日本語文解析システム ibukiC/S について

山田 佳裕，髙松 大地，石原 吉晃，水野 智美
大口 智也，佐藤 芳秀，松本 忠博，池田 尚志
岐阜大学工学部

1 はじめに

我々は日本語文解析システムとして，文節構造解析システム ibukiC と構文解析システム ibukiS を開発している．またそれを用いて点字翻訳システム IBUKI-TEN[5] を開発し公開しているほか，機械翻訳システム jaw[4] の入力文解析部で用いるなどの応用を行っている．

ibukiC についてはこれまでにも報告しているが [1]，今回新たに機能語の構造分割部分や接続属性などを構築し直して新バージョンとしたので報告する．

構文解析解析システムについても，その基本となるアルゴリズム等に関しては既に報告してあるが [2][3]，ibukiC の改版と合わせて ibukiS として構築し直したので合わせて報告する．

2 ibukiC

文節構造解析システム ibukiC は，入力日本語文を文節に分割し，文節の構造を出力するシステムである．文節に現れる機能語部分を，小さな語の単位に分割せずにできるだけそのまま辞書に登録しておいて，逆に辞書上で適切な単位に分割し，あるいはその他の情報を加え，文節構造解析を行う点に特徴がある．もちろん登録機能語の間での接続も可能であり，完全にあらゆる機能語部をそのまますべて登録しようとするわけではない．できるだけ長い単位のまま扱うことによって接続規則に関するあいまいさやゆれの問題を回避することに加え，逆に辞書上で適切な意味的な単位に分割し，さらに標準的な形に言い換えるなどの処理が可能であり，さまざまな応用に対応できると考えた．現在機能語辞書は，約 2 万語のサイズとなっている．以下，今回の主な改良点について述べる．

2.1 機能語要素分割

用言系のとりたて機能「行き は しなかった」の「は」や，体言系の接続機能「〜に対して の」の「の」などを含め，体言系・用言系ともに 6 つの要素に分割することとした．

表 1: 機能語部の分割

		分類	例
体言系	要素 1	とりたて機能 (前)	だけ，すら
	要素 2	格機能	に，を，で
	要素 3	とりたて機能 (後)	こそ，だけ
	要素 4	提題機能	は，も
	要素 5	接続機能	の (=に対して の)
	要素 6	終助詞	ね，な，よ
用言系	要素 1	受身，使役等の機能	させる，られる
	要素 2	時制，肯否等の機能	た，ている，ない
	要素 3	とりたて機能	も，さえ
	要素 4	判断等の機能	だ，だろう，らしい
	要素 5	接続機能	が，のに，ので
	要素 6	終助詞	ね，な，よ

2.2 文節の分割

2.2.1 分割処理

ibukiC では「これは本だ」といった繋辞文などの場合，構文解析の便のために，(これ/は)(本)(だ) のように (本だ) の文節を (本) と (だ) に分割する処理を行っている．

これまでは例 1 のように登録機能語の先頭部分が文節分割の自立語になる場合しか対応できていなかった．新しい ibukiC では例 2,3 のように登録機能語の先頭部分でなくても文節分割できるように整備した．

例 1）(彼/だけ/だったが) → (彼＋だけ)(だ＋た＋が)

例 2）(彼/だけだったが) → (彼＋だけ)(だ＋た＋が)

例 3）(動/い/たからであった)
→ (動く＋た)(からである＋た)

2.2.2 「する/なる」の扱い

「する」はサ変名詞だけでなく，副詞・形容詞などいろいろな語に付属する．「なる」も同様にさまざまな語に付属する．「する/なる」については例 4,5 のように文節分割することとした．

例 4）(ゆっくりする) → (ゆっくり)(する [副詞])

例 5）(ほっそりとなった)
→ (ほっそりと)(なる [副詞] ＋た)

2.3 動詞のとりたて表現の扱い

動詞に後接する機能語の中で,「行き は しない」「動いて さえ いる」といったとりたて詞を含むものがある.

