機能文節を導入した文節構造解析システム ibukiC(v0.20)について

池田尚志　脇田貴之　大口智也
（岐阜大学 工学部）

１．　はじめに

文節は、日本語の構造にとって基本的で重要な文法要素である。しかし、いわゆる形態素解析システムの多くは、単語を切り出すこととその品詞などの属性を同定することを目的としており、文節あるいは文節構造に焦点を当ててはいない。文節については、構文解析(係り受け解析)を行う際に、形態素列の中に文節の区切りを入れる処理というところで簡単に扱われているようである。

我々は、入力された日本語テキストを、日本語の基本的な文法的単位である“文節”に分割し、さらに文節の内部構造を分析するシステムとして ibukiC を開発してきた[1]｛開発途上版として公開している(v0.10: 07/03), (v0.20: 07/11)｝。V0.10 では、長単位の機能語約 2 万語を登録していたが、V0.20 ではこれを約 4,500 語に整理しなおした。そのうち約 3,500 語は、「V かもしれない」「V コトモノ（V の）」などの特殊な文節を構成するものとして通常の文節から分離し機能文節として分析した。鳥バンク[2]の 15 万文を ibukiC(v0.20)で解析して、その中からランダムに抽出した 100 文について手作業で点検した結果では約 96%程度の正確度であった。

２．　ibukiC(v0.20)による解析例

図１に ibukiC(v0.20)による解析例を示す。図１の各行は１個の文節を表現している。(1;2;1)の文節”FuncP3;V/たい/,….”や、(1;2;2)の文節”FuncP1;Y/ん/です/,….”は、(1;2;0)の「見たかったんですが」という文節から分離された“機能文節”である。「た」「が」は、これらの機能文節の“文節内要素”として分析されている。各行の内容は次のとおりである。

A : 通常モード

```
1;0;0;わたし/は;N;わたし;名/代/人;Φ;Φ;は;Φ;Φ;Φ;連用;Φ
1;1;0;紅白/歌/合戦/が;N;紅白;名/一般;Φ;Φ;Φ;Φ;Φ;Φ;複合語;Φ
1;1;1;特殊文節;TailN;歌;尾/名|名・一般/161.81;Φ;Φ;Φ;Φ;Φ;Φ;複合語;Φ
1;1;2;特殊文節;N;合戦;名/サ;Φ;が;Φ;Φ;Φ;Φ;連用;Φ
1;2;0;見/た/かっ/た/んです/が/、;P1;見る;動/１段;Φ;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;2;1;特殊文節;FuncP3;V/たい/;機/動|形イ/117.37;た;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;2;2;特殊文節;FuncP1;Y/ん/です/;機/用|文末/134.164;Φ;Φ;が;Φ;Φ;Φ;連用;、
1;3;0;弟/は;N;弟;名/一般;Φ;Φ;は;Φ;Φ;Φ;連用;Φ
1;4;0;裏番組/の;N;裏番組;名/一般;Φ;Φ;Φ;Φ;の;Φ;連体;Φ
1;5;0;映画/を;N;映画;名/一般;Φ;を;Φ;Φ;Φ;Φ;連用;Φ
1;6;0;見/たが/り/ました/。;P1;見る;動/１段;Φ;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;6;1;特殊文節;FuncP1;V/たい/がる/;機/動|ラ行/117.28;ます/た;Φ;Φ;Φ;Φ;Φ;文末;。
```

B:標準化モード

```
1;0;0;わたし/は;N;わたし;名/代/人;Φ;Φ;は;Φ;Φ;Φ;連用;Φ
1;1;0;紅白/歌/合戦/が;N;紅白;名/一般;Φ;Φ;Φ;Φ;Φ;Φ;複合語;Φ
1;1;1;特殊文節;TailN;歌;尾/名|名・一般/161.81;Φ;Φ;Φ;Φ;Φ;Φ;複合語;Φ
1;1;2;特殊文節;N;合戦;名/サ;Φ;が;Φ;Φ;Φ;Φ;連用;Φ
1;2;0;見/た/かっ/た/んです/が/、;P1;見る;動/１段;Φ;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;2;1;特殊文節;FuncP3;V たい;機/動|形イ/117.37;た;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;2;2;特殊文節;FuncP1;Y のだ;機/用|文末/134.164;Φ;Φ;けれども;Φ;Φ;Φ;連用;、
1;3;0;弟/は;N;弟;名/一般;Φ;Φ;は;Φ;Φ;Φ;連用;Φ
1;4;0;裏番組/の;N;裏番組;名/一般;Φ;Φ;Φ;Φ;の;Φ;連体;Φ
1;5;0;映画/を;N;映画;名/一般;Φ;を;Φ;Φ;Φ;Φ;連用;Φ
1;6;0;見/たが/り/ました/。;P1;見る;動/１段;Φ;Φ;Φ;Φ;Φ;Φ;直後;Φ
1;6;1;特殊文節;FuncP1;V たがる;機/動|ラ行/117.28;た;Φ;Φ;Φ;Φ;Φ;文末;。
```

