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 謙一　　伊佐治 和哉　　高木 優紀江　　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再分割して出力する，文節構造解析システムibukiBを構築した．

また，このibukiBを用いて，2種類の大量テキストデータを文節構造解析し，機能語や文節パターンの出現頻度分布を調査した．

2 文節構造解析システム ibukiB

2.1 システムの概要

ibukiBの概要を図1に示す．

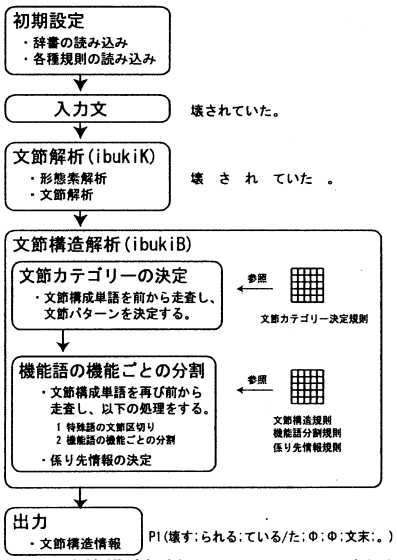

図1　文節構造解析システムibukiBの概要

ibukiBの出力例を図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．

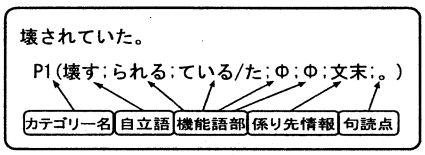

図2　文節構造情報

2.2 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すべて求めた上で，従来の単語単位の方法ではなく，文節および隣接する文節間にコストを与えて，文節単位のコスト最小法によって文節解析を行っている．

2.3 ibukiB

ibukiBでは以下のことを行っている．

- 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．

2.3.1 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	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

表2　用言系文節（4種類）

P1	動詞文節
P2	ダ系文節
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

2.3.2 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．

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，次頁の1~4のように文節を分割することで対処した．

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

1. 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．

2. 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．

3. 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．

4. 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．

2.3.3　係り先情報

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 1 2 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．以下に例を示す．

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

- 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)

2.3.4　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．

2.3.5　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．

- 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 N(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　大規模コーパス中の文節機能語要素および文節パターンの分布調査

3.1　概要

ibukiB を用いて大規模コーパス (毎日新聞記事 67 万文, 辞書用例文等 15 万文) を文節構造解析して，文節カテゴリー，機能語要素，文節パターン等の出現頻度分布調査を行った．

3.2　新聞記事 1 年分に対する調査

'00 年毎日新聞記事 1 年分中の約 67 万文を文節構造解析して，種々の頻度統計を行った．

表7に文節カテゴリーの出現頻度を，表8に文節パターンの出現頻度を示す．

表7　文節カテゴリー統計 (毎日新聞)

文節カテゴリー	説明	出現頻度	割合
N	名詞	3,342,512	62.41%
P1	動詞	1,177,915	21.99%
SN	形式名詞	156,536	2.92%
A	副詞	155,487	2.90%
P2	ダ系	138,625	2.59%
P4	形容動詞	117,094	2.19%
P3	形容詞	113,909	2.13%
T	連体詞	63,021	1.18%
C	接続詞	50,762	0.95%
UN	未知語	27,152	0.51%
TO	引用機能語	11,539	0.22%
I	感動詞	1,380	0.03%
合計		5,355,959	100.00%

表8　文節パターン統計 (毎日新聞)

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	3,342,512	480	15	25	86	231
	SN	156,536	208	9	13	40	120
	TO	11,539	86	13	21	40	75
用言系	P1	1,177,915	9,373	126	366	2,354	8,196
	P2	138,625	1,993	112	229	866	1,855
	P3	113,909	1,371	28	83	489	1,258
	P4	117,094	1,776	21	89	714	1,659
他	A	155,487	45	3	5	8	13

名詞文節のほうが動詞文節よりも出現頻度は大きいが，名詞文節のパターン数は動詞文節と比べてはるかに小さくなる．また，名詞文節は上位15パターン，動詞文節は上位126パターン，形式名詞文節，引用機能語文節，ダ系文節，形容詞文節，形容動詞文節，副詞文節はそれぞれ，上位9,13,112,28,21,3パターンを見るだけで，機能語部の90 %を把握できることが分かる．

