

構造化チャート法に基づく日本語構文解析器 schartj*

宮崎 正弘^{†‡} 川辺 諭[‡] 武本 裕[†]

[†]新潟大学大学院自然科学研究科 [‡]株式会社ラングテック

1 はじめに

構造を含む生成規則を扱える拡張型のチャートパーザである schart パーザ [1] を用いた日本語構文解析器 schartj を提案する。schartj パーザは CFG 文法規則を Lisp 言語の S 式形式で記述することで、生成規則の右辺に部分木構造を埋め込むことができ、不要な部分木の生成を抑止し、解析時に発生する構造多義を削減することができる。schartj は時枝誠記の文法理論を発展的に継承した三浦文法に基づいた、意味との親和性が高い構文木を出力する。日本語形態素解析器 jampar¹によって分かち書きされた形態素情報を用いて節の境界を推定し、複文や重文における接続表現から節同士の依存関係を推定することで、日本語構文解析時に発生する構造多義を抑制する。

2 三浦文法に基づいたパーザ

2.1 三浦文法の言語モデル

時枝誠記 [2] は、言語表現は“対象 ⇄ 認識 ⇄ 表現”という過程的な構造をその背景に持っており、話者の対象を認識するという作用を通じて言語が表現されるという“言語過程説”を提唱した。三浦つとむ [3] は、時枝の言語過程説とそれに基づく日本語品詞体系を発展的に継承し、言語表現の持つ意味とは言語表現自体が持っている対象と認識との客観的な関係であるとする“関係意味論”を提唱した。

言語過程説によれば、話者は対象世界を概念化された実体や属性、またそれらの関係の集合体として捉える。そこに話者の主体的な判断を加えることによって、対象世界の認識構造を形成する。さらに認識構造を構成している実体や属性、関係に言語規範を適用することで、言語表現が生成される。その際、認識構造を再帰的に捉え直し主体的な判断を次々に加えることで、

観念的に多重化した世界を表現することができる。時枝は言語を話者の主観的な判断を直接に表現する主体的表現の語(辞)と、対象を概念化して表現する客体的表現の語(詞)に分類し、文は主体的表現が客体的表現を包み込んだ形の入れ子構造で表されるとした(図1)。

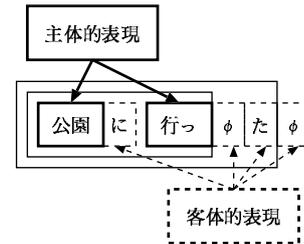


図1: 文の入れ子構造

2.2 日本語形態素の品詞体系

本パーザの日本語の品詞体系は、三浦文法に基づいた品詞体系 [4] を若干修正したものを採用している。品詞体系の特徴としては、単語を客体的表現である詞と主体的表現である辞に分類していることがあげられる。具体的には詞としては名詞、動詞、形容詞、連体詞、副詞、接辞を、辞としては助詞、助動詞、陳述副詞、接続詞、感動詞を分類している。学校文法とは、接続助詞「て、たり」「で、だり」を既定判断の助動詞「た」「だ」の連用形としていること、形容動詞を状態名詞に肯定判断の助動詞「だ」や格助詞「に」が結合したものであることなどが異なる。

2.3 構文木シンボル

本パーザで用いる構文木シンボルは、文や節、句を示す非終端記号を 12 種類、形態素(とその相当語)や記号を示す前終端記号を 21 種類使用している。終端記号としては、日本語形態素解析器 jampar によって分割された形態素データを使用している。

*“エスチャートジェイ”と読む。S は構造化 (structured) の意。
1 “ジャンパー”と読む。JApanese Morphological PARser。