今回の改版で，とりたて詞部分を次のように文節要素として切り出した構造に解析することとした.

(行きはしない) → (行く＋ない＋ は)
(動いてさえいる) → (動く＋ている＋ さえ)

2.4 接続属性・品詞

ibukiC では辞書登録語のすべてに左連接属性 (表 2) と右連接属性 (表 3) をもたせ，その間の接続規則によって形態素・文節解析を行っている．今回の改版でこの左右の属性を長単位機能語の方式に合わせて見直し，左連接属性 106 種，右連接属性 140 種に整備した．これらの組み合わせとしての品詞が現在約 600 種となっている (表 4).

表 2: 左連接属性の一部

左コード	説明
…	
131	機能辞/動詞形容詞/終止形
132	機能辞/動詞・イ型・ダ型/終止形・終止形・語幹
133	機能辞/動詞イ型・ダ型/終止形ウ・終止形・語幹
134	機能辞/動詞イ型・ダ型/連体形
136	機能辞/動詞/終止形
…	

表 3: 右連接属性の一部

右コード	説明
…	
157	機能辞/体言連体格文節
158	機能辞/体言連体格文節/拡張格
159	機能辞/用言連体格文節
160	機能辞/体言並列格文節
161	機能辞/体言—連用格?並列格—文節
…	

表 4: 品詞の一部

左コード	右コード	品詞
…		
13	59	副/トニノ
13	60	副/トニ
13	61	副/トノ
13	62	副/ト
13	63	副/ニノ
13	64	副/ニ
13	65	副/ノ
13	66	副/φ
14	68	接/一般
15	68	接/文頭
…		

2.5 解析例

ibukiC 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.

例） 彼が驚いたのはあのロボットがいきなり動いたからであった。

```
0;0;N; 彼; 名/代/人; Φ; が; Φ; Φ; Φ; Φ; 連用; Φ
1;0;P1; 驚く; 動/カ行; Φ; た; Φ; Φ; Φ; Φ; 直後; Φ
1;1;SN; の; Φ; Φ; Φ; Φ; Φ; は; Φ; 連用; Φ
2;0;AN; あの; 連体; Φ; Φ; Φ; Φ; Φ; Φ; 連体; Φ
3;0;N; ロボット; 名/一般; Φ; が; Φ; Φ; Φ; Φ; 連用; Φ
4;0;AV; いきなり; 副/ダノ; Φ; Φ; Φ; Φ; Φ; Φ; 連用; Φ
5;0;P1; 動く; 動/カ行; Φ; た; Φ; Φ; Φ; Φ; 直後; Φ
5;1;P2; からである; Φ; た; Φ; Φ; Φ; Φ; 文末;。
```

3 ibukiS

ibukiS は,「係り受けは近傍の n ブロック内でなされる (現在 n=3 としている)」という仮説に基づいて，係り受け関係を前向きと後ろ向きに進めていくという言語的なヒューリスティックに基づく方法で解析を行う.

与える規則としては，文節間の係り受け可能性に関する規則と，n ブロック内での係り受けのありようを規則化したブロック化規則を用いる．今回，大きく変更したのは文節間の係り受け関係を規定する規則である．ibukiC が出力する文節構造結果を利用して，より細かな規則化が可能となった.

3.1 ibukiS の係り受け規則

ibukiS では文節間の係り受け関係の可能性を係り受け規則として記述するが，旧システムでは文節カテゴリとして固定的に定められた 196 種類の文節カテゴリ間での規則としていた．今回の改良では，ibukiC の文節構造解析の結果をそのまま用いて文節クラスを定義できるようにして，文節クラス間の規則として記述することとした．図 5 に文節クラスの定義の例を示す．表 5 に現れる「Φ」はその要素が空であることを,「＊」は何でも良いことを表している．現状で 295 種類の文節クラスを定義しているが，必要に応じてさらに文節クラスを定義することは容易である．ibukiS は，この文節クラス定義テーブルのすべての条件にあてはまったとき，その文節に文節クラス ID を割り振る．複数の割り振りの可能性があるときは数値の小さい方を優先する.