表9　名詞後節機能語部上位15パターン

要素1	要素2	要素3	要素4	係り先情報	パターン数	割合
Φ	Φ	の	Φ	連体	635,857	19.02%
Φ	を	Φ	Φ	連用	440,946	13.19%
Φ	に	Φ	Φ	連用	328,087	9.82%
Φ	が	Φ	Φ	連用	319,261	9.55%
Φ	Φ	Φ	は	連用	317,552	9.50%
Φ	Φ	Φ	Φ	並列	199,806	5.98%
Φ	で	Φ	Φ	連用	179,520	5.37%
Φ	Φ	Φ	Φ	文末	136,414	4.08%
Φ	Φ	Φ	Φ	ダ系	105,239	3.15%
Φ	と	Φ	Φ	並列/連用	98,398	2.94%
Φ	Φ	Φ	も	連用	75,412	2.26%
Φ	Φ	Φ	Φ	独立	64,885	1.94%
Φ	から	Φ	Φ	連用	57,063	1.71%
Φ	や	Φ	Φ	並列	49,060	1.47%
Φ	で	Φ	は	連用	30,297	0.91%

表10　動詞後節機能語部上位15パターン

要素1	要素2	要素3	要素4	係り先情報	パターン数	割合
Φ	Φ	Φ	Φ	連体	119,043	10.11%
Φ	た	Φ	Φ	文末	116,309	9.87%
Φ	Φ	Φ	Φ	連用	103,864	8.82%
Φ	た	Φ	Φ	連体	91,260	7.75%
Φ	Φ	Φ	Φ	文末	82,934	7.04%
Φ	Φ	Φ	て	連用	66,600	5.65%
Φ	Φ	Φ	Φ	形式名詞	62,948	5.34%
Φ	ている	Φ	Φ	文末	38,616	3.28%
Φ	た	Φ	Φ	形式名詞	30,686	2.61%
Φ	Φ	Φ	と	連用	21,829	1.85%
Φ	ない	Φ	Φ	連体	13,698	1.16%
られる	た	Φ	Φ	連体	12,508	1.06%
られる	Φ	Φ	Φ	連用	11,958	1.02%
Φ	ている/た	Φ	Φ	文末	11,838	1.01%
Φ	た	Φ	が	連用	11,142	0.95%

表9,10に名詞後節機能語部の上位15パターンと動詞後節機能語部の上位15パターンを示す．

機能語部要素の出現頻度　名詞文節と動詞文節の各要素ごとにどれくらいのパターン数が存在するかを調査するために機能語部要素の統計を行った．結果と，そのときの上位5パターンを表11,12に示す．

表11　名詞要素別統計 (毎日新聞)

	副助詞 (前)		格助詞等		副助詞 (後)		提題助詞	
出現頻度	16,131		1,751,046		698,173		484,812	
パターン数	23		124		18		5	
90%到達位	12		7		1		2	
95%到達位	14		12		1		2	
99%到達位	18		33		2		2	
99.9%到達位	22		75		7		3	
上位5位 & 割合 (%)	だけ	39.2	を	25.3	の	97.1	は	79.8
	ばかり	7.3	に	21.4	も	1.9	も	20.0
	しか	7.1	が	18.3	とは	0.6	も/の	0.2
	くらい	6.6	で	13.4	と	0.1	も/また	0.1
	とか	5.1	と	6.4	まで	0.1	は/また	0.0

表12　動詞要素別統計 (毎日新聞)

	使役等		時制等		判断等		接続	
出現頻度	137,359		548,362		30,648		251,943	
パターン数	98		275		661		499	
90%到達位	5		8		84		29	
95%到達位	9		12		139		48	
99%到達位	26		33		375		123	
99.9%到達位	51		107		631		300	
上位5位 & 割合 (%)	られる	72.6	た	56.2	だろう	9.7	て	28.7
	させる	8.8	ている	15.2	R そうだ/だ	8.8	と	14.8
	たい	5.4	ない	7.2	なければならない/ない	4.9	が	12.4
	てくれる	2.5	ている/た	5.7	べき	4.0	という	6.6
	てもらう	1.5	てくる/た	2.1	らしい	3.8	ように	2.8

表11より格助詞等の出現頻度数の高さ，表12より時制等の出現頻度数の高さ，判断等のパターン数の多さが目に付く．また，副助詞 (後) の「の」，提題助詞の「は」の割合が特に大きいことと，使役等の「られる」，時制等の「た」が50%を越えていることが分かる．

3.3 辞書用例文等の大量データに対する調査

比較的短文が多い辞書用例文等の大量データ (約 15 万文) に対して，前節と同様な出現頻度分布の調査を行った．

表 13 文節カテゴリー統計 (辞書用例文等)