①文 id;
②文節 id;
③文節 subId;
④文節;
⑤文節カテゴリ;
⑥内容語;
⑦内容語の品詞;
⑧文節内要素(1~6);
⑨係り先;
⑩句読点

図１　解析例
(A:通常モード
B:標準化モード)
「わたしは紅白歌合戦が見たかったんですが、弟は裏番組の映画を見たがりました。」

ibukiC の解析には、通常モード(図１;A)と標準化モード(図１;B)がある。標準化モードでは、「見たかったんですが」は「見たかったのだけれども」と、「見たがりました」は「見たがった」のように解析

されている。標準化は辞書の記述に基づいて単純に語彙の変換をしているだけであるが、多くの表記のゆれや表現のゆれをある程度統一的な表現に変換してくれので、応用の場面によっては有効に使えるものと考えている。

“文節内要素”は文節のカテゴリ毎に定義してあり、否定(ない)/時制(た)や文節間の接続に関わる機能語(格助詞、接続助詞の類)の情報を中心に配置している(表 1)。

表 1. 文節内要素の定義の概要

文節の種類	要素 1	要素 2	要素 3	要素 4	要素 5	要素 6
体言	取り立て（格の前）	接続（→用言/格）	取り立て(格の後)	接続（→体言）	Φ	終助詞
用言	時制・極性	接続（→用言/格）	接続（→用言/節）	接続（→体言）	モード	終助詞
副詞	Φ	接続（→用言）	取り立て(格の後)	接続（→体言）	Φ	終助詞

ibukiC はまた、通常の形態素解析のように形態素の列を表示することも出来る(図 2)。表示される内容は次のとおりである。

①文 id; ②文節 id; ③誤り可能性の有無; ④文節 subId; ⑤形態素; ⑥形態素 id; ⑦品詞; ⑧読み

```
1;0;Φ;0;わたし;530256;名/代/人;わたし
1;0;Φ;1;は;70951;機/名|連用/101.141;わ
1;1;Φ;0;紅白;556298;名/一般;こーはく
1;1;Φ;1;歌;800147;尾/名|名・一般/161.81;か
1;1;Φ;2;合戦;557575;名/サ;かっせん
1;1;Φ;3;が;70893;機/名|連用/101.141;が
1;2;Φ;0;見;551614;動/１段;み
1;2;Φ;1;た;87940;機/動|形イ/117.37;た
1;2;Φ;2;かっ;70427;機/形イ|活尾/70.122;かっ
1;2;Φ;3;た;88916;機/用|文末/118.170;た
1;2;Φ;4;んです;93040;機/用|文末/134.164;んです
1;2;Φ;5;が;91598;機/用|連用ｃ/131.147;が
1;2;Φ;6;、;96591;記/読点;てん
```

図 2　形態素列の表示例

```
1;2;Φ;0;見;551614;動/１段;み
1;2;Φ;1;た;88916;機/用|文末/118.170;た
1;2;Φ;2;か;91802;機/用|文末/137.168;か
1;3;#;0;っ;0;名/未知語/ひらがな;っ
1;4;#;0;で;516985;動/１段;で
1;5;Φ;0;すが;2000336;名/個/人名/姓;すが
1;5;Φ;1;、;96591;記/読点;てん
```

図 3 誤り可能性の指摘例

“誤り可能性の有無”は、ヒューリスティックな規則によってその部分の解析が誤っている可能性があると判断された場合に“#”を、そうでない場合には“Φ”を表示する(図 3)。“誤り可能性”には、入力文は正しいが解析が誤っている可能性だけでなく、入力文が誤っている可能性も含む。要するに、解析結果中の何らかの乱れや稀な事象の検出を意図している。この“誤り可能性の有無”については、再現率が未だ十分ではないが、自動点訳システム ibukiTenC において、点訳結果の点検をする場合の点検箇所を誘導するための情報の一部として利用している。十分な精度と再現率が得られれば、利用範囲はいろいろ広がると考えている。

