統語情報と意味情報を統合した日本語句構造ツリーバンクの構築

田中 貴秋 ◇　永田 昌明 ◇　松崎 拓也 ♠　宮尾 祐介 ♠　植松 すみれ †

NTT コミュニケーション科学基礎研究所 ◇
国立情報学研究所 ♠
東京大学知の構造化センター †

{tanaka.takaaki, nagata.masaaki}@lab.ntt.co.jp◇
{takuya-matsuzaki, yusuke}@nii.ac.jp♠
uematsu@cks.u-tokyo.ac.jp†

1　はじめに

日本語の構文解析は，文節係り受けにより実用的な解析精度を実現し，多くの自然言語処理の基盤として発展してきた．この枠組みに基づいて，京都大学テキストコーパス（京都コーパス）[1] 等の様々な大規模コーパスが構築されている．しかしながら，これらのコーパスが持つ構文に関する情報は，基本的に文節間の依存関係の有無の情報に限られており，句や節などの統語機能に関する情報は付与されていない．例えば，単文における後置詞句と述語句との関係，複文における主節と従属節の区別，従属節の機能（副詞節，関係節，補足節等）等の情報である．これらの情報の一部は，Penn Treebank[2] 等では既に付与されている．

文節係り受けに基づくコーパスがこのような情報を持たないのは，文節，あるいは複数の文節の結合した単位が，統語的に機能を持つ単位とは必ずしも一致しないということが一因であると考えられる．またこの単位の不一致は，構文解析と，述語項構造解析・意味解析との親和性を下げることにもつながっている．典型的なものが，並列構造を含む場合であり，例えば,「赤い車と黒いバイクを買った」という文の述語項構造解析を考えると，述語「買った」の項として,「赤い車」「黒いバイク」名詞句の並列構造を対応づけたいが，これらの名詞句を文節の結合単位として抽出することはできない．

統語機能に関する詳細な情報を持つツリーバンクには，HPSG や CCG など語彙化文法に基づくものが存在する [3],[4],[5] ．これらから自動獲得した文法に基づく構文解析器は，強力な表現力を持つが，複雑なアノテーションを必要とするためツリーバンク構築のコストがかかるという点と，単一化に基づく解析機構が高い計算コストを必要とするという点も有する．

そこで我々は，比較的低いアノテーションコストで統語機能の情報を適切な単位に付与することができる Penn Treebank 様式の句構造情報と，それに対応した述語項構造情報を持つ日本語ツリーバンクの構築を進めている．また，本ツリーバンクは，Uematsu らの方法 [5] により CCG 導出木に変換することも可能である．本稿では，ツリーバンクの基本的な設計・構築方法および，構文解析器へ適用した結果について述べる．

2　ツリーバンクの設計

2.1　基本方針

構文木の構造は，機械処理での扱いやすさ，既存の文節係り受けに基づく言語資源からの変換のしやすさから，2 分木とする．用言と 1 語以上の付属語の連続の文節で表されるような句は右分岐の木構造とし，それ以外の句は左分岐の木構造とする．前終端記号は品詞の大分類を元にした定義したラベル（約 30 種類）を付与し，他の非終端記号は，句の文法カテゴリに基づいて定義し基本ラベルとして付与する（表 1 左）．

格情報や主節・従属節の情報を構文木に付加するために文法機能ラベルを導入する．文法機能ラベルを導入するに当たっては，吉本らの欅ツリーバンク [6] を参考にした．我々のツリーバンクの構造とは対照的に，欅は述語を中心としたフラットな構造を採用している．

述語項構造に関する詳細な情報は，句構造とは別形式で用意し，句 ID [1] により対応付けを行う．欅はゼロ代名詞を要素として明示しているが，本ツリーバンクでは既存の構文解析器への適用のしやすさを考慮し，句構造にはアノテーションせず，述語項構造情報にゼロの要素を記述する．

まとめると，本ツリーバンクは，構文木を含む句構造情報と，句 ID により対応づけられた述語項構造情報から構成される (図 1).

