係り受け構造を利用した日本語句構造文法の構築

Construction of Japanese phrase structure grammar by using dependency structure

新谷 宥希 †　　岡留 剛 †
Yuki Shintani　Takeshi Okadome

概 要

文の構造である構文には，主に「句構造」と「依存構造」の 2 つがあり，日本語文においては後者を「係り受け構造」と呼ぶ．日本語文の句構造解析の精度を上げることは難しく，また，強力な句構造文法も存在しない．本研究では，日本語句構造文法を新たに構築することを目的とし，係り受け構造を用いた句構造規則の構築方法を提案する．自然な係り受け構造木を変換することで得られる句構造木を作成し，そこからの抽出によって句構造規則を構築する．提案手法によって生成した句構造規則で，新たな文を解析できるかを確認する予備実験を行ない，句構造解析木が出力として得られる確率である解析可能率を測定した．結果として，93.0 %という高い数値を得ることができ，汎用的な文法を作成できている可能性が高い．ただし，書き換え規則の集合の表現力が強すぎるため，日本文の句構造としては適切でない木も生成してしまう場合がある．そのため，書き換え規則の制約を強めて非文を生成しないようにする必要があり，その手法の検討が今後の課題として挙げられる．

"Syntax" is the structure of a sentence, and there are mainly 2 type syntaxes, which are "Phrase-Structure" and "Dependency-Structure". However, it is difficult to improve the accuracy of Japanese phrase structure analysis, and there is no strong Japanese-Phrase-Structure-Grammar. Therefore, in this research, we aim to construct a new Japanese-Phrase-Structure-Grammar, and we propose a method to construct phrase structure rules using dependency structure. First, we create phrase structure trees transformed by dependency structure trees that seem to be natural, and then construct phrase structure rules by picking out them from the trees. We conducted a preliminary experiment to confirm whether we can parse a new sentence by our Japanese-Phrase-Structure-Grammar, and measured the probability of analysis, which is probability that phrase structure analysis tree is obtained as output. As a result of the preliminary experiments, we found that the probability of analysis is 93.0%, so we could probably create the generic grammar. However, there are cases where there are many outputs of phrase structure analysis trees for one input sentence. Therefore, we have to narrow down phrase structure analysis trees that are candidates for correct answers, and the study of its method is a future task.

1. はじめに

自然言語処理において，文の構造である構文は人間の知的作業を支援する目的で利用されており，機械翻訳や対話システムなどの分野で応用されている．

構文には，主に「句構造」と「依存構造」の 2 つがあり，日本語文においては後者を「係り受け構造」と呼ぶ．比較的容易に精度を上げることができる係り受け解析は，句構造解析に比べて多くの場面で利用されている．それに対して，文中の語順の自由性や省略の頻発などといった日本語元来の性質のため，日本語は句構造解析の精度を上げることが難しく，また，強力な句構造文法も存在しない．しかし，句構造から得られる情報と，係り受け構造から得られる情報を比較すると，前者の方が多いため，日本語文において句構造解析の精度を上げることは，構文解析結果から得られる情報を増やすという側面において非常に有意義である．

そこで本研究で，日本語句構造文法を新たに構築することを目的とする．新たな日本語句構造文法を作成するために，自然な係り受け構造木を作成し，それを変換することで得られる句構造木から句構造規則を抽出していく手法を提案する．予備実験として，サンプル文 100 文から句構造規則を生成し，新たなテストデータ 20 文を句構造解析した．その結果，112 の句構造規則が構築され，93.0 %という高い確率で解析が可能であることを示した．従来手法と比較すると，かなり少ない規則で高い解析可能率を実現できている．しかし，書き換え規則の集合の表現力が強すぎることにより，日本文の句構造としては適切でない木も生成してしまう場合があるため，書き換え規則の制約の検討が今後の重要な課題である．

