長単位の機能語を辞書に持たせた文節構造解析システム ibukiC

伊佐治 和哉，山田 将之，池田 尚志
岐阜大学工学部

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．

さらに，機能語辞書に登録する機能語の粒度についての試みを行った．すなわち，新聞記事１年分の解析データをもとにした，長単位の機能語を持たせた解析システム ibukiC を作成した．これによって，解析の容易さと高精度化，また応用システムでの意味処理が煩瑣にならないことを目指したシステムを構築する見通しを得た．ibukiC ではまた，応用システムでの意味処理等を考慮して，機能語部を標準的な表現に置き換えて出力する機能も持たせた．

2　文節構造解析システム ibukiB

ibukiB の概要を図 1 に，出力例を図 2 に示す．

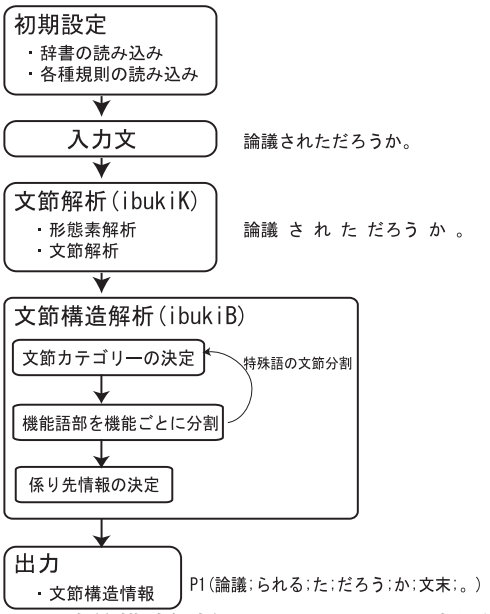

図 1: 文節構造解析システム ibukiB の概要

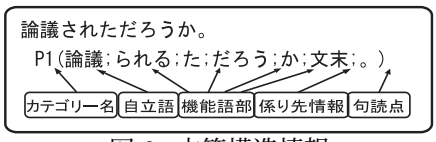

図 2: 文節構造情報

2.1　ibukiK

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辞書，規則を参照し求めた上で，文節および隣接する文節間にコストを与え，単語単位ではなく文節単位のコスト最小法によって形態素・文節解析を行っている．

2.2　ibukiB

ibukiB は ibukiK の出力をもとに「文節構造」を作り出す．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」などから構成される．

ibukiB では以下のことを行っている．

1. 文節カテゴリーを与える．(表 1)

表 1: 文節カテゴリー

体言系	N	名詞文節
	SN	形式名詞文節
	KA	カ系文節
	Q	」文節（引用の終わり）
	TO	引用機能語文節
用言系	P1	動詞文節
	P2	ダ系文節
	P3	形容詞文節
	P4	形容動詞文節
その他	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

2. 機能語部を機能ごとに４つの要素に分割する．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．(表 2)

表 2: 機能語部の分割

		分類	例
体言系	要素１	格助詞相当語に前接する副助詞等	だけ，すら
	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	こそ，だけ
	要素４	提題助詞	は，も
用言系	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

3. 「ぐらい」「くらい」などの字面上は違うが意味が同一のものは，統一する．

4. その文節がどのような文節に係るかという係り先情報を与える．(表 3)

表 3: 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

「並列／連用／疑問」は，並列か連用か疑問のいず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．以下に例を示す．

- 並列：食うか 食われるかの時が来た．
  P1(食う; Φ; Φ; Φ; か; 並列/連用/疑問; Φ)
- 連用：だれか 来たように思ったが空耳だった．
  N(だれ; Φ; か; Φ; Φ; 並列/連用/疑問; Φ)
- 疑問：どちらが 強いか 勝負しよう．
  P3(強い; Φ; Φ; Φ; か; 並列/連用/疑問; Φ)

