# 文節の内部構造解析と出現頻度統計

一ノ瀬 友紀夫，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．本論分は，文節にどんなパターンがあるのか，文節列にはどんなパターンがあるのか，ということを調べる試みである．

これらのパターンの数が扱える範囲内のものであるならば，パターンデータベースに基づく種々の言語処理の可能性に期待ができることになる．

毎日新聞記事 ('99) 1年分 (約100万文) を文節解析し，パターンを調べた結果を報告する．

## 2 文節構造解析

文節は自立語と機能語からなるが，自立語と機能語の区分は必ずしも明確ではなく，いろいろな分け方が考えられる．例えば，

①NPOとして活動しているに違いない

に対して，次の二つは考えられる分析の例である．

②(NPO と)(し て)(活動 し て いる に)(違い ない)

③(NPO として)(活動し ている に違いない)

我々は，あまり細かな語構成に分割するよりも，意味的に考えて，大きめの意味的なまとまりにするほうを選んだ．そのほうが，その後の構文解析や機械翻訳等への応用に有利であると考えたからである．上の例では，③のように分析する．

さらに，機能語の部分はその働きを考慮していくつかの要素に分割することとし，文節構造として図1のような形に分析することとした．

文節カテゴリーは以下の15種類を設けた．

体言系文節 (5種類)

N:名詞文節，SN:ノ系文節，KA:カ系文節
Q:」文節 (引用の終わり)，TO:ト系文節

用言系文節 (4種類)

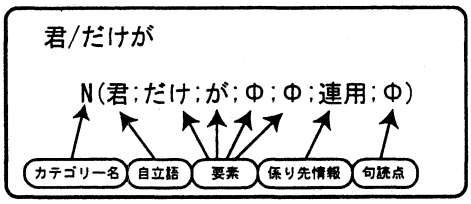


図 1: 文節構造情報

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

基本的には自立語部の品詞を用いているが，名詞に「ダ，カモシレナイ」などがついて名詞述語分となる場合には，

・「太郎かもしれない」→ (太郎)(かもしれない)

のように，名詞文節とダ系文節に分割した．動詞に「の，こと」などの形式名詞が付いて名詞化する場合には，

・「言ったのは」→ (言った)(のは)

のように，動詞系文節とノ系 (形式名詞) 文節に分割した．また，「カ，カドウカ」などはダ系の疑問形と考えることができるが名詞化する場合もあることよりカ系文節としてあつかった．

・「君かどうかが」→ (君)(かどうかが)

機能語の要素の分割は，名詞系では格助詞相当語を中心として4つの構成部分に分けた．

①格助詞相当語の前に位置する副助詞等
②格助詞相当語
③格助詞相当語の後ろに位置する副助詞等，「の」
④提題助詞「は，も」

用言系では次の4つの構成部分に分けた．

①受身使役等の助動詞 (させる，られる ...)

②時制，肯否等の助動詞 (ない，なかった ...)
③判断等の助動詞 (だろう，に違いない ...)
④接続助詞 (けれども，ので ...)

係り先情報はその文節がどのような属性で後ろの文節に係っていくかという情報であり，以下のような 12 種類がある．

> 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．

また，文節構造解析では簡単な一般化という作業もおこなった．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言う．

## 3 文節パターンの出現頻度統計

毎日新聞記事 ('99) 1 年分 (約 100 万文) を文節構造解析を行い，2 節で述べたような形の文節構造データを集積した．以下の点に注目し統計をおこなった．

①比較統計 : 文節構造解析によって機能語部パターン数はどれくらいに収束するか．

②文節カテゴリー統計 : 各文節カテゴリーはどのくらいの頻度で出現しているか．

③機能語部統計 : 機能語部はどれくらいのパターン数があるか．体言後接機能語と用言後接機能語の比較．

文節パターンの例を以下に示す．文節パターンとしては自立語や係り先情報，句読点情報を抜き機能語部だけに注目をした．

> - 「君に関しては,」
>  → N(Φ; にかんして; Φ; は)
> - 「動いていたらしいので」
>  → P1(Φ; ている/た; らしい; ので)

