Vol.2020-CH-124 No.3
2020/9/5

形態素解析部の付け替えによる 近代日本語(旧字旧仮名)の係り受け解析

安岡孝一 (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 はじめに

明治から戦前にかけての日本の文学作品は，いわゆる旧字旧仮名で発表されている．これらの作品を，できれば旧字旧仮名のままで解析したい．

旧字旧仮名の係り受け解析システムは，管見の限り作られていないようだ．ただし，UDPipe [1] をはじめとする多言語係り受け解析システムの多くは，形態素解析部と係り受け解析部に別々の言語モデルを導入可能な設計となっている．ならば，これらの解析システムの形態素解析部を，旧仮名口語 UniDic [2] や近代文語 UniDic [3] に付け替えて，係り受け解析部には現代日本語 (新字新仮名) モデルを使い回すと，一体どうなるだろう．どの程度の解析精度が得られるのだろう．

本稿では，7 種類の現代日本語係り受け解析システムに対し，それぞれの形態素解析部を，近代文語 UniDic・旧仮名口語 UniDic・近世口語 UniDic [4] に付け替える実験をおこなった．結果について報告する．

2 係り受け解析システムにおける形態素解析部の付け替え

本稿で実験に用いた現代日本語係り受け解析システムを，表 1 に示す．これらのシステムは，いずれも，UTF-8 で書かれた現代日本語テキストの係り受け解析 (単語間係り受け) をおこない，日本語 Universal Dependencies [5] に準拠した出力をおこなうもので，形態素解析部と係り受け解析部が分離可能なものである．出力に用いられる品詞体系は，UPOS (Universal Part-Of-Speech) と呼ばれる Universal Dependencies 準拠のものだが，形態素解析部から係り受け解析部への受け渡しに用いられる内部品詞体系は，それぞれに異なっている．

内部品詞体系が UniDic の場合 (表 1 では spaCy [6] と GiNZA [7]) は，近代文語 UniDic・旧仮名口語 UniDic・近世口語 UniDic への付け替えは，比較的容易である．一方，内部品詞体系が UniDic 以外の場合，付け替え後に品詞変換をおこなわなければ，後続の係り受け解析部が動作しない．筆者が作成した品詞変換ルールを，表 2 に示す．これに加え，IPADic [8] や JUMAN [9] への品詞変換に際しては，活用型・活用形の変換もおこなうべきなのだが，筆者はこれを諦めた．IPADic や JUMAN の活用型には,「文語四段」だの「文語助動詞」だのは存在しないからである．具体的には，UniDic 品詞体系から IPADic 品詞体系への変換に際しては，諦めて活用型をそのまま渡すことにした．UniDic 品詞体系から JUMAN 品詞体系への変換に際しては，表 3 に示したもの以外，活用型を一括して「母音動詞」とした．例として「石炭をば早や積み果てつ」という文における品詞変換の様子を，図 1 に示す．

表 1: 実験に用いた現代日本語係り受け解析システム (単語間係り受け)

	形態素解析部	内部品詞体系	係り受け解析部
spaCy 2.3.2	SudachiPy 0.4.9	UniDic	ja_core_news_lg 2.3.2
GiNZA 3.1.2	SudachiPy 0.4.9	UniDic	ja_ginza 3.1.0
spaCy-SynCha 0.4.7	MeCab 0.996	IPADic	CaboCha 0.69 / SynCha 0.3.1.1
spaCy-ChaPAS 0.4.0	MeCab 0.996	IPADic	CaboCha 0.69 / ChaPAS 0.742
Camphr-KNP 0.6.3	JUMAN 7.01	JUMAN	KNP 4.19
Stanza 1.0.1	Stanza 1.0.1	UPOS / Penn	UD Japanese-GSD 2.5
UDPipe 1.2.0	MorphoDiTa 1.9.1	UPOS / Penn	UD Japanese-GSD 2.5