5. 用言文節と体言文節が相互に転化しているような場合には，元は同じ文節であるという情報を保持したまま，以下のように文節分割を行う．

- ノ系
  例：君+の (ノ系)+だけ (副助詞)+に (格助詞)+は (提題助詞)　→　[君／の (ノ系)][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，対処することができる．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 (ダ系)][だ／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」や「かもしれない」などを「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]
  用言文節に形式名詞が後接して述語名詞化文節となっている．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とした．
- カ系
  例：君+かどうか (疑問)+が (格助詞)
  →　[君 (カ系)][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ので,「か」「かどうか」の後に体言後接語が続く場合にカ系文節として文節を区切ると定義した．

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．

- 彼のは見易い筆跡だ．

```
0 0 N(彼; Φ; Φ; の; Φ; ノ系; Φ)
0 1 N(の; Φ; Φ; Φ; は; 連用; Φ)
1 0 P1(見る; やすい; Φ; Φ; Φ; 連体; Φ)
2 0 N(筆跡; Φ; Φ; Φ; Φ; ダ系; Φ)
2 1 P2(だ; Φ; だ; Φ; Φ; 文末;. )
```

- 君に会えたのも何かの縁でしょう．

```
0 0 N(君; Φ; に; Φ; Φ; 連用; Φ)
1 0 P1(会う; Φ; た; Φ; Φ; 形式名詞; Φ)
1 1 SN(の; Φ; Φ; Φ; も; 連用; Φ)
2 0 N(何; Φ; Φ; Φ; Φ; カ系; Φ)
2 1 KA(か; Φ; Φ; の; Φ; 連体; Φ)
3 0 N(縁; Φ; Φ; Φ; Φ; ダ系; Φ)
3 1 P2(だ; Φ; Φ; でしょう; Φ; 文末;. )
```

3 ibukiC

ibukiB でもそうであるが一般に機能語部は，自立語および機能語間の接続規則によって機能語の連接の可否を判定することで解析する．ところで機能語として切り出す単位は，さまざまに可能である．たとえば「壊されてしまったのかもしれませんが」を「壊す＋れる＋て＋しまう＋た＋の＋か＋も＋しれる＋ます＋ん＋が」のように細かく切り出すことも可能であるし,「壊す＋れる＋てしまう＋た＋のかも＋しれません＋が」のようにやや長めの単位で切り出すことも，

また「壊す＋れる＋てしまったのかも＋しれませんが」あるいは「壊す＋れてしまったのかもしれませんが」のように長く切り出すことも可能である．短い単位で切り出せば，登録する機能語の種類は少数（数十個程度）ですむが，接続規則の設定が複雑になり，また機械翻訳など応用場面での意味的な扱いは，それらを組み合わせて要素合成的に処理していかなければならず複雑になる．一方,「壊す＋れてしまったのかもしれませんが」のように長い単位で扱うと，接続規則は単純になり，また意味的扱いが錯綜し複雑になることはないという点で有利である．しかし，対応すべき表現の数が膨大になり現実的ではなくなる恐れがある．

このように機能語辞書にどのような長さで機能語を登録するかということは文節解析における一つの問題点である．ibukiB では比較的長い単位の機能語を登録しているが，我々はさらに長い単位で扱うことを試みた．

3.1 ibukiB の辞書

機能語には,「ている，てはいる，てもいる」「かもしれない，かも知れない」のように，表記の違い，活用，助詞の付加による意味の添加などによる派生的な語が多数存在する．ibukiB ではこのような機能語を１つのグループとして扱い，機能語の整理・収集を行って辞書に登録している．図 3 に，グループ「ざるをえない」に対する辞書表現を示す．

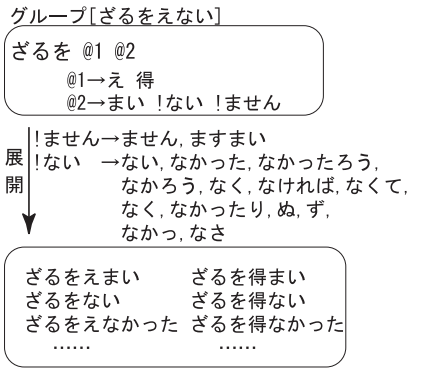

図 3: ibukiB の機能語辞書

3.2 ibukiC の辞書

ibukiB でもこのように長めの単位で機能語を登録しているが，我々は特定の連語だけではなく，実際に現われる文節中の機能語部を，ほとんどすべて辞書に登録することを考えた．