このように文節クラスを定義する方式によって，旧システムより容易に，また細かく係り受け関係の有無

表 5: 文節クラスの定義テーブル

ID	文節カテゴリ	係り先	要素 1	要素 2	要素 3	要素 4	要素 5	要素 6	句読点
3	名詞	連体	Φ	Φ	＊	の	＊	＊	Φ
5	名詞	連体	＊	＊	＊	＊	＊	＊	Φ
13	名詞	連用	＊	を	＊	＊	＊	＊	Φ
21	名詞	連用	＊	が	＊	＊	＊	＊	Φ
57	名詞	並列/連用	＊	と	＊	＊	＊	＊	Φ
66	形式名詞	＊	＊	＊	＊	＊	＊	＊	Φ
103	動詞	連用	Φ	Φ	＊	Φ	て	＊	Φ
183	動詞	文末	＊	＊	＊	＊	＊	＊	Φ
230	ダ系	連用	＊	＊	＊	＊	から	＊	Φ
243	動詞	直後	＊	＊	＊	＊	＊	＊	Φ
251	引用機能語	連用	＊	＊	＊	＊	＊	＊	Φ
499	名詞	＊	＊	＊	＊	＊	＊	＊	Φ
503	名詞	連体	Φ	Φ	＊	＊	の	＊	、
513	名詞	連用	＊	を	＊	＊	＊	＊	、
617	動詞	連用	Φ	＊	＊	Φ	ので	＊	、

を規則化できるようになった．

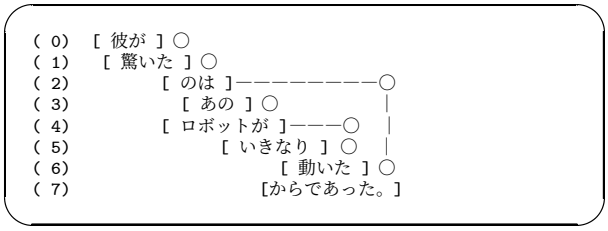
表 6 に係り受け規則テーブルの一部を示す．現在この規則のサイズは約 3 万である．

表 6: 係り受けテーブルの一部

係り元	係り先
13	183
13	243
21	183
21	243
60	122
60	622
66	183
122	102
122	103
243	66

3.2 解析例

例） 彼が驚いたのはあのロボットがいきなり動いたからであった。



4 自動点訳システム ibukiTenC

我々は ibukiK を応用して自動点訳システム IBUKI-TEN を開発しホームページ上で公開してきた．今回 ibukiC の改版と，自立語辞書を独自の辞書として新たに構築し直したことにより，IBUKI-TEN も ibuki-TenC として構築し直した．ibukiTenC は，英語点訳部分の改良など，ユーザの要望も聞きながら進めてきた新たな改良も含んでいる．

5 おわりに

我々が開発している日本語解析システム ibukiC, ibukiS および応用システムの一つである点訳システム ibukiTenC の現状について述べた．今後も改良・整備を続けていく予定である．

参考文献

[1] 伊佐治, 山田, 石原, 髙松, 松本, 池田，文節構造解析システム ibukiC，言語処理学会 第 11 回年次大会 pp.719-722, 2005

[2] 若田，兵藤，池田，文節ブロック間規則による浅い係り受け解析と精度評価，情報処理学会第 57 回全国大会 pp.207-208, 1998

[3] 兵藤，若田，池田，あいまいさを許すロバストな係り受け解析システム， 言語処理学会 第５回年次大会ワークショップ論文集 pp.23-28, 1999

[4] 宇野，福本，田中，松本，池田，日本語から多言語への機械翻訳エンジン jaw， 言語処理学会 第 11 回年次大会 pp.538-541, 2005

[5] 服部，髙松，伊佐治，松本，池田，自動点訳システム IBUKI-TEN の改良と現状，言語処理学会 第 11 回年次大会 pp.217-220，2005

*岐阜大学 池田研究室：http://ikd.info.gifu-u.ac.jp/