以下各説で，関連研究，提案手法，予備実験について詳細に述べ，まとめを記す．

† 関西学院大学, Kwansei Gakuin University

2. 関連研究

一般に公開されているオリジナルの日本語句構造文法は長谷川 [1] の規則以外に見当たらない．しかし，長谷川の句構造規則には規則間に矛盾があり，従属節に関する定義が未定義であったり，意味が不明瞭な記号があるほか，記述力が弱いために，小説文に関してはほとんど解析ができないなどと大幅な改善が必要である．

また，日本語句構造文法の文法理論を詳細に解説したのが郡司 [2] である．一般化句構造文法（GPSG）と，その発展形の主辞駆動句構造文法（HPSG）の方針を取り入れつつ，日本語文独自の特徴に対応できるような日本語句構造文法（JPSG）を提案している．しかし，JPSG では文節による構文の分析には限界があるとして，文節を取り入れずに句構造の理論を展開しているが，その場合は係り受け構造との互換性に無理が生じる．本研究では JPSG とは異なる手法として，係り受け構造を自然に表現できる句構造を構築することを条件としている．

3. 提案手法

3.1 句構造文法作成方法

句構造は隣接する語の間の関係に基づいて文の構造を表現したもので，通常，句構造文法の生成規則（句構造規則 phrase structure rule）によって定義される．句構造は語および文法的カテゴリ（非終端記号）を節点とする木構造で表される．句構造規則群を中心とする文法を句構造文法（phrase structure grammar）と呼ぶ [3]．すなわち，句構造文法とは終端記号・非終端記号・生成規則の 3 つの集合を指し，これを定めることが本研究の目的である．

文から句構造規則に矛盾しない句構造木を生成する操作を構文解析（parsing）と呼ぶ [3]．句構造解析では，1 つの入力文に対して複数の句構造解析木が出力されてしまい，明らかな誤りを含む場合があるため，これを回避することが課題の 1 つである．日本語文の場合，句構造解析木が直感的に正しいかどうかの判断基準は，係り受け関係の矛盾の有無にあるため，係り受け関係が自然に表現できている句構造解析木を正しいと判断する（図 1）．そのような正しい句構造木を生成できるような句構造文法の構築を目指す．

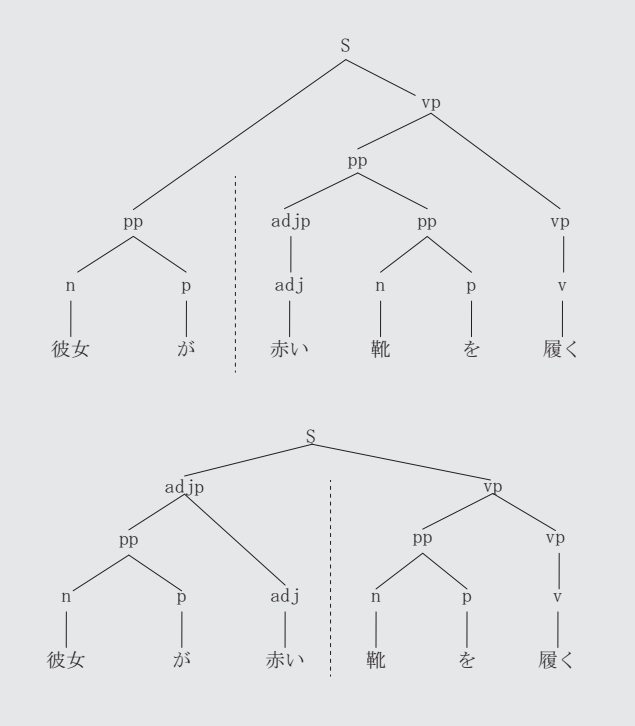

図 1: 例文「彼女が赤い靴を履く.」における自然な句構造木（上）と不自然な句構造木（下）．
上の木は「彼女が」と「赤い靴を履く」の間を文の切れ目としているのに対して，下の木は「彼女が赤い」と「靴を履く」の間を切れ目としている．