- 句構造情報： 構文木，表層格出現形，随意格 (一部)，節情報
- 述語項構造情報：表層格基本形 (ゼロ代名詞含む)，随意格，態 (格交替)

[1] 句構造の修正によって必要になる句 ID の変更を最小限にするため，句 ID は文 ID と句の文字範囲（開始位置と終了位置）の組み合わせにより表す．これにより，述語項構造との対応づけに与える影響を抑える．

NP	名詞句	-SBJ	主格
PP	後置詞句	-OBJ	対格
VP	動詞句	-OB2	与格
VNP	動詞句 (名詞+判定詞)	-TMP	時間格
ADJP	形容詞句	-LOC	場所格
ADVP	副詞句	-COORD	並列
		-APPOS	同格

表 1: 句のラベル / 文法機能のラベル

IP-MAT	主節
IP-ADV	副詞節
IP-REL	関係節（内の関係）
IP-REL_sbj	主名詞が主格
IP-REL_obj	主名詞が対格
IP-REL_ob2	主名詞が与格
IP-ADN	内容節・補充節（外の関係）
CP-NNF	補文（名詞化）
CP-THT	引用節
CP-QUE	疑問節

表 2: 節のラベル

2.2 句構造情報

構文木に文法機能ラベルを付加することにより，格，並列構造の情報を記述する．また，節に特定のラベル（節ラベル）を付与することにより，節の機能を記述する．

2.2.1 格

必須格（主格，対格，与格）の表層格出現形および随意格のうち場所格，時間格を，述語と格関係のある句に文法機能ラベルとして付与する．(表 1 右)．付与対象となるのは，主に後置詞句 PP や名詞句 NP である．関係節と主名詞間の格関係については後述する．

2.2.2 並列構造

2 分木を採用しているため，並列句は左分岐の構造となる．句のラベルに並列構造を示す文法機能ラベル -COORD(並列句) あるいは，-APPOS(同格句) を付加することにより，並列要素が明確になるようにする．並立助詞や接続詞，記号等を含む場合，左側の要素と並立助詞等を結合した構造に文法機能ラベルを付加する（図 1）．

2.2.3 節

主節と従属節に付与するラベルは表 2 のように定義する．日本語では，項が頻繁に省略されることにより句と節の区別が明確でないため，述語句の支配する最大投射の範囲を，IP あるいは CP を付与する対象と見なす．

副詞節　副詞節を含む複文の場合，主節が従属節を包含する構造でアノテーションする．以下の例では，副詞節を動詞句の「跳んで」が支配する「塀から〜跳んで」と見なして，節ラベル IP-ADV を付与し，主節「犬は驚いた」と副詞節を含んだ範囲に IP-MAT を付与している．

格ラベル	必須格	ARG0(ガ格), ARG1(ヲ格), ARG2(二格)
	随意格	TMP(時間格), LOC(場所格), EXT(外の関係), ARGM(その他)
	追加格	CAUS(使役主), BENEF(受益者)
態		ACT(能動態), PASS(受動態), POTN(可能態), BENEF(やりもらい)

表 3: 述語項構造情報

「塀から猫が跳んで　犬は驚いた」
(IP-MAT
(IP-ADV (PP 塀 から)
(VP (PP-SBJ 猫 が) (VP 跳ん で)))
(VP (PP-SBJ 犬 は) (VP 驚い た)))

連体修飾節　名詞句を修飾する節要素は，関係節（内の関係）による修飾関係とそれ以外を区別する．関係節の場合は，ラベル IP-REL を付与し，空所 (gap) になっている名詞の格を_sbj，_obj などのラベルにより明示する．図 1 は，主名詞が主格になる関係節の例になっている．外の関係になっている内容節（e.g.「さんまを焼く+においし」）と補充節（e.g.「家に帰る+途中」）とは境界が曖昧であるので，区別せず IP-ADN を付与する．

補足節　補語になる節要素は，補文 CP-NNF(e.g.「会場に着くのは午後だ」)，引用節 CP-THT(e.g.「早く帰りたいと思った」)，疑問節 CP-QUE(e.g.「いつ来るか知らない」) に分類し，ラベルを付与する．

