文節構造解析システム ibukiC

伊佐治 和哉，山田 佳裕，石原 吉晃，高松 大地，松本 忠博，池田 尚志
岐阜大学工学部

1　はじめに

日本語には，自立語と機能語からなる文節という構文単位があり，文は文節の列として構成される．

我々は，日本語解析システム IBUKI の開発を行っている．IBUKI には形態素･文節解析システム ibukiK，文節構造解析システム ibukiC，構文解析システム ibukiS があり，開発・整備を続けている．本稿では ibukiC を中心に述べる．

ibukiK は形態素・文節単位で切り出し，品詞等の情報を付与する．ibukiC は ibukiK の出力する機能語部をさらに解析し，機能語部を意味的・機能的な観点からいくつかの要素に分割し，また構文解析の観点から必要に応じて文節を再分割する．さらに，標準的な言い回しに言い換えるなどの指定もできる．ibukiS は ibukiC の出力する文節構造列を入力として構文解析を行う．

解析システムの機能語辞書には，解析の容易さと高精度化，および解析結果の応用の際の扱いやすさを考慮して，新聞記事１年分を解析して文節機能語部を取り出し，そのうちの一定以上の頻度を持つ約１万の機能語部をそのまま機能語辞書のエントリーとした．自立語辞書には，フリーで公開されている ICOT の形態素辞書の自立語やフリーの固有名詞データ，我々が日常的に収集した語彙などをベースに作成した．(約 18 万語)

また，構築した解析システムの評価実験として京都大学テキストコーパスを参照データとし，解析結果の比較を行った．

2　日本語解析システム IBUKI

IBUKI の概要を図 1 に示す.

2.1　ibukiK

形態素・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コストを与え，コスト最小法により最適解を決定している．ibukiK では一般的な形態素解析システムのような単語やその連接にコストを与えるのではなく，文節や文節間にコストを与えて解析を行うことで，より構文的・大域的な接続規則を与えて解析を行っている．

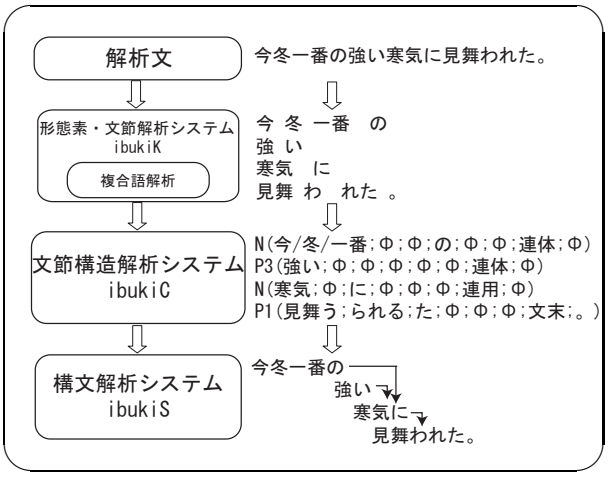

図 1: 日本語解析システム IBUKI の概要

2.2　ibukiC

2.2.1　文節構造

文節構造は「文節カテゴリ (主に自立語の品詞を表す)」,「自立語」,「機能語部を最大６つに分割した各要素」,「係り先情報」,「句読点」の情報から構成することとした (図 2).

図 2: 文節構造

2.2.2　文節カテゴリ

ibukiC では，文節構成単語を走査し，文節カテゴリを付与する．文節カテゴリは主に自立語の品詞を表しており，表 1 に示すようなものがある．

2.2.3　機能語部を各要素に分割

機能語部を最大６つの要素に分割する．機能語部の分割は，表 2 に示すような意味的・機能的な観点からの分割であり，語順は保たせている．

表 1: 文節カテゴリ

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
	NUM	数詞文節
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

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
	要素 5	終助詞	ね，な
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	とりたて詞	も，さえ，たり
	要素 4	判断等の助動詞	だ，だろう，らしい
	要素 5	接続助詞	が，のに，ので
	要素 6	終助詞	ね，な

2.2.4 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 3 に示すように，10 種類がある．

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	直後		

「並列/連用」などのように，係り先が一意に決まらない曖昧な場合は複数の係り先を表記している．

2.2.5 文節の分割

名詞述語文や用言の名詞化文節など，用言文節と体言文節が相互に転化するような場合は，構文解析の観点から，文節を分割して整えた方がよい場合がある．以下のような場合，ibukiK の文節を再分割した．この分割は辞書上で指定できる．

