文節構造解析システム ibukiC の解析仕様および精度の比較と評価

山田 佳裕，脇田 貴之，大口 智也，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 を開発している．またそれの点字翻訳システム IBUKI-TEN[5]，機械翻訳システム jaw[4] などへ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][2][3]，今回システムや自立語辞書，機能語の構造分割部，接続属性，規則等を改編し整備した．本稿では，これらの改編・改良について述べ，さらに京都テキストコーパスやその他のコーパス，解析システムとの解析仕様の違いについて比較する．また，解析精度の評価実験について述べる．

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．まず，形態素・文節解析システム ibukiK において形態素および文節に分割し，その後 ibukiC にて文節構造 (図 2) を付与している (図 1)．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．詳細は参考文献 [1][2][3] 参照．次に，今回の主な改良点について述べる．

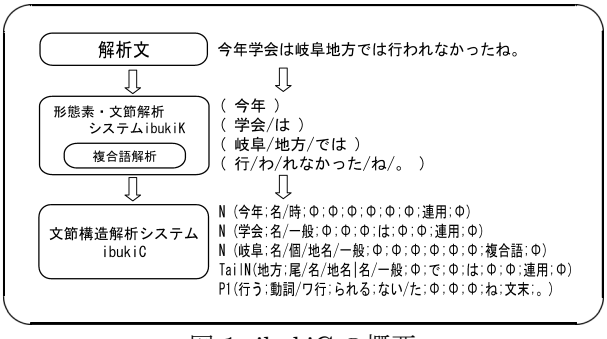

図 1: ibukiC の概要

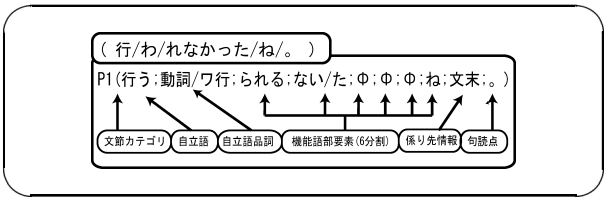

図 2: 文節構造

2.1 文節分割部の改良

2.1.1 文節分割処理

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．

(だ) のほか，現在，延べ 4,848 語，272 種の文節分割処理を行っている．例を図 3 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．

```
(動/い/た/からでしょう/。)

0;0;P1(動く; 動/カ行; Φ; た; Φ; Φ; Φ; Φ; 直後; Φ)
0;1;P2(Y からだ; 機/用―文末; Φ; Φ; Φ; でしょう; Φ; Φ; 文末;。)

(彼/か否かが)(問題/だ/。)

0;0;N(彼; 名/代/人; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
0;1;KA(T か否か; 機/名―連用; Φ; が; Φ; Φ; Φ; Φ; 連用; Φ)
1;0;N(問題; 名/一般; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
1;1;P2(T だ; 機/名―ダ系文末; Φ; Φ; Φ; Φ; Φ; Φ; 文末;。)

(開/く/しかない/。)

0;0;P1(開く; 動/カ行; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
0;1;P3(Y しかない; 機能辞/動詞/終止形; Φ; Φ; Φ; Φ; Φ; Φ; 文末;。)

(失敗/する/はずはない/。)

0;0;N(失敗; 名/サ; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
0;1;P1(T する; 機/活尾; Φ; Φ; Φ; Φ; Φ; Φ; 直後; Φ)
0;2;P3(Y はずはない; 機能辞/動詞形容詞/連体形; Φ; Φ; Φ; Φ; Φ; Φ; 文末;。)
```

図 3: 文節分割の例

2.1.2 複合語の処理

複合語は，まず形態素解析の際に漢字列として抽出し，その漢字列に対して改めて複合語となる形態素の組み合わせを解析する複合語解析を行っている．

今回 ibukiC において，その複合語となる単語一つ一つに対して，文節構造を与えた (下記例参照)．このとき，文節構造の係り先に「複合語」と出力することで他の文節構造と区別している．