新たな日本語句構造文法を作成するために，まず正解の句構造木を作成し，その句構造木から句構造規則を抽出していく手法を取る（図 2）．正解の句構造木は，係り受け構造木を変換することで得られる，係り受け関係を保存した句構造木と定義する．一般に句構造木から係り受け構造木への変換は一意にでき，句構造木の各兄弟ノードにおいて，左ノードが右ノードに係るように変換することによって係り受け構造木を生成できる．句構造木を上記の方法で変換した結果，直感的に正しい係り受け構造木が得られるような句構造木を正解と定義する．これにより，修飾-被修飾関係における直感的に正しい句構造木が得られ，これまでの解析では相容れなかった句構造と係り受け構造を，矛盾なく結びつけることができる．

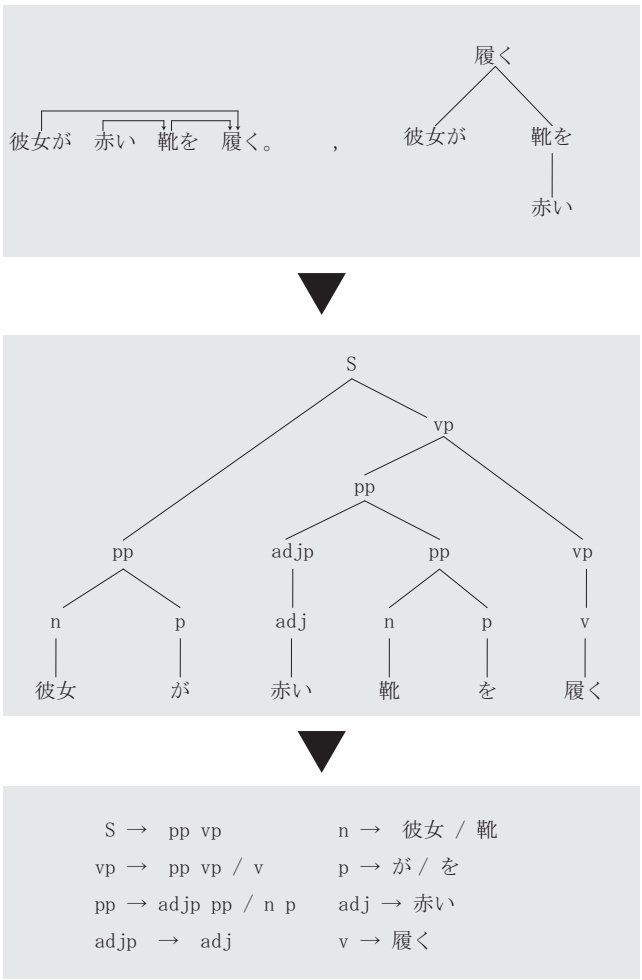

図 2: 例文「彼女が赤い靴を履く.」における係り受け木とそれをもとにした句構造規則の抽出

3.2 句構造木作成方法

句構造規則を作成するための途中過程として必要となる句構造木は以下の手順で作成する（図 3）.

句構造木作成手順

1) 葉ノードとして終端記号である単語を並べる.
2) 各終端記号に非終端記号である品詞記号（統語範疇）をつなげる.
3) 各文節について左側（文頭側）から順に 2 語ずつを兄弟としてつなげ，句記号（句範疇）をつける.
4) 係り受け構造に矛盾しないように，係り受け関係を示す枝をつけていく.
5) 4) において従属節の最右端の兄弟の非終端記号に “*” をつけたものを非終端記号として導入する.