表 2: UniDic 品詞体系から IPADic・JUMAN・UPOS / Penn への変換

UniDic				IPADic	JUMAN	UPOS / Penn
名詞	普通名詞	一般		名詞, 一般	名詞 6 普通名詞 1	NOUN / NN
		形状詞可能		名詞, 形容動詞語幹		
		サ変可能		名詞, サ変接続	名詞 6 サ変名詞 2	
		サ変形状詞可能				
		助数詞可能		名詞, 接尾, 助数詞	接尾辞 14 名詞性名詞助数辞 3	
		副詞可能		名詞, 副詞可能	名詞 6 時相名詞 10	NOUN / NR
	助動詞語幹			名詞, 特殊, 助動詞語幹	助動詞 5 ∗ 0	AUX / RB
	固有名詞	一般		名詞, 固有名詞, 一般	名詞 6 固有名詞 3	PROPN / NNP
		人名	一般	名詞, 固有名詞, 人名, 一般	名詞 6 人名 5	
			姓	名詞, 固有名詞, 人名, 姓		
			名	名詞, 固有名詞, 人名, 名		
		地名	一般	名詞, 固有名詞, 地域, 一般	名詞 6 地名 4	
			国	名詞, 固有名詞, 地域, 国		
	数詞			名詞, 数	名詞 6 数詞 7	NUM / CD
代名詞				名詞, 代名詞, 一般	指示詞 7 名詞形態指示詞 1	PRON / NP
動詞	一般			動詞, 自立	動詞 2 ∗ 0	VERB / VV
	非自立可能			動詞, 非自立		
助動詞				助動詞	助動詞 5 ∗ 0	AUX / AV
形容詞	一般			形容詞, 自立	形容詞 3 ∗ 0	ADJ / JJ
	非自立可能			形容詞, 非自立		
形状詞	一般			名詞, 形容動詞語幹		
	タリ					
	助動詞語幹			名詞, 接尾, 助動詞語幹		
連体詞				連体詞	連体詞 4 ∗ 0	DET / JR
副詞				副詞, 一般	副詞 8 ∗ 0	ADV / RB
助詞	格助詞			助詞, 格助詞, 一般	助詞 9 格助詞 1	ADP / PS
	係助詞			助詞, 係助詞	助詞 9 副助詞 2	ADP / PK
	副助詞			助詞, 副助詞		ADP / PH
	準体助詞			名詞, 非自立, 一般	名詞 1 形式名詞 8	ADP / PN
	終助詞			助詞, 終助詞	助詞 9 終助詞 4	PART / PE
	接続助詞			助詞, 接続助詞	助詞 9 接続助詞 3	SCONJ / PC
接続詞				接続詞	接続詞 10 ∗ 0	CCONJ / CC
接頭辞				接頭詞, 名詞接続	接頭辞 13 名詞接頭辞 1	NOUN / XP
接尾辞	名詞的	一般		名詞, 接尾, 一般	接尾辞 14 名詞性名詞接尾辞 2	NOUN / NN
		サ変可能		名詞, 接尾, サ変接続		
		副詞可能		名詞, 接尾, 副詞可能	接尾辞 14 名詞性特殊接尾辞 4	
		助数詞		名詞, 接尾, 助数詞	接尾辞 14 名詞性名詞助数辞 3	NOUN / XSC
	動詞的			動詞, 接尾	接尾辞 14 動詞性接尾辞 7	PART / AV
	形容詞的			形容詞, 接尾	接尾辞 14 形容詞性述語接尾辞 5	PART / JN
	形状詞的			名詞, 接尾, 形容動詞語幹	接尾辞 14 形容詞性名詞接尾辞 6	
感動詞	一般			感動詞	感動詞 12 ∗ 0	INTJ / UH
	フィラー			フィラー		
補助記号	句点			記号, 句点	特殊 1 句点 1	PUNCT / SYM
	読点			記号, 読点	特殊 1 読点 2	
	括弧開			記号, 括弧開	特殊 1 括弧始 3	
	括弧閉			記号, 括弧閉	特殊 1 括弧終 4	
	ＡＡ	一般		記号, アルファベット	特殊 1 記号 5	
		顔文字		記号, 一般		
	一般					
記号	文字					SYM / SYM
	一般					
空白				記号, 空白	特殊 1 空白 6	

表 3: UniDic 活用型・活用形から JUMAN への変換

(a) 動詞の活用型・活用形

UniDic	JUMAN
五段-カ行	子音動詞カ行 2
文語四段-カ行	
五段-ガ行	子音動詞ガ行 4
文語四段-ガ行	
五段-サ行	子音動詞サ行 5
文語四段-サ行	
五段-タ行	子音動詞タ行 6
文語四段-タ行	
五段-ナ行	子音動詞ナ行 7
五段-バ行	子音動詞バ行 8
文語四段-バ行	
五段-マ行	子音動詞マ行 9
文語四段-マ行	
五段-ラ行	子音動詞ラ行 10
文語四段-ラ行	
五段-ワア行	子音動詞ワ行 12
文語四段-ハ行	

UniDic	JUMAN
語幹	語幹 1
未然形	未然形 3
意志推量形	意志形 4
連用形	基本連用形 8
終止形	基本形 2
連体形	
仮定形	基本条件形 7
已然形	
命令形	命令形 6

