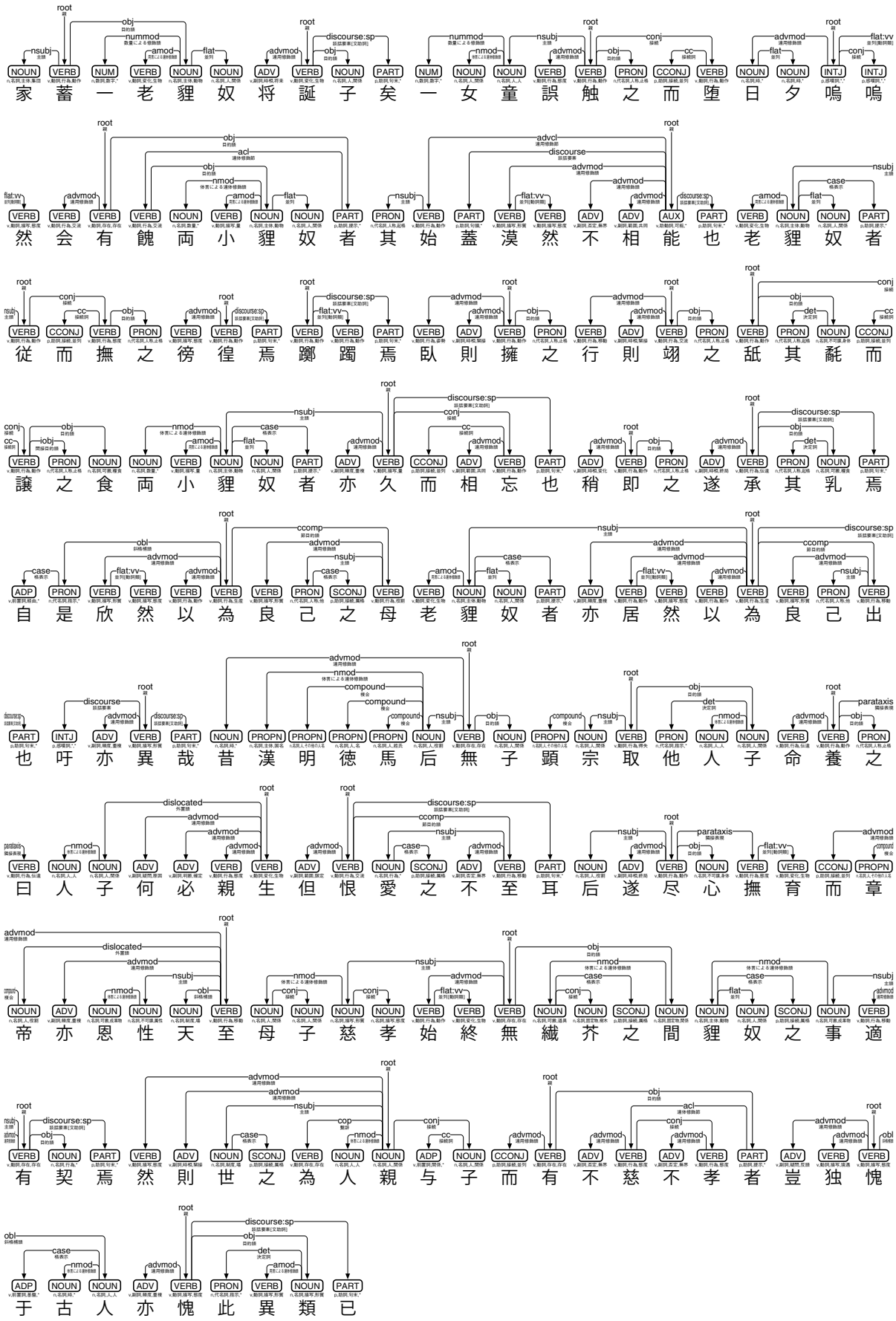


図 102: 手作業で作成した「正解」UD (2015 年)