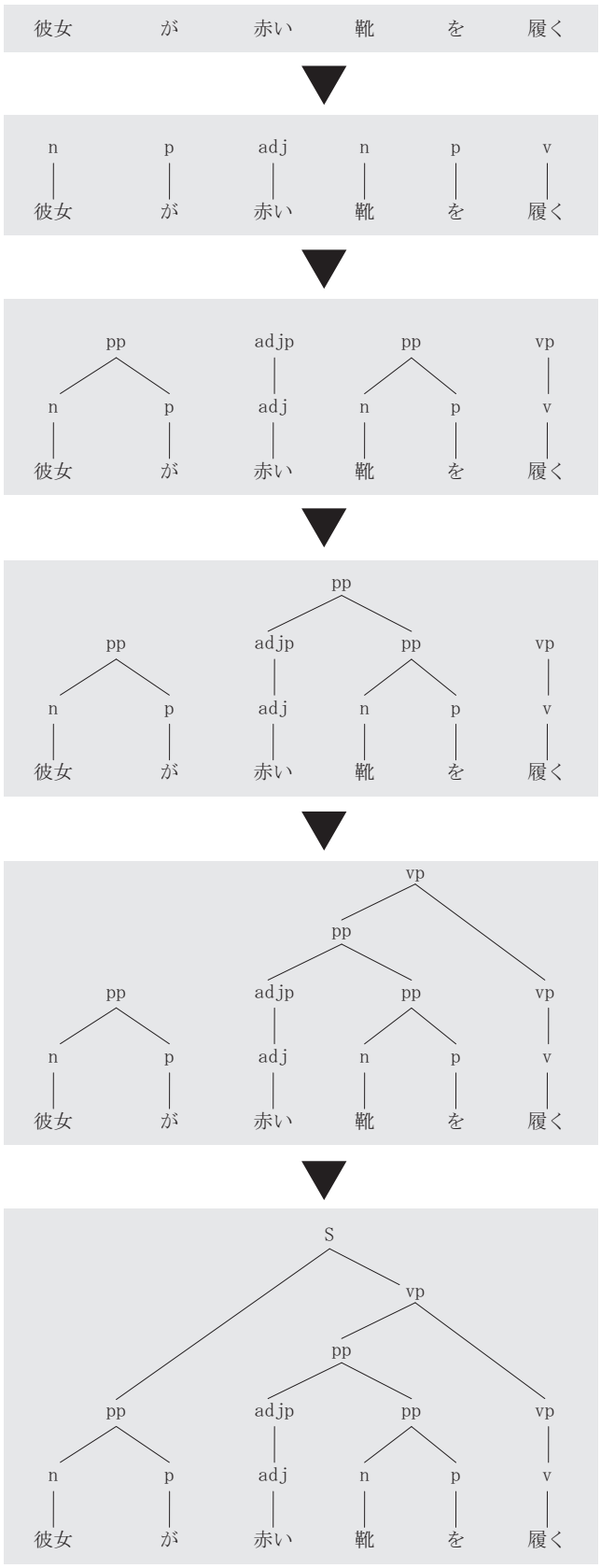

図 3: 「彼女が赤い靴を履く.」における句構造木の作成

手順 3 について，非終端記号である句記号は右側の非終端記号の頭文字に p をつけたものとする．ただし右側の非終端記号が助動詞や補文標識の場合は，左側の非終端記号の頭文字に p をつけたものを非終端記号とする．

また，一般に日本語文において，句構造木の各ノードは最右端の兄弟ノードに係るように，一意に係り受け構造に変換される．そのため，手順 4 の「係り受け構造に矛盾しないような係り受け関係を示す枝」とは,「変換によって人間が見て自然な係り受け構造木を生成できるような句構造木の枝」を指す．

また，手順 5 の「従属節の最右端の兄弟の非終端記号に “*” をつけたものを非終端記号として導入する」により，従属節であることを決定づける部分木のルートノードに “*” が記載される（図 4）．次節で示す非終端記号として明記されていない記号にも，条件を満たせば “*” がつけられることになる．