(b) 形容詞の活用型・活用形

UniDic	JUMAN
形容詞	イ形容詞イ段 19
文語形容詞-シク	
文語形容詞-ク	イ形容詞アウオ段 18

UniDic	JUMAN
語幹	語幹 1
未然形	文語未然形 19
意志推量形	基本推量形 4
連用形	基本連用形 7
終止形	基本形 2
連体形	文語連体形 21
仮定形	基本条件形 6
已然形	
命令形	文語命令形 22

```
石炭	名詞,普通名詞,一般,*,*,*,セキタン,石炭,石炭,セキタン,セキタン,漢,石炭,セキタン,セキタン,セキタン,*,*,*,*,*,*,3,C2,*
をば	助詞,格助詞,*,*,*,*,ヲバ,をば,をば,オバ,ヲバ,和,をば,オバ,ヲバ,ヲバ,*,*,*,*,*,*,*,*,*
早や	副詞,*,*,*,*,*,ハヤ,早,早や,ハヤ,ハヤ,和,早や,ハヤ,ハヤ,ハヤ,*,*,*,*,*,*,1,*,*
積み	動詞,一般,*,*,文語四段-マ行,連用形-一般,ツム,積む,積み,ツミ,ツミ,和,積む,ツム,ツム,ツム,*,*,*,*,*,*,0,C4,*
果て	動詞,非自立可能,*,*,文語下二段-タ行,連用形-一般,ハテル,果てる,果て,ハテ,ハテ,和,果つ,ハツ,ハツ,ハツ,*,*,*,*,*,*,1,C1,*
つ	助動詞,*,*,*,文語助動詞-ツ,終止形-一般,ツ,つ,つ,ツ,ツ,和,つ,ツ,ツ,ツ,*,*,*,*,*,*,*,動詞%F4@0,*
EOS
```

(a) 近代文語 UniDic による出力

```
石炭	名詞,一般,*,*,*,*,石炭,セキタン,セキタン
をば	助詞,格助詞,一般,*,*,*,をば,ヲバ,オバ
早や	副詞,一般,*,*,*,*,早,ハヤ,ハヤ
積み	動詞,自立,*,*,文語四段-マ行,連用形,積む,ツミ,ツミ
果て	動詞,非自立,*,*,文語下二段-タ行,連用形,果てる,ハテ,ハテ
つ	助動詞,*,*,*,文語助動詞-ツ,基本形,つ,ツ,ツ
EOS
```

(b) IPADic 品詞体系へ変換

```
石炭 せきたん 石炭 名詞 6 普通名詞 1 * 0 * 0 NIL
をば をば をば 助詞 9 格助詞 1 * 0 * 0 NIL
早や はや 早 副詞 8 * 0 * 0 * 0 NIL
積み つみ 積む 動詞 2 * 0 子音動詞マ行 9 基本連用形 8 NIL
果て はて 果てる 動詞 2 * 0 母音動詞 1 基本連用形 8 NIL
つ つ つ 助動詞 5 * 0 母音動詞 1 基本形 2 NIL
EOS
```

(c) JUMAN 品詞体系へ変換

```
1	石炭	石炭	NOUN	NN	_	_	_	_	SpaceAfter=No|Translit=セキタン
2	をば	をば	ADP	PS	_	_	_	_	SpaceAfter=No|Translit=ヲバ
3	早や	早	ADV	RB	_	_	_	_	SpaceAfter=No|Translit=ハヤ
4	積み	積む	VERB	VV	_	_	_	_	SpaceAfter=No|Translit=ツミ
5	果て	果てる	VERB	VV	_	_	_	_	SpaceAfter=No|Translit=ハテ
6	つ	つ	AUX	AV	_	_	_	_	SpaceAfter=No|Translit=ツ
```

(d) UPOS / Penn 品詞体系へ変換

図 1: 「石炭をば早や積み果てつ」の各品詞体系フォーマット

表 4: 形態素解析部の付け替えを『舞姫』冒頭部で評価 (LAS / MLAS / BLEX)

	近代文語 UniDic	旧仮名口語 UniDic	近世口語 UniDic	オリジナル
spaCy	52.83 / 33.90 / 33.90	52.34 / 33.90 / 33.90	50.00 / 30.00 / 33.33	44.25 / 24.62 / 18.46
GiNZA	52.83 / 32.14 / 32.14	52.83 / 32.14 / 32.14	50.00 / 31.58 / 31.58	42.48 / 22.95 / 16.39
spaCy-SynCha	84.91 / 70.37 / 74.07	81.13 / 66.67 / 70.37	72.22 / 57.14 / 57.14	35.59 / 19.35 / 16.13
spaCy-ChaPAS	79.25 / 59.26 / 62.96	77.36 / 59.26 / 62.96	70.37 / 53.57 / 53.57	28.81 / 15.87 / 9.52
Camphr-KNP	54.72 / 29.09 / 32.73	50.94 / 25.45 / 29.09	51.85 / 32.14 / 32.14	26.67 / 8.96 / 8.96
Stanza	66.04 / 47.27 / 50.91	62.26 / 43.64 / 50.91	59.26 / 45.61 / 45.61	26.42 / 10.71 / 7.14
UDPipe	69.81 / 47.27 / 54.55	67.92 / 47.27 / 54.55	57.41 / 42.11 / 42.11	28.04 / 10.71 / 7.14