- ダ系
  例：彼+だけ+だったが　→　(彼; だけ) (だ; た; が)
  　このように体言文節に判定詞「だ」を含む機能語が後接すると，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「である」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するようにした．
- 形式名詞
  例：謝罪する+こと が　→　(謝罪する)(こと; が)
  　これは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例である．このような文節は，文節区切りを行い，形式名詞以下を体言文節として分割した．
- ノ系
  例：私+の だけは　→　(私; の)(の; だけ; は)
  　この文節は「わたしの (もの) だけは」という意味を持っているが,「もの」は省略している文である．これを何かが省略されている「の」とし「ノ系」と定義して文節を分割することとした．
- カ系
  例:走る+かどうか が →　(走る)(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もできるが，例のように名詞化する場合があるので，新たに「カ系」と定義し,「か」や「かどうか」の後に体言後接語が続く場合にはカ系文節として文節を区切ると定義した．

2.2.6 解析結果の例

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．

- 私のだけは入賞だよ．

```
0;0;N; 私; Φ; Φ; Φ; の; Φ; Φ; 直後; Φ
0;1;N; の; だけ; Φ; Φ; Φ; Φ; は; Φ; 連用; Φ
1;0;N; 入賞; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ
1;1;P2; だ; Φ; Φ; Φ; Φ; Φ; よ; 文末;.
```

- 選挙を実施できたことがイラク移行政府の正統性につながるでしょう．

```
0 0 N(選挙; Φ; を; Φ; Φ; Φ; Φ; 連用; Φ)
1 0 N(実施; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
1 1 P2(できる; Φ; た; Φ; Φ; Φ; Φ; 直後; Φ)
1 2 SN(こと; Φ; が; Φ; Φ; Φ; Φ; 連用; Φ)
2 0 N(イラク/移行/政府; Φ; Φ; Φ; の; Φ; Φ;
連体; Φ)
3 0 N(正統性; Φ; に; Φ; Φ; Φ; Φ; 連用; Φ)
4 0 P1(つながる; Φ; Φ; Φ; だろう; Φ; Φ; 文
末;. )
```

ibukiC では，2 つ目の例のように，機能語要素の「でしょう」を「だろう」に置換する，標準的な表現などへの言い換えの指定もできる．

2.3 DLL

ibukiK・ibukiC はライブラリ (DLL) 化を行っており，様々な応用アプリケーションの開発が可能である．ibukiC の解析用インターフェースは，このライブラリを利用して作成した．

3 解析用辞書

3.1 機能語辞書

機能語辞書には，新聞記事を解析し，実際に解析結果に現れる文節機能語部をそのまま登録することを試みた．

3.1.1 登録する機能語の単位

一般に文節機能語部は自立語および機能語間の接続規則によって連接可否を判定することで解析を行っている．

機能語として辞書に登録する単位は様々であり,「て」「しまう」「た」のように短い単位で登録することも可能であるし,「てしまった」のように長い単位で登録することもできる．前者の場合，登録する機能語はごく少数で済むが，接続規則の設定が複雑になり，誤った機能語部の表現を生成してしまう可能性がある．また，機械翻訳等の応用システムでは，それらを要素合成的に処理しなくてはならなくなり，複雑になる．一方，後者の場合，接続規則は単純になり，意味的扱いが錯綜し複雑にはならないという点では有利である．しかし，登録すべき表現が膨大となって，現実的ではなくなる可能性がある．

我々は，実際に出現する機能語部は有限の数に収まると仮定し，新聞記事から文節機能語部を収集し，そのままの形で機能語辞書に登録する試みを行った．

3.1.2 新聞記事１年分を解析

長単位の機能語辞書を作成するために，毎日新聞 2000 年度の記事 1 年分（約 67 万文，約 1400 万文節）を長単位機能語辞書を採用する以前の IBUKI で解析した．表 4 に，1 年分の解析結果を示す．

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．登録しなかった機能語は，長単位の機能語を接続することで解析することになるが，その接続コストは高めに設定した．再度新聞記事１年分を解析したところ，この規則により，未登録である出現頻度の下位約１万語の約 85 %に対しても正しく解析を行うことができた．

表 4: 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

3.1.3 作成した機能語辞書

機能語辞書に登録した機能語の詳細を表 5 に示す．

表 5: 出現頻度別機能語登録数

出現頻度	語数	例
1,348,014	1	の
892,213	1	を
646,579	1	に
...		
4	828	ないより
3	1,255	そうだなと
2	2,007	なって下さい
合計	10,364	−