2.3 述語項構造情報

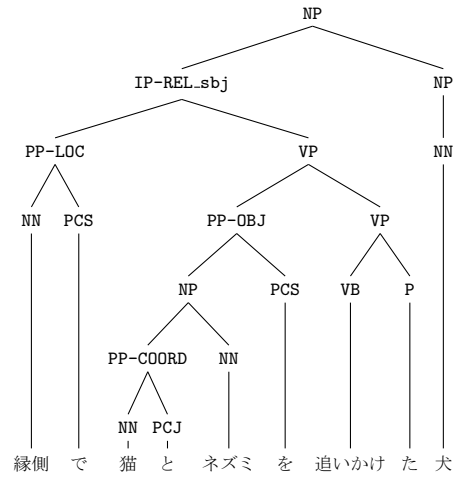
句構造情報と対応する形で，必須格の表層格基本形 (述語を能動態に直した時の形) および随意格，追加格，態に関する情報を持つ (表 3)．

項や述語の単位は単語ではなく句とし，句 ID により構文木中の句に対応する形で持つ．同一文内に該当する句がない場合，文外照応の場合は zero，外界照応の場合は，NAIST テキストコーパス (NTC)[7] の記法に準じて，exo1，exo2，exog を付与する．

また，各述語項構造について，別の辞書で定義される格フレームから適合するものを選択し付与する．同じ述語でも語義により適合する格フレームは異なる．本ツリーバンクでは，日本語語彙大系 [8] の構文体系辞書の情報を元に定義する．

3 ツリーバンクの構築

ベースとなるツリーバンクは，既存の言語資源の情報を活用して構築した．京都コーパスの文節係り受け構造を，句構造に自動的に変換したのち，品詞タグ情報と NTC の述語項構造情報,「と」コーパス [9] の格助詞「と」でマークされた項と述語の関係の情報，日本語語彙大系の格フレーム情報を参照して，句・節ラベルを自



voice: ACT
case frame ID: 024454

	phrase	phrase ID	marker
PRED	追いかけた	P9500000_009-013	
ARG0	犬	P9500000_013-014	adnom
ARG1	猫とネズミを	P9500000_003-009	を
LOC	縁側で	P9500000_000-003	で

図 1: 句構造情報と対応する述語項情報の例

動的に付与した．ベースのツリーバンクを人手で修正し，20,000 文のツリーバンクを構築した．

3.1 文節係り受け構造から句構造への変換

係り受け構造から句構造への変換は，(1) 文節内の構造の認定 (2) 文節間の係り受け関係から句の間の階層関係への変換の 2 段階に分けて行った．(1) では，文節内の形態素を 2 分木としてまとめ上げる．この際，体言を中心とした文節の場合は，中心となる複合名詞に右下がりの 2 分木の構造を与え，最後に助詞を右上がりの順に付加した．中心となる用言に助動詞や形式名詞が後続する形の文節の場合は，全体を右上がりの 2 分木の構造とすることを原則とした．(2) の係り受け関係から句構造への変換は，(1) で決定した各文節に対する部分木を，係り受け構造に従って組み合わせることで行う．この際，係り側の句の部分木を，受け側の部分木のどの位置に接合するかを決定する必要がある．この接合位置の決定は，係り側の句の種類，係り側の句の最右の形態素の活用形，受け側の木の接合位置の句の種類、京都コーパスの係り受け関係ラベル (並列句か否か) などを基にしたルールに従って行った．自動変換で作られた句構造は，人手で修正を行った．

3.2 句ラベルの変換

句のラベル（非終端記号+文法機能ラベル）は，以下のように付与した．

Tag set	LF_1	Comp	UF_1	Comp
Base	88.4	34.0	89.6	37.9
$\mathbf{Base_{inf}}$	88.5⋆	33.5	90.0⋆	39.3
Full	81.0	**15.6**	88.5	**37.3**
$\mathbf{Full_{inf}}$	**81.3**⋆	15.3	**88.8**	37.2
$\mathbf{Full_{lex}}$	80.3⋆	14.2	87.9⋆	33.6⋆
$\mathbf{Full_{vsup}}$	81.2	15.5	88.5	35.2
$\mathbf{Full_{vsup+alt}}$	77.9⋆	11.7⋆	86.0⋆	29.9⋆