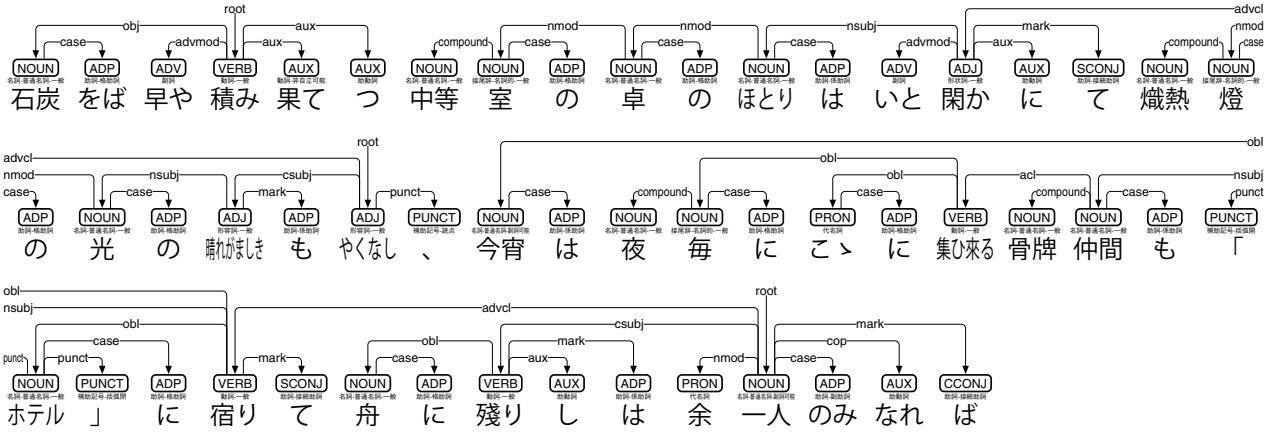
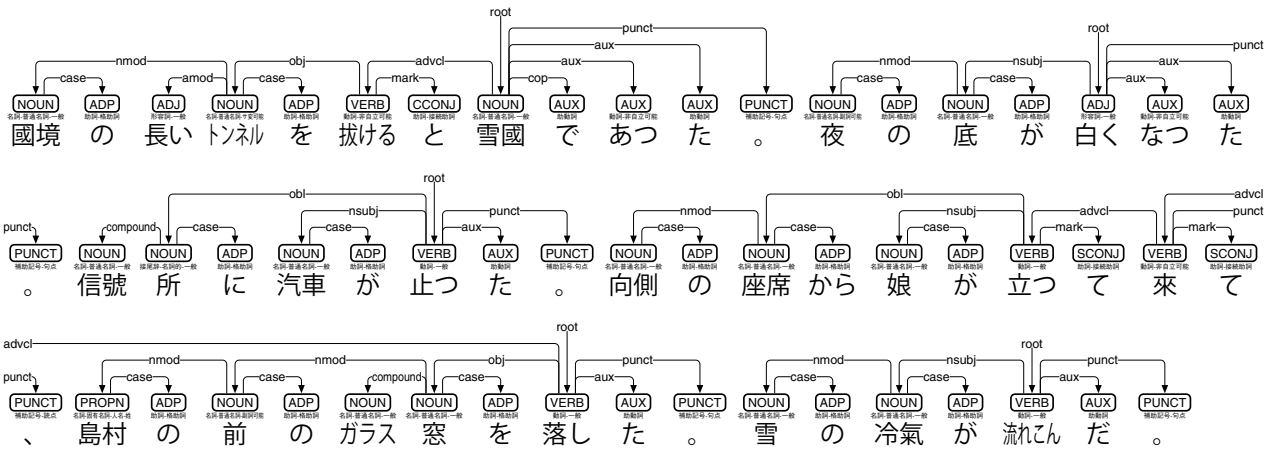


表 5: 形態素解析部の付け替えを『雪國』冒頭部で評価 (LAS / MLAS / BLEX)

