E-25

文節構造解析システムibukiBについて

Bunsetsu Structure Analysis System ibukiB

伊佐治 和哉† 岸井 謙一† 池田 尚志†
Kazuya Isaji Ken-ichi Kishii Takashi Ikeda

1 はじめに

日本語には文節という構文単位がある．文節は自立語と機能語からなり，文は文節の列からなる．

我々は，形態素・文節解析システムibukiKを開発している．また，ibukiKが出力する文節の機能語部をさらに解析し，意味的・機能的な観点から機能語部をいくつかの要素に分割して出力する，文節構造解析システムibukiBを構築した．これらについて報告する．

2 システムの概要

構築したシステムの概要を図1に示す．

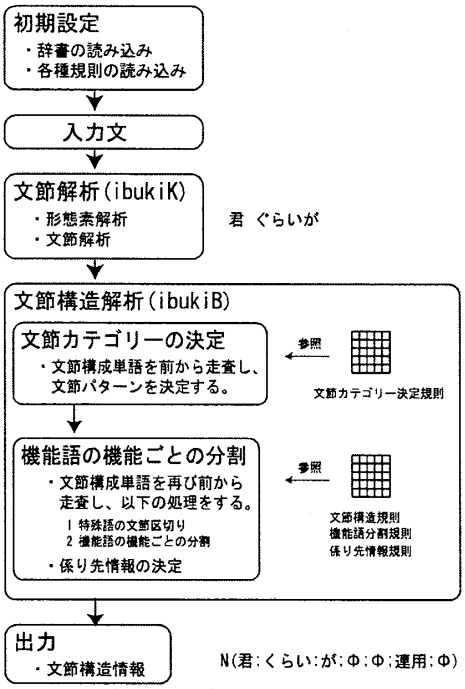

図1 文節構造解析システムの概要

システムが出力する結果の例を図2に示す．文節構造は「文節カテゴリ（主に品詞）」,「自立語」，および「機能語部を機能ごとに分割した要素」から構成される．

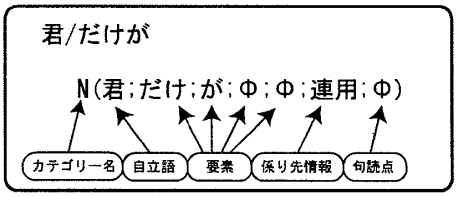

図2 文節構造情報

3 ibukiK

我々の開発している，日本語文節解析システムibukiKでは，文節として可能性のあるものをすべて求めた上で，コストに基づく解析を行い，解を導き出している．

形態素解析で一般的に用いられているコスト最小法では，個々の単語に与える単語コストと隣接する単語の接続に対する連接コストを用いて，総コストの少ない単語列を優先解として出力する（単語単位の手法）．これに対し，我々のシステムでは，個々の単語および単語間にコストを与えるのではなく，文節および隣接する文節間にコストを与えることとしている（文節単位の手法）．

4 ibukiB

2節で述べた文節構造の定義に沿って，文節構造解析システムを構築し，以下のことを実現した．

- 文節カテゴリーを与える．
- 機能語部を機能ごとに分割する．
- 「ぐらい」「くらい」などの字面上は違うが意味が一緒のものは，統一して一般化を行う．
- 形式名詞，判定詞を含み構成が複雑になっている文節を分割することにより，文節の統語的役割を明確にする．

4.1 文節カテゴリーの決定

文節構造解析システムでは，まず文節に文節カテゴリーを与える．文節カテゴリーは主に品詞を表し，以下の15種類を設けた．（表1，表2，表3）

表1 体言系文節（5種類）

N	名詞文節
SN	ノ系文節
KA	カ系文節
Q	」文節（引用の終わり）
TO	ト系文節

表2 用言系文節（4種類）

P1	動詞文節
P2	タ系文節
P3	形容詞文節
P4	形容動詞文節

表3 その他の文節（6種類）

A	副詞文節
T	連体詞文節
C	接続詞文節
I	感動詞文節
QF	「文節（引用の始まり）
UN	未知語文節

文節カテゴリーは，ibukiKの文節解析情報から，文節を構成する全単語の左連接属性及び，右連接属性を元に決定している．

4.2 機能語部を機能ごとに分割

機能語部の分割は，意味的・機能的な観点からの分割であるが，語順を保たせている．その定義を体言系は表4，用言系は表5に示す．

†岐阜大学工学部

表 4　体言後接機能語部の分割

	分類	例
要素 1	格助詞相当語に前接する副助詞等	だけ，すら，さえ
要素 2	格助詞相当語	に，を，で
要素 3	格助詞相当語に後接する副助詞等，ノ格	に，だけ，の
要素 4	提題助詞	は，も，はまた