そのような新しい機能語辞書を作成するために，新聞記事１年分（約 67 万文，約 1400 万文節）を ibukiB で解析し，そこに出現した機能語をほぼそのまま機能語見出しとして登録することを試みた．表 4 に，1 年分の解析結果を示す．

表 4: ibukiB による新聞記事解析結果

文節		出現頻度	機能語部パターン数	到達位 (%) 90	95	99	99.9
体言系	N	6,822,001	1,351	9	16	62	265
	SN	301,768	322	8	12	33	138
	TO	43,389	192	12	22	53	149
	Q	903,816	451	7	10	31	129
	KA	14,996	119	13	19	49	104
用言系	P1	2,313,180	9,913	66	225	1,684	7,600
	P2	275,013	2,013	71	153	620	1,738
	P3	222,219	1,427	10	44	324	1,205
	P4	238,301	1,844	8	33	392	1,606
その他		3,193,628	1,944	10	36	403	1,832
合計		14,328,311	19,576	135	381	2,348	14,466

異なり数で約 2 万件の機能語部パターンが現れたが，このうちの出現頻度の上位約１万件を辞書に登録した．この辞書には長単位の機能語を持たせているため，機能語間の接続は基本的に行う必要はないが，未登録な機能語の解析を考慮し，コストは高めではあるが接続を許可している．この辞書によるシステムを ibukiC とした．

ibukiB と ibukiC の辞書の登録語の関係を図 4 に示す．

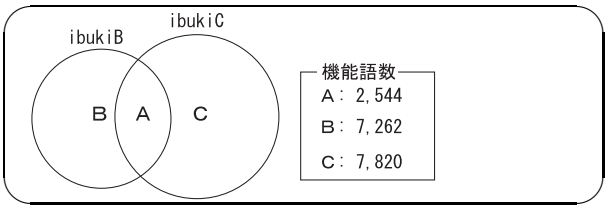

図 4: ibukiB と ibukiC の辞書の関係

ibukiB と ibukiC で共通している単語はそれほど多くはなく，このことから ibukiB では新聞記事の場合，ほとんどの機能語部は複数の単語を接続させることで解析していることが分かる．

3.3 ibukiC による解析

ibukiC によって，新聞記事１年分を再度解析した結果，次のことが分かった．表 4 の機能語部パターン約２万件のうち，出現頻度の上位約１万文節に現れる機能語はそのまま登録し，残り約１万文節に現れる機能語は登録しなかったのであるが，登録しなかった１万語の機能語のうち約 85%は登録した機能語間の接続により ibukiC で正しく解析された．

辞書に未登録であるが正しく解析できた例と，誤った解析の例を以下に示す．

- 正解例１ (見え//にくく/なっていたが)

```
0 0 P1(見える; にくい/なる; ている/た; Φ; が; 連用; Φ)
```

- 正解例２ (伺//わ/れる/にもかかわらず)

```
0 0 P1(伺う; られる; Φ; Φ; にもかかわらず; 連用; Φ)
```

- 誤り例１ (存在)(出来//なくなっている)

```
0 0 N(存在; Φ; Φ; Φ; Φ; 独立; Φ)
1 0 P1(出来る; ない/なる; ている; Φ; Φ; 連体; Φ)
```

- 誤り例２ (付与//し/なくては/なる)(まい)

```
0 0 P1(付与; Φ; ない; Φ; ては/なる; 連体; Φ)
1 0 N(まい; Φ; Φ; Φ; Φ; 独立; Φ)
```

例のような正しく解析されなかった残りの約15%は，新聞記事１年分の全文節中の0.02%程度であった．結局，新聞記事１年分に対するibukiBとibukiCによる解析結果は，ほぼ同じとなった．

さらに新聞記事とは違うジャンルでの試みとして小説「坊ちゃん」を解析してみた結果では，古い言葉など辞書未登録の語が多かったため正解率は低かったが，解析結果はほぼ同じであった．