	近代文語 UniDic	旧仮名口語 UniDic	近世口語 UniDic	オリジナル
spaCy	74.34 / 60.38 / 64.15	75.00 / 65.38 / 69.23	75.00 / 65.38 / 65.38	51.28 / 33.90 / 33.90
GiNZA	76.11 / 76.92 / 73.08	76.79 / 76.92 / 73.08	76.79 / 76.92 / 69.23	52.99 / 42.86 / 39.29
spaCy-SynCha	77.88 / 69.39 / 69.39	80.36 / 69.39 / 69.39	83.93 / 77.55 / 73.47	43.33 / 32.14 / 28.57
spaCy-ChaPAS	83.19 / 77.55 / 73.47	87.50 / 81.63 / 77.55	89.29 / 85.71 / 77.55	45.00 / 31.58 / 31.58
Camphr-KNP	63.72 / 46.15 / 46.15	67.86 / 50.00 / 50.00	73.21 / 57.69 / 53.85	26.09 / 13.33 / 13.33
Stanza	72.57 / 65.38 / 61.54	76.79 / 69.23 / 65.38	75.00 / 69.23 / 61.54	78.18 / 64.00 / 40.00
UDPipe	72.57 / 65.38 / 61.54	76.79 / 69.23 / 65.38	76.79 / 69.23 / 61.54	58.93 / 45.28 / 33.96



3 評価と改良

『舞姫』[10] と『雪國』[11] の冒頭部を，事前に手作業で解析して準備しておき，それらを例題として用いるやり方で，形態素解析部の付け替えに対する評価をおこなった．例題の品詞体系は UPOS / UniDic であり，単語長は短単位 [12] である．評価指標は，LAS (Labeled Attachment Score) / MLAS (Morphology-aware Labeled Attachment Score) / BLEX (Bi-LEXical dependency score) の 3 つの指標 [13] を用いた．ただし，Universal Dependencies の第 6 (FEATS)・第 9 (DEPS) フィールドは，評価に用いていない．

『舞姫』冒頭部による評価結果を表 4 に,『雪國』冒頭部による評価結果を表 5 に示す．比較のために，形態素解析部を付け替える前の評価値を「オリジナル」として示した．全ての付け替えにおいて，付け替え前より MLAS / BLEX が上昇している．本稿の手法の有効性が示された，と言っていいだろう．

全体として高い評価値を示しているのは，spaCy-SynCha である．SynCha の土台となっている述語項構造のアノテーションにおいては，述語の基本形 (見出し語形) にタグを付与する，という方針 [14] が取られており，これが高い評価値に繋がったと考えられる．というのも，UniDic においては，書字形 (表層形) が旧字旧仮名の場合でも，語彙素 (見出し語形) は新字新仮名で表されている．したがって，入力が旧字旧仮名であっても，SynCha での処理は新字新仮名でおこなわれており，形態素解析が適切であれば，SynCha 本来の性能が発揮されるわけである．ただ，同様の方針は ChaPAS でも採用 [15] されており，モデルによって得手不得手があるようだ．

他と比べて評価値が低いのは Camphr-KNP だが，これは，単語長のミスマッチが大きいと考えられる．UniDic は単語長に短単位を採用しているが，JUMAN の単語長は異なっている．たとえば「立って来て」を，UniDic は「立っ」「て」「来」「て」の 4 語として扱うが，JUMAN は「立って」「来て」の 2 語として扱う.「落した」を，UniDic は「落し」「た」の 2 語とするが，JUMAN は「落した」の 1 語とする．これを敷衍すると,「積み果てつ」は JUMAN において「積み」「果てつ」の 2 語として扱うべきだが，筆者の変換ツールは「積み」「果て」「つ」の 3 語のまま (図 1(c)) である．この単語長のミスマッチが，後続の KNP での解析 [16] に悪影響を及ぼし，評価値が上がらない結果になっていると思われる．ただ，単語長を変更すると，それだけで LAS / MLAS / BLEX は下がってしまうことから，なかなか評価が悩ましい．

『雪國』冒頭部に対する Stanza「オリジナル」の評価値は，かなり特異なケースである．Stanza の形態素解析部は，事前に Chinese Gigaword Corpora で鍛えられており [17]，中国語の繁體字 (および簡化字) で書かれた単語を数多く知っている．これにより Stanza は，日本語の旧字「國境」「雪國」「信號」「冷氣」も，難なく読める結果になっているようだ．

UDPipe については，係り受け解析部を UD Japanese-GSD 2.5 から UD Japanese-Modern 2.5 に入れ替えることで，さらなる改良をおこなってみた．UD Japanese-Modern 2.5 は近代日本語 (旧字旧かな) の係り受けコーパスであり，品詞体系に UPOS / UniDic を採用している [18] からである．なお，コーパス中で見出し語形にあたる第 3 (LEMMA) フィールドは，もちろん新字新仮名とした．UniDic2UD 2.0.0 としてパッケージ化をおこなった上,『舞姫』『雪國』の冒頭部で評価した結果を，表 6 に示す．全ての評価値において，UD Japanese-GSD 2.5 より UD Japanese-Modern 2.5 の方が改善されている．ただし，この結果は，形態素解析部の付け替えだけでなく，係り受け解析部も入れ替えた方がよい，という事実を示している．当たり前の結果とは言え，本稿の立場としてはクヤシイ．