登録語の文字長の平均は 5.539 文字で，最大は 16 文字の「なければならなくなるかもしれない」であった．

3.1.4 他の年度の新聞記事の解析

機能語辞書には 2000 年度の新聞記事から収集した機能語部を登録したのであるが，同様に 97〜99 年度の記事の解析結果もあわせて機能語部の統計を行ったところ，表 6 に示すような結果となった．

表 6: 97 年度から 2000 年度までの機能語部の統計

機能語部数 (異なり)	機能語部数 (述べ)	カバー率 (%)
1,868	33,283,136	99.0
3,584	33,451,191	99.5
8,122	33,552,021	99.8
13,988	33,585,631	99.9
37,855	33,619,250	100

出現頻度が 1 のものが異なり数で 16,351 件あるため，約 38,000 件の機能語部が出現したが，出現頻度の上位約 14,000 件で全体の 99.9 %の文節を網羅できることは同じであった．よって実際に現れる機能語部は「ほぼ有限」の数に収束することが分かった．

3.2 自立語辞書

自立語辞書は，フリーで公開されている ICOT 形態素辞書中の自立語 (約 15 万語，表 7 の情報を持っている) やフリーの固有名詞データ (人名，地名，駅

名など)，また我々が日常的に収集した語彙などを集積して作成した．以下のような辞書項目を持たせている．

表 7: ICOT 形態素辞書が持つ単語情報

情報	説明
見出し	単語の見出し
読み	単語の読み
形態素コード	品詞情報

- 辞書登録番号
  各単語に通し番号を付け，管理している．
- 品詞情報
  IBUKI では，単語の品詞およびその細分類を左右連接属性対で表現している．ICOT 形態素辞書の品詞情報に関しては，それを左右連接属性対に変換した．
- 文字種コード
  各単語が，どのような文字で構成されているかを表現する値である．文字の種類はひらがな，カタカナ，漢字，数字，アルファベットである．ひらがな＋カタカナ等の表記も一つの種類とし，合計で３３種類の文字種に分類した．
- コスト
  文節コストのもとになる値を付与した．
- 読みコスト
  各単語の読みが複数存在した場合，その読みの優先度を示す値を追加した．
- 読み (点訳表記)
  データ中の読み情報を，我々が開発している自動点訳システム IBUKI-TEN で利用するための点訳表記に変換した．単語内を区切るかどうかを示す，分かち書き情報を含む仮名表記を登録した．(例：あくせん/くとー)

4 評価実験

ibukiC のシステムの解析精度を評価するため，京都大学テキストコーパス Version3.0(以下，京大コーパスと略す) を参照データとし，解析結果の比較を行った．京大コーパスの文節開始文字位置と，解析結果の文節開始位置が一致した場合，その直前の文節は正解であるとした．

なお，今回は文節切りの評価を目的としており，形態素の品詞等は考慮していない．

4.1 評価尺度

文節解析の精度を調べるために，以下の尺度を用いた．

- C_Count：京大コーパスの文節区切り数
- I_Count：ibukiC の文節区切り数
- R_Count：文節の区切りが一致する数

$$\text{適合率} = \frac{R\_Count}{I\_Count} * 100$$

$$\text{再現率} = \frac{R\_Count}{C\_Count} * 100$$

$$F \text{ 尺度} = 2/(\frac{1}{\text{適合率}} + \frac{1}{\text{再現率}})$$

4.2 実験結果

評価実験の結果を表 8 に示す．

表 8: 文節解析精度

C_Count	I_Count	R_Count	文節解析 適合率	文節解析 再現率	F 尺度	文節解析 異なり個所
333,653	307,002	301,241	98.1235	90.2857	94.0416	5,761

4.3 考察

適合率に比べて再現率が低い結果となった．ibukiC の機能語辞書には，(年頭)+「にあたり」のような，長めの単位の機能語を登録し，1 文節としている．これに対して京大コーパスでは，(年頭に)(あたり) のような 2 文節で表現されており，再現率に影響している．

5 おわりに

長単位の機能語辞書をベースに機能語部の内部構造を逆に分解・出力でき，あるいは標準的な表現などに言い換えて出力できるなどの機能を持った文節構造解析システム ibukiC を開発した．なお，ibukiK，および ibukiC は近日 Web 上*で公開する予定である．今後は解析精度の向上のために，自立語辞書においては専門語彙の追加や一般語彙の充実，機能語辞書においては個々の登録語の見直しなどを進めていく予定である．

参考文献

[1] 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, 山田, 池田　言語処理学会 第 10 回年次大会 発表論文集 (2004)

[2] 京都大学テキストコーパス Version3.0 東京大学 西田・黒橋研究室 (http://www.kc.t.u-tokyo.ac.jp/)

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