3 意味と親和性のある構文解析の例

本パーザは意味と親和性のある統語構造を出力する [5]。解析例を以下に示す。

3.1 状態性名詞

三浦文法では形容動詞語幹+活用語尾 (例「熱心な」) を状態性名詞+肯定判断の助動詞「だ」として扱う。また schartj では、これらの状態性名詞が格要素を伴う場合は、節として部分木を構成する。図 2 では「研究に熱心な」の部分が節「研究に熱心だ」として判定され、名詞「学生」を修飾している。

```
|-ss
|-c1
| |-np
| | |-cx
| | | |-c1
| | | | |-pp
| | | | | |-np
| | | | | | |-n-< @NVS:研究 >
| | | | | | |-pP-< @pP13:に >
| | | | | | |-np
| | | | | | | |-na-< @NAD3:熱心 >
| | | | | | | |-x-< @xa17:な >
| | | | | | |-np
| | | | | | | |-n-< @NNG:学生 >
| | | | | | |-sp-< @SP1:. >
```

図 2: 「研究に熱心な学生。」

3.2 連体格助詞「の」

連体格助詞「の」に関しては、文節が修飾する名詞の範囲を正しく判定する。図 3 では「梅の花」という名詞句が意味的に正しく解析されている。

```
|-ss
|-c1
| |-pp
| | |-np
| | | |-ppT
| | | | |-np
| | | | | |-n-< @NNG:梅 >
| | | | | | |-pT-< @pP21:の >
| | | | | | |-np
| | | | | | | |-n-< @NNG:花 >
| | | | | | | |-pP-< @pP11:が >
| | | | | | |-vp
| | | | | | | |-v
| | | | | | | | |-@V5K6
| | | | | | | | | |-< @V5K0:咲 >
| | | | | | | | | |-< @i5K6:< >
| | | | | | |-sp-< @SP1:. >
```

図 3: 「梅の花が咲く。」

3.3 格助詞を伴う用言性名詞

動詞連用形から転成した名詞は、転成前の動詞とその格要素を節として判定し、部分木として構造化する。図 4 では「魚を釣り」の部分が節「魚を釣る」として判定されている。

```
|-ss
|-c1
| |-pp
| | |-c1
| | | |-pp
| | | | |-np
| | | | | |-n-< @NNG:魚 >
| | | | | | |-pP-< @pP12:を >
| | | | | | |-vp
| | | | | | | |-v
| | | | | | | | |-@V5R3
| | | | | | | | | |-< @V5R0:釣 >
| | | | | | | | | |-< @i5R3:り >
| | | | | | | | |-pP-< @pP13:に >
| | | | | | | |-vp
| | | | | | | | |-v
| | | | | | | | | |-@V5Z6
| | | | | | | | | | |-< @V5Z0:行 >
| | | | | | | | | | |-< @i5Z6:< >
| | | | | | | | |-sp-< @SP1:. >
```

図 4: 「魚を釣りに行く。」

3.4 係助詞「は」と助動詞のスコープ

係助詞「は」や助動詞に関しては、指示対象 (スコープ) を正しく反映した木構造を出力する。図 5 では係助詞「は」、助動詞「た」のスコープが正しく解析されている。

```
|-ss
|-cx
| |-cx
| | |-pp
| | | |-np
| | | | |-n-< @NNP:太郎 >
| | | | | |-pW-< @pK11:は >
| | | | | | |-cc
| | | | | | | |-c1
| | | | | | | | |-pp
| | | | | | | | | |-np
| | | | | | | | | | |-n-< @NNG:山 >
| | | | | | | | | | | |-pP-< @pP12:を >
| | | | | | | | | | | |-vp
| | | | | | | | | | | | |-v
| | | | | | | | | | | | | |-@V5R3
| | | | | | | | | | | | | | |-< @V5R0:下 >
| | | | | | | | | | | | | | | |-< @i5R3:り >
| | | | | | | | | | | | | | |-c1
| | | | | | | | | | | | | | | |-pp
| | | | | | | | | | | | | | | | |-np
| | | | | | | | | | | | | | | | | |-n-< @NNG:村 >
| | | | | | | | | | | | | | | | | | |-p-< @pP13:に >
| | | | | | | | | | | | | | | | | | |-vp
| | | | | | | | | | | | | | | | | | | |-v
| | | | | | | | | | | | | | | | | | | | |-@V5K4
| | | | | | | | | | | | | | | | | | | | | |-< @V5K0:着 >
| | | | | | | | | | | | | | | | | | | | | | |-< @i5K4:い >
| | | | | | | | | | | | | | | | | | | | | | |-x-< @xp16:た >
| | | | | | | | | | | | | | | | | | | | | | |-sp-< @SP1:. >
```