表 6: UDPipe 1.2.0 の係り受け解析部を UD Japanese-Modern 2.5 に入れ替えた場合 (LAS / MLAS / BLEX)

		近代文語 UniDic	旧仮名口語 UniDic	近世口語 UniDic
UniDic2UD 2.0.0 (UD Japanese-Modern 2.5)	『舞姫』	69.81 / 58.18 / 61.82	67.92 / 58.18 / 61.82	66.67 / 56.14 / 59.65
	『雪國』	84.96 / 78.43 / 78.43	91.07 / 86.27 / 86.27	89.29 / 86.27 / 82.35

4 おわりに

近代日本語 (旧字旧仮名) の係り受け解析において，形態素解析部の付け替えにより，現代日本語の係り受け解析システムを使い回しても，十分な解析精度が得られることを示した．spaCy-SynCha については，結果がかなり良好だったことから，Web 茶まめ [19] に接続するモジュールを搭載する形で，spaCy-SynCha 0.5.0 にアップデートした．spaCy-ChaPAS も同様に，Web 茶まめに接続した．UniDic2UD については，UDPipe を内蔵するのみならず，spaCy・GiNZA・Stanza へのインターフェースも追加した．Camphr-KNP については，単語長を変更する機能など色々と試作してみたものの，付け替えモジュールの Pull Request 等は出していない．

なお，本稿の実験途中に，UD Japanese-GSD 2.6 がリリースされた．UD Japanese-GSD 2.6 は，品詞体系が UPOS / UniDic に変更 [20] されており，単語長も短単位に変更されている．UDPipe や Stanza の日本語モデルが，バージョンアップで UD Japanese-GSD 2.6 を採用したら，本稿の手法もさらに解析精度が上がるのではないかと期待しつつ，それは別稿に譲ることにする．

参考文献

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

[2] 小木曽智信: 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.

[3] 小木曽智信, 小町守, 松本裕治: 歴史的日本語資料を対象とした形態素解析, 自然言語処理, Vol.20, No.5 (2013 年 10 月), pp.727-748.

[4] 小木曽智信, 市村太郎, 鴻野知暁: 近世口語資料の形態素解析の試み, 第 4 回コーパス日本語学ワークショップ予稿集 (2013 年 9 月), pp.145-150.

[5] 浅原正幸, 金山博, 宮尾祐介, 田中貴秋, 大村舞, 村脇雄吾, 松本裕治: Universal Dependencies 日本語コーパス, 自然言語処理, Vol.26, No.1 (2019 年 3 月), pp.3-36.

[6] Matthew Honnibal, Mark Johnson: An Improved Non-monotonic Transition System for Dependency Parsing, EMNLP 2015: Conference on Empirical Methods in Natural Language Processing (September 2015), pp.1373-1378.

[7] 松田寛, 大村舞, 浅原正幸: 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習, 言語処理学会第 25 回年次大会発表論文集 (2019 年 3 月), pp.201-204.

[8] 浅原正幸, 松本裕治: ipadic version 2.7.0 ユーザーズマニュアル, 奈良: 奈良先端科学技術大学情報科学研究科自然言語処理学講座 (2003 年 11 月).

[9] 日本語形態素解析システム JUMAN version 7.0, 京都: 京都大学大学院情報学研究科黒橋・河原研究室 (2012 年 1 月).

[10] 鷗外森林太郎: 舞姫, 國民之友, 第 6 巻, 第 69 號 (1890 年 1 月) 附錄, pp.45-61.

[11] 川端康成: 雪國, 東京: 創元社 (1937 年 6 月).

[12] 近藤明日子: 近代文語 UniDic 短単位規程集, Ver.1.1, 立川: 国立国語研究所コーパス開発センター (2016 年 3 月).

[13] 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.

[14] 飯田龍, 小町守, 井之上直也, 乾健太郎, 松本裕治: 述語項構造と照応関係のアノテーション, 自然言語処理, Vol.17, No.2 (2010 年 4 月), pp.25-50.

[15] 渡邉陽太郎, 浅原正幸, 松本裕治: 述語語義と意味役割の結合学習のための構造予測モデル, 人工知能学会論文誌, Vol.25, No.2 (2010 年 2 月), pp.252-261.

[16] 笹野遼平, 河原大輔, 黒橋禎夫, 奥村学: 構文・述語項構造解析システム KNP の解析の流れと特徴, 言語処理学会第 19 回年次大会発表論文集 (2013 年 3 月), pp.110-113.

[17] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning: Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, The 58th Annual Meeting of the Association for Computational Linguistics: Proceedings of the System Demonstration (July 2020), pp.101-108.

[18] Mai Omura, Yuta Takahashi, Masayuki Asahara: Universal Dependency for Modern Japanese, Proceedings of the 7th Conference of Japanese Association for Digital Humanities (September 2017), pp.34-36.