複合語内部の係り受け解析は今後の課題である．

2.1.3 解析例

ibukiC が出力する解析結果の出力例を以下に示す．

例） (佐藤/太郎/さん/に)(会/っ/た/。)

```
＜従来＞
0;0;N(佐藤/太郎/さん; 複合語; Φ; に; Φ; Φ; Φ; Φ; 連用; Φ)
1;0;P1(会う; 動/ワ行; Φ; た; Φ; Φ; Φ; Φ; 文末;。)

＜改良後＞
0;0;N(佐藤; 名/個/人名/姓; Φ; Φ; Φ; Φ; Φ; Φ;複合語; Φ)
0;1;N(太郎; 名/個/人名; 名; Φ; Φ; Φ; Φ; Φ; Φ;複合語; Φ)
0;2;TailN(さん; 尾/名/人名; Φ; に; Φ; Φ; Φ; Φ; 連用; Φ)
1;0;P1(会う; 動/ワ行; Φ; た; Φ; Φ; Φ; Φ; 文末;。)
```

例） (応用/情報/学科/の)(友人/。)

＜従来＞
0;0;N(応用/情報/学科; 複合語; Φ; の; Φ; Φ; Φ; Φ; 連体; Φ)
1;0;N(友人; 名/一般; Φ; Φ; Φ; Φ; Φ; Φ; 文末;。)

＜改良後＞
0;0;N(応用; 名/サ; Φ; Φ; Φ; Φ; Φ; Φ;複合語; Φ)
0;1;N(情報; 名/一般; Φ; Φ; Φ; Φ; Φ; Φ;複合語; Φ)
0;2;N(学科; 名/一般; Φ; の; Φ; Φ; Φ; Φ; 連体; Φ)
1;0;N(友人; 名/一般; Φ; Φ; Φ; Φ; Φ; Φ; 文末;。)

3 京都テキストコーパスとの比較と評価

3.1 はじめに

ibukiC の解析精度の評価として，京都テキストコーパス (以下，K コーパス) との比較実験を行った．

K コーパスは人手で修正済みのものであり，正解データとして考えられるが，形態素や文節の区切り，品詞等について考え方が異なっている場合もあり，解析結果が一致しないことが直ちに ibukiC の誤りということにならない．今回は,「文節区切り」のみを対象として精度の比較，違いの分析を行った．

3.2 京都テキストコーパス

K コーパスは，毎日新聞 95 年 1 月 1 日から 17 日までの全文記事約 2 万文，1 月から 12 月までの社説記事約 2 万文の，計約 4 万文に対して，形態素解析システム JUMAN，構文解析システム KNP で自動解析を行い，その結果に対し，さらに人手で修正を施したテキストコーパスである

```
# S-ID:950101003-001 KNP:96/10/27 MOD:2005/03/08
* 0 26D
村山 むらやま * 名詞 人名 * *
富市 とみいち * 名詞 人名 * *
首相 しゅしょう * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 2D
年頭 ねんとう * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 6D
あたり あたり あたる 動詞 * 子音動詞ラ行 基本連用形
```

図 4: 京都テキストコーパスより一部抜粋

3.3 ibukiC の解析仕様と京都コーパスとの比較

ibukiC では機能語をできるだけ長い単位で扱っていたり，前節で述べた文節分割処理を行うなどの処理を行っているため，解析結果に違いが出てくる．次に例を示す.

(1) 繋辞文など ibukiC で文節分割する文節
(2) 括弧 (「」() 等) の処理
(3) 「このため」「このところ」などの指示語を含めた語
(4) 「気が付く」「誇り高い」などの慣用句として扱う語
(5) 「〜にあたり」「〜に対して」などの複合辞

(1) については，ibukiC では (本)(だ) と 2 文節で解析しているのに対し，K コーパスは (本だ) と 1 文節で解析している．

(2) は，括弧があった場合の解析処理が次のように異なっている．

K コーパス : (「/諸悪/の)(根源/」/の/ように)
ibukiC : (「)(諸悪/の)(根源)(」のように)

K コーパスでは，括弧と括弧の中の語とを同じ文節として表しているため，2 文節で解析している．しかし，ibukiC の場合は，括弧は括弧内の語と別の文節で表すようにしているため 4 文節として解析している．

(3)(4) では，辞書登録の違いにより，次のような解析の違いがある．

K コーパス : (この)(ため)(この)(ところ)
ibukiC : (このため)(このところ)

K コーパス : (気が)(付く)(誇り)(高い)
ibukiC : (気が付/く)(誇り高/い)

ibukiC では比較的長い単位での辞書登録が多いため，指示語や，慣用句的な扱いをする語については長い単位で登録している．このため文節数が K コーパスより少ない．

(5) の「〜にあたり」「〜に対して」といった機能的な役割をする語については，ibukC ではこれを複合辞として 1 文節として解析しているが，K コーパスでは 2 文節として解析している．

K コーパス : (年頭/に)(あたり)
ibukiC : (年頭/にあたり)

これについては，複合辞ではなく自立語として使用される場合もあるため問題が生じる．例えば，例の「年頭にあたり」の場合は，機能的な「〜にあたり」

で正しいが,「ボールにあたり」といった場合は (ボールに)(あたり) のように「あたる」と自立語として解析しなければならない.