表 4: 構文解析結果

まず，各単語に付与された品詞情報を変換テーブルに基づいて前終端記号に変換した．次に，簡単な CFG ルールにより上位の句のラベルをボトムアップに決定した．その後，述語の支配する最大範囲を節と認定し，節のラベルは，中心となる述語句構造の形や，句の末尾の付属語等に基づくルールより自動付与した．最後に，次節の述語項構造情報からの格の情報等を統合することにより後置詞句や関係節に付与する格関係に関する文法機能ラベルに反映させた．

3.3 述語項構造情報の統合

述語項構造は，NTC や「と」コーパスの情報から機械的に項候補を作成し，人手で修正した．最初に，NTC から “PRED” のラベルの付いている述語を抽出し，対応する句 ID を求める[2]．各項に当たる句は，NTC と「と」コーパスの情報から対応する句を抽出し，同一文内にある句について句 ID として記述した．文外照応，外界照応になっている項は，それぞれ記号 `zero`, `exo(1|2|g)` を付与した．また，その他，述語と依存関係にある句に含まれる項に関しては，時間格 `TMP`，場所格 `LOC`，その他の随意格 `ARGM` のいずれかを付与した．

4 構文解析への適用評価

本ツリーバンクを構文解析器 Berkeley Parser[10] の訓練データとして使用することにより評価を行った．京都コーパスの一部を対象として構築した 20,000 文のうち，14,895 文を訓練セット，1,860 文を評価セット，残りを開発セットとした．

構文解析器の訓練時に，前終端記号のラベルをそのまま使うだけではなく，ラベルに次のような付加情報を，構文解析に使用する情報として加えたものでも評価した：用言の活用形の追加 ($\mathbf{_{inf}}$)，助詞，助動詞の語彙化 ($\mathbf{_{lex}}$)，用言の下位範疇化情報 ($\mathbf{_{vsub}}$)，下位範疇化情報+格交替の情報 ($\mathbf{_{vsub+alt}}$)．

また，ラベルの詳細度の影響を確認するため，基本ラベルセット **Base** (`IP`, `CP` 等の節ラベルおよび，文法機能ラベルを除くラベルを使用) と全ラベルセット **Full** の 2 種類のラベルセットで評価した．

[2]事態性名詞は対象外とした．

Tag	P	R	F_1	Tag	P	R	F_1
PP-SBJ	69.6	81.5	75.1	IP-REL_sbj	48.4	54.3	51.1
PP-OBJ	72.6	83.5	77.7	IP-REL_obj	27.8	22.7	24.9
PP-OB2	63.6	71.4	67.3	IP-REL_ob2	17.2	29.4	21.7
PP-TMP	45.0	48.0	46.5	IP-ADN	50.9	55.4	53.1
PP-LOC	21.3	15.9	18.2	CP-THT	66.1	66.6	66.3

表 5: 文法機能ラベルの判別結果（格 / 節）

句構造解析の精度を見る前に，文節係り受けの解析と比較するため，解析結果の句構造木を文節係り受け相当に変換[3]して，ラベルなしの依存構造の精度を算出した．その結果，$\mathbf{Base_{inf}}$ で 89.4，$\mathbf{Full_{inf}}$ で 88.5 となり，工藤らの文節係り受けの精度 [11] 90.46 には及ばないものの，同等程度の水準にあると考えられる．

句構造解析の評価結果を，表 4 に示す．付加情報については，活用形情報は効果が見られたが，他の情報は副作用による精度低下が生じている．以下では，$\mathbf{Base_{inf}}$，$\mathbf{Full_{inf}}$ のセットを用いた結果について述べる．F 値は **Base** では，ラベル付き (LF_1) で 88，ラベルなし (UF_1) で 90 程度となっている．ラベル有無であまり差がないのは，**Base** のレベルでは，木の形に決定すると句カテゴリラベルの曖昧性が小さいからだと考えられるが，**Full** では，文法機能ラベル，節ラベルの誤りにより，LF_1 は UF_1 より 7-8 ポイント程度低下している．

文法機能ラベルのうち格の判別に関する結果を，表 5 左に示す[4]．後置詞句 (PP) のうち，表層格は格助詞による判別が可能なものが多数あるために，適合率で 60-70 程度はあるが，時間格，場所格については，かなり低い値となっている．特に場所格は，項に入る名詞のバリエーションが多く，誤りの 85%が PP あるいは PP-OB2 への誤判別であるため，名詞の分類情報を使用することが必要であると考えられる．