[19] 川口寛治, 薦田龍輝, 提智昭: 形態素解析ソフトウェア『Web 茶まめ』の改良と Web API の試作, 言語資源活用ワークショップ 2016 発表論文集 (2017 年 3 月), p.265-272.

[20] 松田寛, 若狭絢, 山下華代, 大村舞, 浅原正幸: UD Japanese GSD の再整備と固有表現情報付与, 言語処理学会第 26 回年次大会発表論文集 (2020 年 3 月), pp.133-136.

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

[22] 工藤拓, 松本裕治: チャンキングの段階適用による日本語係り受け解析, 情報処理学会論文誌, Vol.43, No.6 (2002 年 6 月), pp.1834-1842.

[23] 吉田将: 二文節間の係り受けを基礎とした日本語文の構文解析, 電子通信学会論文誌, Vol.55-D, No.4 (1972 年 4 月), pp.238-244.

[24] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoro: Applying Conditional Random Fields to Japanese Morpho-

logical Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (July 2004), pp.230-237.
[25] Kira Droganova, Daniel Zeman: Towards Deep Universal Dependencies, Proceedings of the Fifth International Conference on Dependency Linguistics (August 2019), pp.144-152.
[26] Jana Straková, Milan Straka, Jan Hajič: Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (June 2014), pp.13-18.

付録　近代日本語における単語間の係り受け解析と文節間の係り受け解析

本稿で扱った係り受け解析 (単語間係り受け) は，日本語 Universal Dependencies に基づくものであり，言語学的には Мельчук 依存文法 [21] の流れにある．一方，日本語における係り受け解析には，CaboCha [22] をはじめとして，文節間の係り受け解析 [23] が多く用いられている．これらの間の関係を,『舞姫』の「産まれん子は君に似て黒き瞳子をや得ん」を例に，筆者なりに記しておく．

本稿で扱った単語間の係り受け解析は，細かく分けると，単語切り・品詞付与・文節切り・文節間の係り受け解析・述語項構造解析・UPOS 品詞付与・Universal Dependencies 係り受け解析，の 7 つのステージから構成される.「産まれん子は君に似て黒き瞳子をや得ん」の解析例を，図 2 に示す．ただし，これらのステージは必ずしもこの順序ではなく，また，複数のステージがまとめておこなわれる場合もある．実際，本稿で形態素解析部と呼んでいるのは，単語切り・品詞付与の 2 ステージを合わせたものである．図 2 は，あくまで概念図なので注意されたい．

spaCy-SynCha を例に挙げると，まず，MeCab [24] が単語切り・品詞付与の 2 ステージをまとめておこなう．それを受けて CaboCha が，文節切り・文節間の係り受け解析の 2 ステージをまとめておこない，続いて SynCha が述語項構造解析をおこなう．最後に，spaCy-SynCha 自身 (内蔵モジュールの syncha2ud) が，UPOS 品詞付与・Universal Dependencies 係り受け解析の 2 ステージをまとめておこなう．この場合において CaboCha は，せいぜい前半 4 ステージまでしかおこなっていない．これが，文節間の係り受け解析と単語間の係り受け解析における差であり，言い換えるなら，単語間の係り受け解析には述語項構造解析が含まれる，ということである．

spaCy-ChaPAS では，まず MeCab が単語切り・品詞付与の 2 ステージをまとめておこなう．それを受けて CaboCha が，文節切り・文節間の係り受け解析の 2 ステージをまとめておこない，続いて ChaPAS が述語項構造解析をおこなう．最後に，spaCy-ChaPAS 自身 (内蔵モジュールの chapas2ud) が，UPOS 品詞付与・Universal Dependencies 係り受け解析の 2 ステージをまとめておこなう．spaCy-ChaPAS においても，単語間の係り受け解析には述語項構造解析が含まれる．

Camphr-KNP では，まず JUMAN が単語切り・品詞付与の 2 ステージをまとめておこなう．それを受けて KNP が，文節切り・文節間の係り受け解析・述語項構造解析の 3 ステージをまとめておこなう．最後に Camphr が，UPOS 品詞付与・Universal Dependencies 係り受け解析の 2 ステージをまとめておこなう．Camphr-KNP においても，単語間の係り受け解析には述語項構造解析が含まれる．

ただし，単語間の係り受け解析に述語項構造解析が含まれる，と言っても，それは部分的である．たとえば図 2 の「子」は，文中の 3 つの動詞 (産まれ・似・得) 全てでガ格となっている．これら 3 つのうち,「産まれ」との関係については acl (連体修飾節) として,「得」との関係については nsubj (主語) として，それぞれ最後の Universal Dependencies に反映されている．しかし「似」との関係については，Universal Dependencies には反映されていない．この点に関して，Universal Dependencies は力不足であり，Deep Universal Dependencies [25] への拡張を検討すべきかもしれない．