現在 ibukiC では，こうした複合辞を含む文は全て複合辞として解析している．対して K コーパスでは，全て自立語として解析しているようである.

3.4 評価の方法

K コーパスの文節の区切りを正解として，ibukiC のそれとを比較することで精度を評価した．ただし，前節で述べた点については次のように対処した.

(1)' 文節分割する前 (ibukiK) を比較対象とする
(2)' 括弧の前後の文節が正しければ正解とみなす
(3)' 正解とみなす
(4)' 正解とみなす
(5)' 人手で正解を判定

(5)' に関しては，機械的には判定できないため，人手で正解かどうかを比較することにした.

互いの文節区切りが一致する文節数と正解である文節数が異なるため，下の評価式で評価を行った．①②は，K コーパスとの違いを考慮してそれぞれに計算した数値を使っている.

$$\text{適合率} = \frac{\textit{ibukiC} \text{ の正解文節数①}}{\textit{ibukiC} \text{ の文節数}} * 100$$

$$\text{再現率} = \frac{\textit{ibukiC} \text{ の正解文節数②}}{K \text{ コーパスの文節数}} * 100$$

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

3.5 評価実験

実験結果を表 1 に示す．前節までに述べた解析の仕様を考慮した結果，96 %強の精度があった.

表 1: K コーパスと比較した解析精度

K コーパス文節数	②	ibukiC 文節数	①	適合率 (%)	再現率 (%)	F 尺度 (%)
372,130	362,060	373,619	355,690	95.20	97.29	96.23

正解とみなした文節例について次の表 2，3，4 に示す。複合辞については，人手で判定した結果，およそ 99 %程度 ibukiC の解析結果が正しかった (約 7,405 文節中 7,312 文節).

表 2: 指示語例 (延べ 707 文節　異なり 50 種)

K コーパス	ibukiC
(こう)(なれば)	(こうな/れば)
(この)(中/で)	(この中/で)
(この)(ほど)	(このほど)
(この)(とき)	(このとき)
(その)(ため/には)	(そのため/には)
(その)(一/つ/が)	(その一つ/が)
(その)(うえ)	(そのうえ)
(それ)(故)	(それ故)
(あの)(まま)	(あのまま)
(あの)(ころ/の)	(あのころ/の)
(ある)(日)	(ある日)
(どう)(なるか)	(どうな/るか)
(どう)(やって)	(どうやって)
(どの)(くらい)	(どのくらい)

表 3: 慣用句例 (延べ数 233 文節　異なり数 71 種)

K コーパス	ibukiC
(史上)(初)	(史上初)
(虫/の)(息)	(虫の息)
(火/の)(気)	(火の気)
(年/の)(瀬)	(年の瀬)
(ほど)(よい)	(ほどよ/い)
(より)(よい)	(よりよ/い)
(きめ)(細かい)	(きめ細か/い)
(気/の)(早い)	(気の早/い)
(気/に)(なる)	(気にな/る)
(身/に)(つける)	(身につけ/る)
(悦/に)(入る)	(悦に入/る)
(あざ)(笑う)	(あざ笑/う)
(あり)(得る)	(あり得/る)

表 4: 複合辞例 (延べ 7,312 文節　異なり 474 種)

K コーパス	ibukiC
(改正/に)(ついて)	(改正/について)
(教団/に)(とって)	(教団/にとって)
(発行/に)(よって)	(発行/によって)
(長年/に)(わたって)	(長年/にわたって)
(政府/に)(対し)	(政府/に対し)
(スポーツ/を)(通じて)	(スポーツ/を通じて)
(距離/を)(おいて)	(距離/をおいて)
(決定/を)(めぐり)	(決定/をめぐり)
(米国/を)(はじめ)(各国/は)	(米国/をはじめ)(各国/は)
(構え/で)(いる)	(構え/でいる)
(進める)(ほか)(ない)	(進め/る/ほかない)
(ばれる)(はず/が)(ない)	(ばれ/る/はずがな/い)
(アジア/と)(いう)(地域)	(アジア/という)(地域)
(通じた)(あげく)	(通じ/たあげく)
(広がり/つつ)(あった)	(広が/り/つつあった)

3.6 考察

K コーパスと比較した結果，主に次のような誤解析が見られた.