また，主な節ラベルの判別結果を，表 5 右に示す連体修飾節と被修飾名詞の関係ラベルの判別は，比較的精度の高い IP-REL_sbj(主格空所の関係節）や IP-ADN(内容節・補充節) でも，F 値で 50 程度と低い．誤りの内訳を見ると，IP-REL_sbj に関しては，句構造の誤りが 36%，IP-ADN と誤判別したものが 22%，VP が 19%，IP-ADN に関しては，句構造の誤りが 52%，VP と誤判別としたものが 22%，IP-REL_sbj が 18%であった．VP と誤判別したものは，対象の句がさらに別の句 (PP 等）を項として取り込むと判断したものが主な原因であり，これも句構造の誤りといえる．IP-REL_sbj と IP-ADN 間の判別は，格関係があるのか人間でも判断が難しいものも多く (e.g.「冷静で感情に左右されない+冷めた視点」)，アノテーションの判断基準を明確にする工夫も必要だと考えられる．

[3]各文節を含む最小の句を抽出し，抽出された句から共通の句を親に持つ句を選び，左の句に含まれる文節から右の句に含まれる文節に係り受け関係を付与することで変換した．

[4]誤りには，句構造（句の範囲）は正解で句のラベルのみが不正解のものと，句構造そのものが不正解であるものの 2 種類が存在する．

5 おわりに

句構造とそれに対応付けられた述語項構造情報を持つ，統語情報と意味情報を統合した日本語ツリーバンクについて述べた．構築した 20,000 文を既存の構文解析器に適用することにより，文節係り受け解析と同程度の解析精度を持ちつつ，詳細な統語情報を出力できることを確かめた．出力する文法機能ラベルの精度向上のため，アノテーションの改良や格フレーム情報の利用等について検討する予定である．

参考文献

[1] Sadao Kurohashi and Makoto Nagao: Building a Japanese parsed corpus – while improving the parsing system. In Abeille (ed.), Treebanks: Building and using parsed corpora, Chap. 14, pp. 249–260. Kluwer Academic Publishers (2003)

[2] Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz: Building a large annotated corpus of English: the Penn Treebank, In *Journal of Computational Linguistics*, Vol.19, No.2, pp. 313–330 (1993)

[3] Yusuke Miyao, Jun-ichi Tsujii: Feature forest models for probabilistic HPSG parsing. In *Journal of Computational Linguistics*, Vol.34, No.1, pp. 35–80 (2008)

[4] Francis Bond, Sanae Fujita and Takaaki Tanaka: The Hinoki syntactic and semantic treebank of Japanese, In *Journal of Language Resources and Evaluation*, Vol.42, No. 2, pp. 243–251 (2008)

[5] Sumire Uematsu, Takuya Matsuzaki, Hiroaki Hanaoka, Yusuke Miyao and Hideki Mima: Integrating multiple dependency corpora for inducing wide-coverage Japanese CCG resources, In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pp. 1042–1051 (2013)

[6] 吉本啓，周振，小菅智也，大友瑠璃子，Alastair Butler: 日本語ツリーバンクのアノテーション方針, 言語処理学会第 19 回年次大会予稿集, pp.924–927 (2013)

[7] 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：NAIST テキストコーパス構築からの経験から，自然言語処理, Vol. 17, No. 2, pp. 25–50 (2010)

[8] 池原悟，宮崎正弘，白井諭，横尾昭男，中岩浩巳，小倉健太郎，大山芳史，林良彦:日本語語彙大系, 岩波書店, (1997)

[9] Hanaoka, H., Mima, H., and Tsujii, J.: A Japanese particle corpus built by example-based annotation, In *Proceedings of the 7th international conference on Language Resources and Evaluation (LREC 2010)* (2010)

[10] Slav Petrov, Leon Barrett, Romain Thibaux and Dan Klein: Learning accurate, compact, and interpretable tree annotation, In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pp. 433-440 (2006)

[11] Taku Kudo and Yuji Matsumoto: Japanese dependency analysis using cascaded chunking, In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, Volume 20, pp. 1–7 (2002)