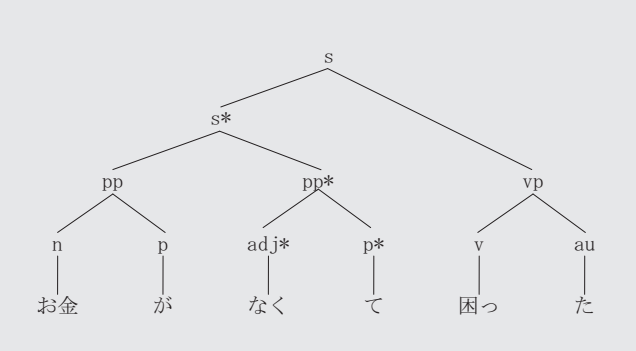

図 4: 「お金がなくて困った.」における句構造木

3.3 非終端記号

第 3.2 節で述べた句構造木作成手順で導入された非終端記号を表に示す（表 1）．基本的には，多くの言語の共通する文法指標である統語範疇と句範疇の範囲に収まるよう設定した．しかし，日本語独自の品詞のうち日本語文の構造において重要な役割を果たす品詞や，他の統語範疇に含めることができない品詞については，品詞そのものを記号として採用している．

表 1: 非終端記号一覧

区分	記号	内容	備考
節	s	文	開始記号
	s*	従属節	
句	dp	決定詞句	d などからなる.
	cp	接続詞句	c などからなる.
	np	名詞句	n(p)+n(p) などからなる.
	vp	動詞句	n(p)+v(p) などからなる.
	adjp	形容詞句	adj などからなる.
	advp	副詞句	adv などからなる.
	pp	後置詞句	n(p)+p(p) などからなる.
品詞	n	名詞	形容動詞語幹も含む.
	v	動詞	
	v*	動詞	文末以外の動詞
	adj	形容詞	
	-	形容動詞	名詞＋「だ」として扱う.
	adv	副詞	
	c	接続詞	
	d	決定詞	連体詞にあたる.
	p	後置詞	助詞にあたる.
	p*	後置詞	接続助詞と引用の助詞
	au	助動詞	
	au*	判定詞	助動詞「だ」とその活用形
	com	補文標識	

3.4 従属節決定方法

日本語文は，節と句の違いが曖昧であるため，句構造解析を行うためには従属節の定義と範囲を定めておく必要がある．本研究では，日本語文において述語となり得る品詞（動詞・形容詞・形容動詞語幹を含む名詞）が，文末以外の位置で以下に示す所定の条件を満たす場合，その語を従属節の述語と定める．そして，係り受け関係において，従属節の述語にかかる語句全体を従属節の範囲と定める（図 5）．

本研究における従属節の述語の条件は以下の通りである．

＜動詞＞

- 文末以外の全ての動詞は，従属節の述語（図 6）．

＜名詞＞（形容動詞語幹も含む）

- 文末以外の名詞で，直後に判定詞「だ」があれば，従属節の述語（図 7）．

＜用言全般＞（動詞・形容詞）

- 文末以外の用言で，直後に接続助詞か引用の助詞「と」があれば，従属節の述語（図 8）．
- 文末以外の連用形の用言は従属節の述語（図 9）．

- 文末以外の用言で，名詞を修飾していれば従属節の述語（図 10）.

また，本研究においては従属節の述語となり得る品詞については，1 文節以上で節としての存在を認める．1 文節の語句のまとまりについては，節より句として取り扱う方が直感に即しているが，

- 1 文節の動詞は節と認めたい場合が多い，（例文：「電話するのを忘れないでね.」）
- 節は句を含む，
- 1 文節の文は存在する，

という点を踏まえると，1 文節の従属節の存在を広義に認める方が自然であると言える.