UDPipe では，最初に単語切りをおこなった後，MorphoDiTa [26] が品詞付与・UPOS 品詞付与の 2 ステージをまとめておこなう．その後，途中の 3 ステージ (文節切り・文節間の係り受け解析・述語項構造解析) をすっとばして，いきなり Universal Dependencies 係り受け解析に入る．つまるところ UDPipe に，文節という概念は無いし，述語項構造も陽には解析されていない．ただ，UniDic2UD においては，UDPipe による Universal Dependencies 係り受け解析の後に，文節切り・文節間の係り受け解析の 2 ステージをまとめて挿入可能とし，udcabocha と名づけた．蛇足かもしれない，との懸念はあるが，よければ試してみてほしい．

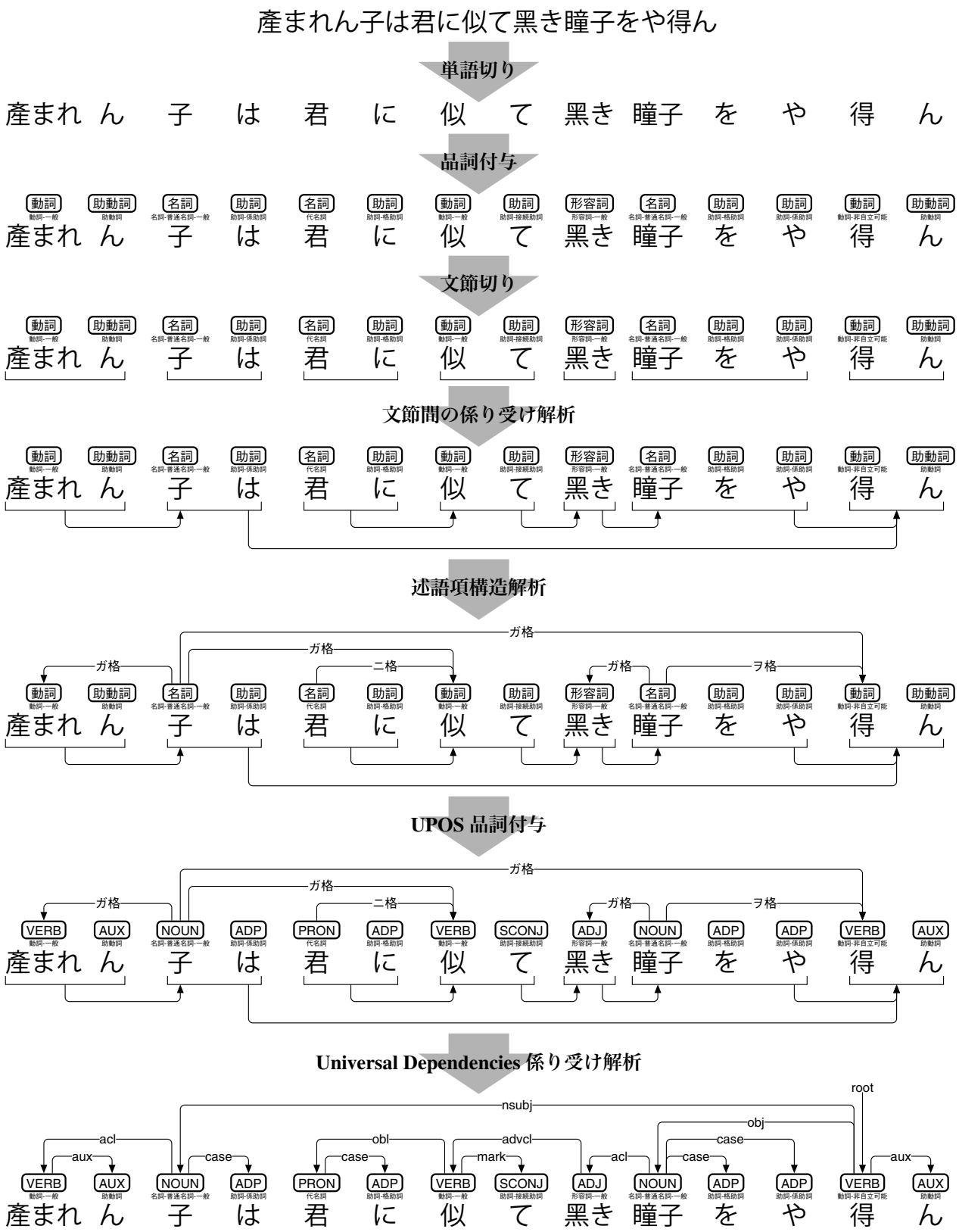


図 2: 近代日本語における単語間の係り受け解析 (概念図)