1. ibukiC または K コーパスの誤解析
2. ibukiC の辞書登録が無いための誤解析

1. は次のように時相名詞を含む場合が多かった.

K コーパス：(毎日/聖書/を)(昨年/暮れ/に)
ibukiC：(毎日)(聖書/を)(昨年)(暮れ/に)

上記の時相名詞は ibukiC のように単独の文節として 2 文節の方が良い．しかし，ibukiC では次の例のように複合語として同一文節として扱うべき文節ま

で分割してしまう誤解析があった.

K コーパス : (日常/生活/に)(戦後/処理/の)
ibukiC : (日常)(生活/に)(戦後)(処理/の)

このように時相名詞に関連して，K コーパスか ibukiC のどちらかが間違っていると思われる文節が約 400 文節あった.

また，2. については例えば次のようなものがあった.

K コーパス : (年老いた)(見直し/問題/は)
ibukiC : (年)(老い/た)(見直/し)(問題/は)

これらは辞書登録語によって改善可能である．ibukiC が正解とならなかった文節の多くが辞書登録に関する問題であった.

4 その他のシステムやコーパスとの比較実験

4.1 はじめに

K コーパスの他に，日英対訳コーパス [6](以下，C コーパス) とも同様に評価を行い精度評価実験を行った．また，他の解析システムとの比較として KNP との比較実験も行った.

4.2 C コーパスとの比較実験

4.2.1 C コーパス

C コーパスは，約 12 万文を機械解析した後，人手で必要な修正を施したものである.

```
INPUT=AC000004-00=
彼のお母さんがああ若いとは思わなかった.
1. /彼 (1710,{NI:23,NI:48,IM:01131,IM:01211})
2. +の (7410)
3. /お母さん (1100,{NI:80,NI:49,IM:01212,IM:01220})
4. +が (7410)
5. /ああ (1110,{NI:1235,IM:05000})
6. /若い (3106,{NY:5,NY:5,IY:9300})
7. +と (7420)
8. +は (7530)
```

図 5: C コーパスの例

4.2.2 ibukiC の解析仕様と C コーパスとの比較

K コーパスと同様，括弧や複合辞，慣用句などの部分で解析仕様が異なっていた.

4.2.3 評価実験

K コーパスとの比較同様，解析仕様を考慮して比較実験を行った．結果を表 5 に示す.

表 5: C コーパスと比較した解析精度

C コーパス 文節数	②	ibukiC 文節数	①	適合率 (%)	再現率 (%)	F 尺度 (%)
714,093	690,650	684,510	647,615	94.61	96.71	95.65

誤解析の多くが，ibukiC の辞書に登録が無いためであった.

C コーパス : (焼きのり/の)
ibukiC : (焼/き/の)(りの)

また K コーパス同様，時相名詞を含む場合などに，C コーパスが間違っていると思われるものがあった.

C コーパス : (そのとき/彼/が)
ibukiC : (そのとき)(彼/が)

4.3 KNP との比較実験

次に，代表的な構文解析システムである KNP の文節まとめ上げ結果との比較を行った．K コーパスと同じ新聞記事を KNP で解析した結果とを比較した.

KNP は，K コーパスの元になっているシステムであるため，解析仕様の違いは K コーパスと同様である．実験結果を表 6 に示す.

表 6: KNP と比較した解析精度

KNP 文節数	②	ibukiC 文節数	①	適合率 (%)	再現率 (%)	F 尺度 (%)
371,929	360,677	373,619	354,548	94.89	96.97	95.92

5 おわりに

各コーパスと比較実験した結果，高い精度があると考えられる.

誤解析の原因として辞書登録語に関する問題が多いため，今後は登録語の拡充をはじめとした辞書の整備を行い，さらなる精度向上を目指したい.

参考文献

[1] 山田, 松本, 池田，文節構造解析システム ibukiC における文節分割について，平成 18 年度電気関係学会東海支部連合大会, 2006

[2] 山田, 髙松, 石原, 水野, 大口, 佐藤, 松本, 池田，日本語文解析システム ibukiC/S について，言語処理学会 第 12 回年次大会, 2006

[3] 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

[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

[6] 村上仁一，池原悟他, 日本語英語の文対応データーベースの作成，第 7 回 LACE 研究会 pp1-10，2002

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