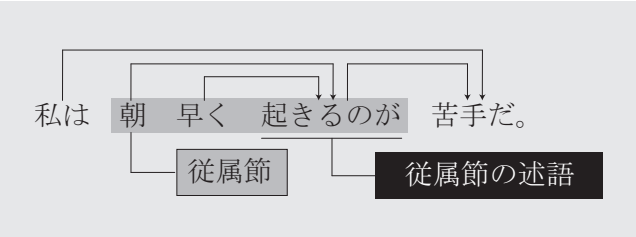

図 5: 例文「私は朝早く起きるのが苦手だ.」における従属節の範囲

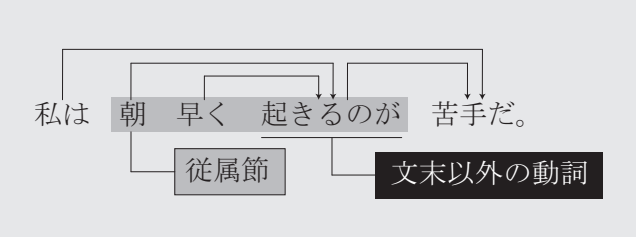

図 6: 例文「私は朝早く起きるのが苦手だ.」における従属節の述語

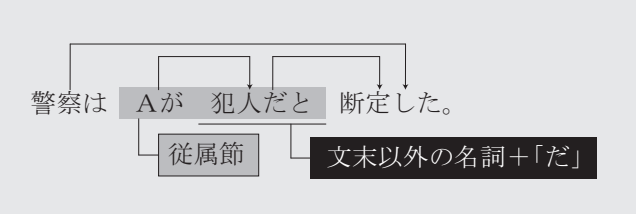

図 7: 例文「警察はAが犯人だと断定した.」における従属節の述語

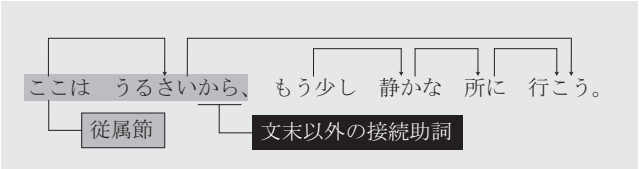

図 8: 例文「ここはうるさいから，もう少し静かな所に行こう.」における従属節の述語

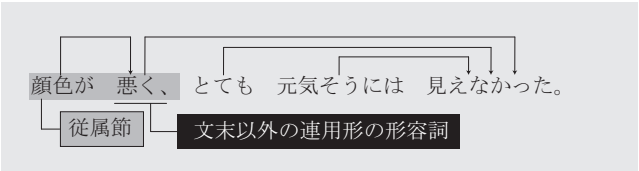

図 9: 例文「顔色が悪く，とても元気そうには見えなかった.」における従属節の述語

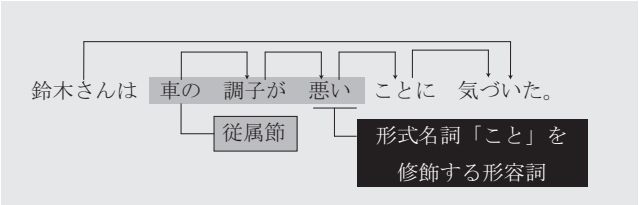

図 10: 例文「鈴木さんは車の調子が悪いことに気づいた.」における従属節の述語

4. 予備実験

4.1 実験手順

以下の手順で予備実験を行った.

1) サンプルデータ抽出
   BCCWJ（現代日本語書き言葉均衡コーパス）のベストセラーのうち，8~12 単語からなる約 2 万文からランダムに 100 文選択する.
2) 句構造規則生成
   1) の 100 文から句構造規則を生成する.
3) テストデータ抽出
   1) と重複の無い文を 1) と同様の手順で 20 文選択する.
4) 句構造解析
   2) の句構造規則で 3) の 20 文を句構造解析する.

各手順において使用した解析器は以下の通りである.

- 形態素解析器：MeCab ver.0.996 [4]
- 係り受け解析器：CaboCha ver.0.69 [5]
- 句構造解析器：NLTK ver.3.2.1 chart parser

今後の実験においてはCaboChaの係り受け解析結果に誤りがある場合は，それを正した上で句構造木を生成する予定であるが，今回の予備実験では行っていない．上記の手順3と手順4を繰り返すことで5回の実験を行い，作成した句構造規則によって何かしらの句構造解析木が出力として得られる確率である，解析可能率を測定した．