図 5: 「太郎は山を下り村に着いた。」

4 構文解析処理

構文解析処理を実行する前に、形態素の品詞や字面、用言の必須格などの情報を用いて、部分木の境界推定、節同士の係り判定、格要素の係り先判定、呼応関係に基づく係り先判定などの前処理を行う。構文解析時にこれらの情報を参照することで、構造的な曖昧性の発生を抑止する。

4.1 部分木境界推定処理

以下の位置に部分木の境界を推定する [6]。推定された境界はスラッシュ記号 “/” で示している。

- 節の接続語となる語の直後
 - 用言の連用中止形
例：山を 下り/村に着いた。
 - 助動詞の連用中止形、仮定形
例：ちょっと横になったら/疲れが取れた。
 - 接続助詞、接続助詞相当の形式名詞
例：ちょうど私が出る とき/彼が着いた。
- 遠くに係りやすい語の直後
 - 係助詞「は」
例：彼 は/疲れていたが手伝いに行った。
 - 文頭の副詞型名詞
例：来月/沖縄へ行く予定だ。
- 節全体に係る語の直前
 - 節の切れ目となる助動詞
例：台風が通過し/て 風雨が収まった。

チャート法による解析処理時は、これらの処理で推定された境界によって分割されたセグメントに関して、セグメントをカバーする不活性弧以外の弧が、境界をまたいで弧を生成しないように抑止する。

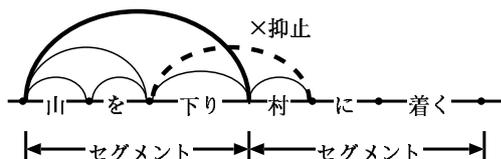


図 6: 部分木境界による制約

図 7 では「下り」と「村」はセグメントをカバーする弧ではないので、境界をまたいだ「下り村」という弧は生成されない。

4.2 節係り判定処理

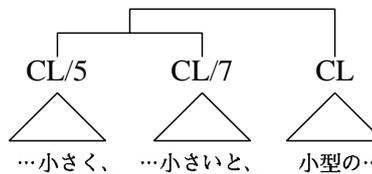
3 節以上からなる重文、複文の構造的曖昧性を解消するために、節間の接続優先度を用いて節同士の係り関係を判定する [6]。

表 1: 節間接続優先度

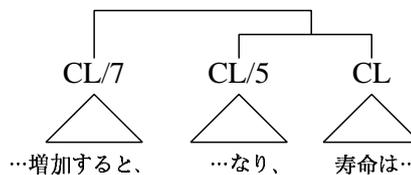
接続優先度	節の分類
7	主節
	展開の接続助詞 + 読点
6	展開の接続助詞
	副助詞「は」 + 読点
5	条件の接続助詞 + 読点
	連用中止形 + 読点
	用言の仮定形 + 読点
	体言止め + 読点
	副助詞「は」
	副詞句 + 読点
4	条件の接続助詞
	連用中止形
	用言の仮定形
	体言止め
3	同時の接続助詞 + 読点
2	同時の接続助詞

節末の接続表現に着目して表 1 に示した接続優先度を与える。チャート法による解析処理時は、それぞれの節が自分の優先度よりも高い優先度を持つ節を飛び越えて、遠方の弧に係ることがないように抑止する。