“読み”は、自動点訳への応用を考慮しており、「は」は「わ」、「おとうと」は「おとーと」など、点字としての表記で読みを表示している。

3. ibukiC の辞書データおよび解析方式の概要

ibukiC では辞書や規則などのすべてのデータは、RDB(関係データベース)上で管理している。辞書や規則は、未だ十分整備された状態ではないが、現在のところ機能語辞書には約 4,500 語が登録してあり、そのうち約 3,500 語は機能文節を分割している。内容語辞書には約 24 万語彙が登録されている。表 2 に機能文節の例を示す。表 3 に文節内要素の一部の例を示す。表 4 は現在登録してある文節内要素の分布である。

ibukiC の解析は、基本的には左右の連接コードによる語と語の接続条件、文節の開始点・終了点の条件、をチェックして接続コストや単語コスト、部分的には bigram コスト、漢字複合語内部のコスト、なども利用して、コストを計算して、最小のコストの単語列/文節列を導き出すという方法で行われる。古典的な解析方法であるが、“形態素”を切り出すことに主眼を置くのではなく文節を切り出し、その構造を分析することに主眼を置いていること、機能語辞書に長めの単位の語を登録していること、”機能文節”を分離していること、などが特徴である。

表 2　機能文節の例

親文節	分離機能文節	機能文節の延べ数	機能文節の異なり数	見出し語の例	機能文節語の例	機能文節語（標準化）の例
体言	体言	230	41	かのを、かは、か否かが、のかと言えば、	N/か/の/、N/か/、N/か/否か/、N/の/か/、、	N かノモノ、N デアルか、N デアルかどうか、N のモノか、、
体言	副言	12	11	きっての、だてらに、なす、らしく、高く、、	N/きっての/、N/だてらに/、N/なす/、N/らしく/、N/高く/、、	N きっての、N だてらに、N なす、N らしく、N 高く、、
体言	用言	293	169	かもしれず、かも知れず、かもしれな、かも知れな、から見ると、から見れば、、	N/かも/しれ/ず/、N/かも/知れ/ず/、N/かも/しれ/ない/、N/かも/知しれ/ない/、N/ノカンテンカラ/、、	N デアルかもしれない、N/ノカンテンカラ/、、
副言	体言	3	3	なの、なん、の	F/なの/、F/な/ん/だ/、A/の/	F であるコトモノ、F なのだ、A のモノ
副言	用言	18	11	かもしれな、だ、だろう、でした、なのだ、、	F/かも/しれない/、F/だ/、F/だ/ろう/、F/です/、F/な/の/だ/、、	F かもしれない、F だ、F だろう、F なのだ、、
用言	体言	631	185	いっぽう、いっぽうで、うえで、うえでの、おりに、かどうかについて、かどうかによって、、	Y/いっぽう/、Y/いっぽうで/、Y/うえで/、Y/おりに/、Y/か/どう/か/、、	Y いっぽうで、Y うえで、Y おりに、Y かどうかトイウコト、、
用言	副言	2	1	がはやいか、が早いか	V/がはやいか/、V/が早いか/	V がはやいか
用言	用言	2320	1382	がたい、がち、がちにな、けばいい、けばよい、けば良い、ことすらな、ずじまいでした、、	V/がたい/、V/がち/だ/、V/がち/に/なる/、V けばいい、V/こと/すら/ない/、V/ず/じまい/です/、、	V がたい、V がちだ、V がちになる、V たらよい、V ことさえない、V ずじまいだ、、
合計		3509	1803			

表 3　文節内の要素の例

	要素 1		要素 2		要素 3		要素 4	
	表層形	標準形	表層形	標準形	表層形	標準形	表層形	標準形
体言文節	くらい、ぐらい、だけ、こそ、さえも、、	くらい、だけ、こそ、さえも、、	が、から、で、を、にあたって、にあたり、にあたりまして、、、、	が、から、で、を、にあたって、、、、	こそ、さえ、しか、すら、は、って、ったら、、、	こそ、さえ、しか、すら、は、、、	からの、たりえる、にかんしての、にかんする、、	からの、たりえる、にかんする、、
用言文節	た、だ、ない、ぬ、ん、ません、ます/た、ません/でした、、	た、ない、なかった、、	しか、ちゃ、じゃ、ては、にあたり、にあたって、にあたりまして、、	しか、ては、てこそ、にあたって、、	あげく、あげくに、が、けど、けれど、けれども、、	あげく、けれども、、	あげくの、こその、て以降の、で以降の、との、、	あげくの、てこその、て以降の、との、、

表 4　現バージョンの辞書に登録してある文節内要素の異なり数

親文節	分離する機能文節	要素 1		要素 2		要素 3		要素 4		要素 5		要素 6	
		表層	標準	表層	標準	表層	標準	表層	標準	表層	標準	表層	標準
体言		13	13	138	83	23	16	87	73	0	0	5	5
体言	体言	6	4	38	33	9	7	11	11	0	0	0	0
体言	用言	4	2	5	3	28	22	14	2	2	2	10	10
副言		0	0	3	3	3	3	1	1	0	0	0	0
副言	用言	1	1	0	0	1	1	0	0	1	1	0	0
用言		16	7	44	37	129	83	15	14	1	1	43	37
用言	体言	9	6	69	64	17	14	18	17	0	0	0	0
用言	用言	15	4	6	6	51	51	4	4	18	18	7	6
合計数		48	37	303	229	261	213	150	122	22	22	65	58