4.2 実験結果

予備実験によって，汎用的な句構造規則が作成できている可能性が高いという結果が得られた．5回の実験のうち，各回ごとの解析可能率は以下の通りである（表2）．なお，今回の実験で得られた句構造規則の個数は112個である．長谷川の規則 [1] が約500個であることに鑑みると，比較的少数の規則で網羅的に解析できていると言える．ただし，今回の句構造解析において，テストデータから得られた句構造解析木の数は多数あり，書き換え規則の制約を強めることは今後の重要な課題と位置づける必要がある．

表 2: 実験結果

回数	解析可能率
1回目	90.0 %
2回目	95.0 %
3回目	95.0 %
4回目	90.0 %
5回目	95.0 %
平均	93.0 %

5. まとめ

本稿では，これまで貧弱であった日本語の句構造文法について，係り受け構造を利用した新たな構築方法を提案した．予備実験において，提案手法を用いて構築した日本語句構造文法を用いて句構造解析をした結果，93.0 %の新たな入力文に関して解析可能であることが分かった．ただし今回は，係り受け構造に誤りがある場合を考慮していないため，句構造規則の精度が低い可能性があるため，今後高めていくべきである．また，入力文の長さを平均的なものに制限したことにより，高い成果が得られた側面もあるはずであり，この点は今後の課題とする必要がある．加えて，日本文の句構造としては適切でない木も生成してしまうことも問題であり，書き換え規則の制約の強め方について，検討を進めていくことが急務とされる．

参考文献

[1] 長谷川守寿. 日本語の句構造文法. **筑波応用言語学会研究**, Vol. 1, pp. 59–71, 1994.

[2] 郡司隆男. **日本語句構造文法**. 人工知能学会誌, Vol. 1, No. 2, 8 1986.

[3] 麻生英樹. 句構造と依存構造について. **The 26th Annual Conference of the Japanese Society for Artificial Intelligence**, 2012.

[4] 工藤拓, 山本薫, 松本祐治. Conditional random fields を用いた日本語形態素解析. Technical Report 47(2004-NL-161), 奈良先端科学技術大学院大学情報科学研究科／ NTT コミュニケーション科学基礎研究所リサーチアソシエイト, CREST JST ／東京工業大学, 奈良先端科学技術大学院大学情報科学研究科, 2004.

[5] 工藤拓, 松本祐治. 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**, Vol. 43, No. 6, pp. 1834–1842, 2002.

A　予備実験で生成された句構造規則（一部抜粋）

s → advp np
s → advp pp
s → advp vp
s → cp np
s → cp pp
s → cp vp
s → np np
s → np pp
s → np vp
s → pp adjp
s → pp np
s → pp pp
s → pp vp
s → s* adjp
s → s* np
s → s* pp
s → s* vp
s* → adjp p*
s* → adjp pp*
s* → advp pp*
s* → advp vp*
s* → n au*
s* → np pp*
s* → np* p
s* → pp pp*
s* → pp vp*
s* → pp* p
s* → v*
s* → v* p*

np → dp np
np → n
np → n au
np → n n
np → np au
np → np n
np → np np
np → pp n
np → pp np
np → s* np
np → v n
np → vp n
np* → v* n
np* → vp* n
vp → adj v
vp → adjp vp
vp → advp vp
vp → n v
vp → pp v
vp → pp vp
vp → s* vp
vp → v
vp → v au
vp → v v
vp → vp au
vp* → advp vp*
vp* → n v*
vp* → np vp*
vp* → pp vp*
vp* → v*
vp* → v* au
vp* → v* au*

pp → adj p
pp → adjp pp
pp → adv p
pp → dp pp
pp → n p
pp → np p
pp → np pp
pp → pp p
pp → pp pp
pp → s* pp
pp → v p
pp → vp p
pp* → adjp* p*
pp* → advp pp*
pp* → np* p
pp* → pp pp*
pp* → pp* p
pp* → v* p
pp* → v* p*
pp* → vp* p
pp* → vp* p*
adjp → adj
adjp → adj au
adjp → advp adjp
adjp → n adj
adjp* → pp* adj
advp → adv
cp → c
dp → d