文節分割については，ibukiBでも同様であるが未だ不適切なものが少なくない．ibukiCの辞書記述によれば個々に適切な分割を与えることができるので，今後対処していく予定である．

- 例１：Vているはずのところなのに
  → (V//ている/はずだ/のところなのに)
  (分割しない)

- 例２：Nであることにほかならない
  → (N//である) (こと//にほかならない)
  (分割する)

3.4 機能語要素の標準表現への変換

ibukiCではまた，登録機能語の要素への分割の際に，標準的な表現に変換することも試みた．機能語を「会話」・「丁寧」などの表現別に分類し，「標準」と同一の表現に置換できるものは統一した．また，分類が「標準」のものに関しても，他の「標準」に置換できるものは同様に標準化を行った．標準化の例を表5に示す．

表 5: 標準化の例

分類	辞書登録数	置換数	例	標準化後
標準	8481	40	でしょう	だろう
会話	427	387	じゃ/ない	では/ない
丁寧	1379	1261	できる/ません	できる/ない
古語	72	48	とて	といっても
方言	5	2	ん/よ	の/よ

毎日新聞１年分の解析結果中の用言文節に着目し，機能語部の各要素別でのパターン数の変化を，標準化前と標準化後で比較した．統計結果を表6，表7に示す．

表 6: 用言要素別統計 (標準化前)

	要素１		要素２		要素３		要素４	
出現頻度	286,017		1,379,515		104,800		549,389	
パターン数	112		521		990		845	
90%到達位	7		15		83		26	
95%到達位	13		26		136		45	
99%到達位	29		74		398		132	
99.9%到達位	60		213		886		439	
上位５位 & 割合 (%)	られる	65.7	た	46.9	ろう	8.3	て	25.1
	させる	8.0	ている	7.5	だろう	6.1	が	15.8
	たい	6.8	だ	6.8	でしょう	4.1	と	15.5
	なる	4.4	ない	6.3	ね	4.0	という	6.0
	てくれる	2.8	ている/た	4.3	よ	3.5	か	4.9

表 7: 用言要素別統計 (標準化後)

	要素１		要素２		要素３		要素４	
出現頻度	286,017		1,378,329		104,666		549,499	
パターン数	**110**		**458**		**934**		**812**	
90%到達位	6		12		77		26	
95%到達位	11		20		127		44	
99%到達位	27		55		368		128	
99.9%到達位	58		170		830		416	
上位５位 & 割合 (%)	られる	65.7	た	47.8	だろう	10.1	て	25.1
	させる	8.0	ている	12.3	ろう	8.4	が	15.8
	たい	6.8	だ	8.3	ね	4.1	と	15.5
	なる	4.4	ない	6.6	よ	4.0	という	6.0
	てくれる	3.5	ている/た	4.4	の/だ	3.9	か	4.9

このように，機能語部の各要素に標準表現を用いることで，応用システムでは標準表現を元にした規則等を作成することで規則の簡略化や点検がしやすくなることが期待できる．

4 おわりに

長単位の機能語辞書を持たせた文節構造解析システムの開発を試みた．これらの試みから，１万語程度の大きな粒度の機能語を登録することで，接続コスト等で高精度化を目指すのではない新たな解析システムを構築できる見通しを得た．現在は未だibukiBとの比較で精度が向上したとはいえないが，今後は個々に登録機能語を精査し精度向上をめざす予定である．文節分割の扱いについても，現在の分割には未だ不適切なものがあるが，ibukiCの機能語辞書によれば，各長単位の機能語毎に分割の仕方を記述することが出来る．辞書記述の作業のコストはかかるが，適切な構文解析を得ることにつながる適切なデータを与えることが出来る．これらによって高精度の文節構造解析システムの開発を進めていく予定である．

参考文献

[1] 文節解析システム ibukiB と大規模コーパス中の文節パターンの分布について　岸井, 伊佐治, 高木, 池田　言語処理学会 第 9 回年次大会 発表論文集 (2003)

[2] 文節解析のための長単位機能語辞書　兵藤, 村上, 池田　言語処理学会 第 6 回年次大会 発表論文集 (2000)