接続規則は手作業で与えており、さほど微妙な調整はしていないがほぼ十分に機能している。ibukiC(v0.20)で毎日新聞記事 10 年分を解析した結果での、機能語/文節機能語の出現割合等を表 5～表 8 に示す。

表 5　新聞記事 10 年分中の機能語

	延べ		異なり	出現しなかった辞書登録語	
	出現数	割合	語数	語数	割合
機能文節を分離する機能語	22,241,932	17.50%	2679	819	76.60%
分離しない機能語	104,855,548	82.50%	941	75	92.60%

表 6　新聞記事 10 年分中の機能文節

出現数	全出現機能語中の割合	機能語
4,890,749	3.85%	Nする
2,311,313	1.81%	Vている
1,752,244	1.38%	Vられる
1,478,812	1.16%	Nだ
1,175,356	0.92%	Q」と
816,024	0.64%	Nなど
674,072	0.53%	Yこと
671,066	0.53%	Qと
651,404	0.51%	Kだ
514,474	0.40%	Kこともの
…	…	…
672	0.00%	Vてやまない
…	…	…
1	0.00%	Vてみたがる

表 7　辞書登録語のうち 10 年分中に出現した語

	異なり数	辞書登録語中の割合
通常形	1,291	73.70%
標準化形	813	75.00%

表 8　10 年分中に出現しなかった語の例

Nときたらない，Nぐむ，Nだらけだ
Nめく，N至極だ，V交わす，V懸ける，V終わる，Vわたる，V果てる，，，

3．　ibukiC(v0.20)の解析精度

鳥バンク[2]の約 15 万文を ibukiC(v0.20)で解析し、その中からランダムに抽出した 100 文を手作業で点検した。その結果、100 文は 781 文節に解析されたが、そのうち筆者が解析誤り部分と判断したのは 35 文節であった（4.5%）。文としては 14 文中に誤りが含まれていた（14%）。誤り 35 文節のうち、33 文節は｛釣りざお、釣竿｝のような異表記の問題を含む辞書登録不備の問題、2 文節は「中止した方が｛ほう／かた｝」のような意味に関わる問題、「縛りあげさるぐつわを」の場合に連用中止として認識できなかった問題、であった。

東西冷戦下で仏の核はソ連に向けられたが、冷戦構造は崩壊し、ソ連は既にない。
[東西]○
  [冷戦]○
  [下(で)]――――○
    [仏(の)]○　　|
      [核(は)]――○
      [ソ連(に)]―○
        [向ける]○
[Vられる(た,が,、)]―――○　★
              [冷戦]○　|
            [構造(は)]―○
                [崩壊]○
            [Nする(し、)]――○
                [ソ連(は)]―○
                    [既に]○
                  [ない(。)]

図 4　ibukiS による解析例

4．　ibukiC の結果を用いた係り受け解析

今回の v0.20 には含ませていないが、我々は、ibukiC の結果を入力とする係り受け解析システム ibukiS も開発している(図 4)。ibukiC が出力する“文節カテゴリ”、“係り先”を基本的な情報として、必要に応じて文節内要素の情報も用いて、文節間の係り受け規則を設定する（現在のところ基本規則 59 個、細則 145 個）。近接 3 ブロック内での係り受け関係を優先して前向きと後ろ向きのブロック化を遂行し、係り受け解析を行う[3]。簡単な方法であるが、現在の段階で、文節ベースで 90%程度の精度を得ている。ibukiS についても ibukiC に含めて近く公開したいと考えている。

5．　おわりに

開発途上の文節構造解析システム ibukiC(v0.20)について簡単に述べた。

我々の研究室では、ibukiC を基礎部分に利用して、係り受け解析システム(ibukiS)、自動点訳システム(ibukiTenC)、機械翻訳システム(jaw)、手話テキストへの翻訳システム(jaw/SL)、テキストマイニングなどについて研究･開発している。ibukiTenC も研究室のホームページで公開している。

ibuikiC の辞書や規則は未だ整備が不十分な部分も多い。今後も整備を続けていく予定である。

文献

[1]　http://www.ikd.info.gifu-u.ac.jp/ibukiC/
[2]　http://unicorn.ike.tottori-u.ac.jp/toribank/
[3]　大口智也, 構文解析システム ibukiS に関する研究, 岐阜大学工学研究科, 修士論文, 2008.3