表 5　用言後接機能語部の分割

	分類	例
要素 1	受身，使役等の助動詞	させる，られる
要素 2	時制，肯否等の助動詞	た，ている，ない
要素 3	判断等の助動詞	だ，だろう，らしい
要素 4	接続助詞	が，のに，ので

文節には複数の意味が含まれているものもあり，以下の場合，上記の定義では対応できない．そこで，元は同じ文節であるという情報を保持したまま文節を分割することで対処した．

- ノ系

  例：君+の (ノ格)+だけ (副助詞)+に (格助詞)+は (提題助詞)　→　(君／の)(の／だけ／に／は)

  この文節は「君の (もの) だけには」という意味を持っているが，「もの」を省略している．これを省略の「の」とし「ノ系」と定義した．「もの」が省略されていなければ「もの」の前で区切られ表 4 で対処できる．そこで，省略の「の」を含む文節は文節区切りを行うことにした．

- ダ系

  例：君＋だけ (副助詞) ＋だ (判定詞) ＋た (時制) ＋が (接続)　→　(君／だけ) (だった／が)

  このように体言文節に判定詞「だ」が後接すると，用言文節のような名詞述語化文節となる．このような体言文節後接機能語の「だ」を「ダ系」と定義する．そして名詞文節とダ系文節に分割した．

- 形式名詞

  例：助ける+こと (形式名詞)+が (格助詞)　→　(助ける)(こと／が)

  このように用言文節に形式名詞が後接すると述語名詞化文節となる．そこで，文節区切りを行い，形式名詞以下を体言として扱うことで対処した．

- カ系

  例：君+かどうか (疑問)+が (格助詞)　→　(君)(かどうかが)

  「か」「かどうか」などはダ系の疑問形と考えることが出来るが，名詞化する場合があることより，カ系文節として扱った．そして「か」「かどうか」の後に体言後接語が続く場合は文節を区切ると定義した．

4.3　係り先情報の決定

係り先情報はその文節がどのような文節に係っていくかという情報であり，表 6 のように１２種類を設けた．

「並列／連用／疑問」は，並列か連用か疑問のいづれかの属性になるという曖昧な場合である．このような場合には，文節情報だけでは，係り先の文節カテゴリーは一意に決まらない．

表 6　係り先情報

連用	連体	独立	並列
仮定	命令	文末	並列／連用
並列／連用／疑問	ダ系	ノ系	カ系

4.4　簡単な一般化

文節構造解析では簡単な一般化という作業を加えた．一般化とは「ぐらい」「くらい」や，「にかんし」「にかんして」「に関して」などの字面は違うが基本的には同じ単語である機能語に，同じ表記を与えることをいう．

これは，意味的に同一な文節構造を同一化するための処置である．

4.5　解析結果の例

文節構造解析システムが出力する解析結果の出力例を以下に示す．解析結果の１つ目のフィールドは文節番号，２つ目のフィールドは文節区切りを行ったときに用いるサブ文節番号を表す．

- 私が心配したところでどうにもならない．

```
0 0 N(私; Φ; が; Φ; 連用; Φ)
1 0 P1(心配; Φ; た; Φ; Φ; 形式名詞; Φ)
1 1 SN(ところ; Φ; で; Φ; Φ; 連用; Φ)
2 0 A(どうにも; Φ; Φ; Φ; Φ; 連用; Φ)
3 0 P1(なる; Φ; ない; Φ; Φ; 文末;. )
```

- 他人のは良くみえるものだ．

```
0 0 N(他人; Φ; Φ; Φ; Φ; ノ系; Φ)
0 1 N(の; Φ; Φ; Φ; は; 連用; Φ)
1 0 A(良く; Φ; Φ; Φ; Φ; 連用; Φ)
2 0 P1(みる; える; Φ; Φ; Φ; 形式名詞; Φ)
2 1 SN(もの; Φ; Φ; Φ; Φ; ダ系; Φ)
2 2 P2(だ; Φ; だ; Φ; Φ; 文末;. )
```

5　おわりに

文節の機能語部を意味的・機能的な観点から解析する，文節構造解析システム ibukiB を構築した．これによって，文節の構造を詳しく，分かりやすく，表現することが可能となった．

ibukiB を大規模コーパスに適用して，機能語部のパターンや文節列のパターン（文パターン）の出現分布について調査してみる予定である．また ibukiB の解析結果を用いる構文解析システム（係り受け解析システム）を構築する予定である．

機能語部のパターンの統計に関しては，意味的な観点でまとめれば（標準形に置き換えれば）機能語部の要素数は現実的に有限の数に収まるのではないかと期待している．そのように収束すれば，パターン変換型機械翻訳や，パターン照合による新しいタイプの構文解析の可能性について検討することが出来る．

参考文献

[1] 文節の内部構造統計と出現頻度統計，一ノ瀬，池田，言語処理学会 第 8 回年次大会 発表論文集 (2002)