- 例 1: 「インゴットサイズが 小さく、(5) かつ流れ応力が 小さいと、(7) 小型の圧延機が使える。」



- 例 2: 「抵抗が増加する と、(7) 光は暗く なり、(5) 寿命はのびる。」



4.3 格要素の係り先判定処理

複数の用言にかかる可能性がある格要素に関しては、用言の必須格となる助詞を用いて係り先を判定する。この場合、意味的制約のチェックは行わない。格要素に係る可能性がある用言が複数ある場合や、必須格では判定できない自由格成分は、直近の用言を係り先とする(ただし係助詞「は」を用いた格要素は例外として、直近を係り先にはしない)。

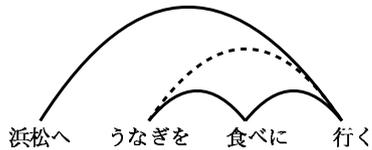


図 7: 格要素の係り先判定処理

図 7では格要素「うなぎを」が2つの用言「食べる」「行く」の格要素となり得るので、直近の「食べる」に係るものとする。

チャート法による解析処理時は、節を構成する格要素と用言に同一の識別子を付与し、識別子が異なる不活性弧を含む弧が生成されないように抑止する。

4.4 呼応に基づく係り先判定

呼応とは、先行する語に対して特定の語が要求される関係のことである。ここでは、語が使われる確実性が非常に高いものを強い呼応、比較的高いものを弱い呼応、呼応する語同士の距離が比較的近いものを局所的な呼応、遠いものを大域的な呼応と定義する。

局所的な呼応の例としては、「油を売る」のように格要素の主名詞が特定の語である強い呼応や、「辞書を引く」「辞書を引く」のような弱い呼応がある。大域的な呼応の例としては、「決して～ない」のように陳述副詞を用いた強い呼応や「～とは～だ」のような弱い呼応がある。

呼応に関しても前節の格要素と同様に係り先判定を行う。すなわち、節を構成する格要素や述語に共通の識別子を付与することで、異なる識別子の不活性弧を含む弧の発生を抑止する。

5 おわりに

三浦文法に基づいた意味と親和性の高い日本語構文木を出力する日本語構文解析器 schartj を実装した。

schartj では従来の係り受け解析の手法では意味的に妥当な統語構造を出力することが難しかった状態性名詞や連体格助詞、用言性名詞などに関して、正しい構文木を出力することが可能となった。また係助詞「は」や助動詞のスコープを正しく判別することが可能となった。

構文解析の内部実装に関しては、複数の節をふくむ重文、複文などで生じる構造多義を抑制するために、チャート解析の前段階で形態素の字面・品詞の情報などを用いて、構文解析補助情報を付与する。具体的には部分木境界推定や節係り判定を行うことで、節同士の関係を推測する。また、格要素の係り先判定や呼応に基づく係り先判定を行うことで、節のスコープを推定する。

今後、用例を利用するなどして格要素のスコープをより正しく判定し、解析正解率を向上させることを検討している。schartj のより詳細な情報は <http://www.languetech.co.jp/> から得られる。

参考文献

- [1] 川辺、宮崎：構造を含む生成規則を扱える拡張型チャートパーザ – Schart パーザの実装 –、言語処理学会第 11 回年次発表論文集、pp. 911~914 (2005).
- [2] 時枝誠記：日本文法 口語篇、岩波全書 (1950).
- [3] 三浦つとむ：日本語とはどういう言語か、講談社学術文庫 (1976).
- [4] 宮崎、白井、池原：言語過程説に基づく日本語品詞の体系化とその効用、自然言語処理 Vol.2 No.3、pp. 3~25 (1995).
- [5] 武本、宮崎：意味と親和性のある統語構造を出力する日本語文パーザ、自然言語処理 Vol.14 No.1 (2007).
- [6] 武本、宮崎：局所・大域的制約による構造的曖昧性抑制機構を持つ日本語文パーザ、言語処理学会第 12 回年次発表論文集、pp. 137~140 (2006).