文節カテゴリー	説明	出現頻度	割合
N	名詞	461,757	48.99%
P1	動詞	262,726	27.87%
A	副詞	44,968	4.77%
SN	形式名詞	41,638	4.42%
P2	ダ系	35,084	3.72%
P3	形容詞	34,389	3.65%
T	連体詞	28,197	2.99%
P4	形容動詞	25,690	2.73%
C	接続詞	2,468	0.26%
UN	未知語	2,126	0.23%
Q	」	1,130	0.12%
GF	「	1,050	0.11%
TO	引用機能語	768	0.08%
I	感動詞	559	0.06%
合計		942,550	100.00%

毎日新聞記事では 60%強を名詞が占めていたが，辞書用例文等では名詞の割合は 50%を切っている．これは文の長さが短いということの結果である．

表 14 文節パターン統計 (辞書用例文等)

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	461,757	342	11	21	83	209
	SN	41,638	124	8	11	26	87
	Q	1,130	50	9	13	39	49
	TO	768	32	8	11	25	32
用言系	P1	262,726	3,571	134	344	1,469	3,309
	P2	35,084	659	47	97	342	624
	P3	34,389	554	20	50	259	520
	P4	25,690	728	30	182	472	703
他	A	44,968	29	1	4	6	12

表 15 名詞要素別統計 (辞書用例文等)

	副助詞 (前)		格助詞等		副助詞 (後)		提題助詞	
出現頻度	2,228		255,390		77,123		95,903	
パターン数	21		115		17		5	
90%到達位	11		6		1		1	
95%到達位	13		12		1		2	
99%到達位	17		35		3		2	
99.9%到達位	21		76		9		4	
上位 5 位 & 割合 (%)	だけ	27.6	を	33.9	の	97.1	は	91.7
	ばかり	18.0	に	23.6	も	1.4	も	8.1
	くらい	8.4	が	20.7	とは	0.8	も/の	0.1
	さえ	7.9	で	6.9	と	0.2	は/また	0.1
	しか	6.9	と	3.8	まで	0.1	も/また	0.0

辞書用例文等の体言系の機能語部パターン数は毎日新聞記事と比べてあまり変わりはないが，用言系ではおおよそ 1/3 程度である．辞書用例文等では用言系は限られたパターンが現れているということが分かる．これは，用言系の要素別統計にも同じことが言える．

表 16 動詞要素別統計 (辞書用例文等)

	使役等		時制等		判断等		接続	
出現頻度	25,245		121,689		11,352		76,031	
パターン数	75		199		302		290	
90%到達位	8		9		54		21	
95%到達位	14		15		83		33	
99%到達位	29		39		189		101	
99.9%到達位	56		116		291		220	
上位 5 位 & 割合 (%)	られる	57.8	た	49.9	なさい	16.4	て	38.5
	させる	8.4	ている	13.5	だろう	11.2	と	16.2
	てくださる	7.9	ない	11.6	なければ なら/ない	7.4	ように	4.0
	たい	6.6	ます	5.2	ましょう	4.1	が	4.0
	てくれる	5.6	ている /た	3.6	べき	3.5	ても	3.4

3.4 表現の標準化

各要素に着目し，意味的に同様と考えられるものは，出現頻度の高い表現に置き換える表現の標準化を行った．例えば「でしょう」を表現頻度の高い「だろう」に置換をした．この標準化を行った上で，同じ分布調査を試みた．

表 17 標準化後の文節パターン統計 (毎日新聞)

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	3,342,512	438	14	24	79	210
	SN	156,536	193	9	13	38	111
	TO	11,539	84	13	21	39	73
用言系	P1	1,177,915	7,536	106	291	1,700	6,359
	P2	138,625	1,575	83	166	608	1,437
	P3	113,909	1,140	25	69	363	1,027
	P4	117,094	1,559	18	74	568	1,442
他	A	155,487	44	3	5	8	13

体言系の機能語部パターン数はあまり減少しなかったが，用言系の機能語部パターン数はかなり減少した。特に動詞の機能語部パターン数は約 1800 も減少し，原文のパターン数に比べて約 22%も減少した．

辞書用例文中の機能語は，新聞記事中の機能語集合に全て含まれると予想したが，用言に後節する機能語についてはこれは正しくなかった．標準化しない場合で 176 個，標準化した場合で 110 個が含まれていなかった．用言後接機能語の多様性が伺われる．

4 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．また 2 種の大規模コーパスを解析し，機能語部に関する種々の出現頻度統計を行った．

この研究の一部は科学技術振興事業団 (JST) の戦略的基礎研究事業 (CREST)(チームリーダ:池原悟 鳥取大教授) の支援を受けており，辞書用例文等 15 万文はこのプロジェクトが作成した対訳コーパスの日本語部分である．

参考文献

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

[2] 大規模コーパスにおける文パターンの分布調査 高木, 他 FIT 情報科学技術フォーラム 講演論文集 (2002)