**比較統計**　文節構造解析により機能語部はどれだけまとまるのかを，文節構造解析を使わない場合の機能語部パターン数，使った場合の機能語部パターン数を比較し，検証した．表 1 に示す．

| 項目 | 機能語 | 文節構造 |
|---|---|---|
| 文数 | 1046401 | 1046401 |
| 文節数 | 9769449 | 10166120 |
| 機能語部パターン数 | 34428 | 15347 |
| 90%到達位 | 61 | 53 |
| 95%到達位 | 222 | 129 |
| 98%到達位 | 1063 | 410 |
| 99%到達位 | 2788 | 887 |
| 99.5%到達位 | 6476 | 1785 |
| 99.9%到達位 | 24610 | 6729 |
| 残りパターン数 | 9818 | 8618 |
| 頻度 5 | 929 | 452 |
| 頻度 4 | 1370 | 636 |
| 頻度 3 | 2179 | 1017 |
| 頻度 2 | 4499 | 1967 |
| 頻度 1 | 18149 | 7020 |

表 1: 比較結果

文節構造解析を用いることにより，機能語部パターンは 45%の 15347 パターンに収束した．さらに，文節構造パターンの上位 887 パターンで文節全体の 99%をカバーできることもわかった．

**文節カテゴリー統計**　各文節カテゴリーはどのくらいの頻度で出現しているかの統計をとった．表 2 に示す．

表より，全体の 60%強もの文節が名詞であることがわかった．次に続くのが動詞で 16%を占めた．

**機能語部統計**　機能語部はどれくらいのパターン数があるか．体言後接機能語と用言後接機能語の比較をした．それぞれの機能語を伴う文節の代表として出現頻度の多かった N:名詞と P1:動詞について比較した．表 3 に示す．

名詞の機能語部パターンは約 340 パターンに収束しているが，動詞の機能語部パターンに関しては，名詞よりも文節数が少ないにも関わらず約 9300 パターンもある．

頻度が小さい機能語パターンをみると「でしょう」→「だろう」などの表現の標準化 (言い換え) をおこなえば，パターン数は収束するかもしれないことが期待される．表現の標準化は次の課題である．

| カテゴリー | 説明 | 数 | 割合 |
|---|---|---|---|
| N | 名詞 | 6429487 | 63.24% |
| P1 | 動詞 | 1618924 | 15.92% |
| QF | 「 | 501299 | 4.93% |
| Q | 」 | 435652 | 4.29% |
| A | 副詞 | 223982 | 2.20% |
| SN | 形式名詞 | 202705 | 1.99% |
| P2 | ダ系 | 183551 | 1.81% |
| P3 | 形容詞 | 161569 | 1.59% |
| P4 | 形容動詞 | 156965 | 1.54% |
| T | 連体詞 | 94536 | 0.93% |
| C | 接続詞 | 79486 | 0.78% |
| UN | 未知語 | 43097 | 0.42% |
| TO | 引用機能語 | 25481 | 0.25% |
| KA | カ系 | 5486 | 0.05% |
| I | 感動詞 | 3898 | 0.04% |
| NIL | 解析ミス | 2 | 0.00% |

表 2: 文節カテゴリー統計

## 4 文節列パターンの抽出

文は文節の並びである．どのような文節の並びで文を構成しているのか，文節列パターンはどのように分布しているのかを調べた．まず，様々の条件で文節を抽出するツールを作成した．

このツールによって、次の①，②，③のパターンを抽出した。

①全ての文節カテゴリーの文節構造を出力

②引用文 (文節カテゴリー QF,Q,TO) を含む文を除き，さらに連体構造を除き，用言文節と主要な係り文節の主要な要素部分のみを出力

③引用文を含む文を除き，さらに連体構造を除き，用言文節の文節構造のみ出力

主要な係り文節とは，格助詞相当語あるいは提題助詞を持つ体言系文節のことである．主要な要素部分とは，体言系なら格助詞相当語と提題助詞と句読点情報，用言系なら接続と係り先情報と句読点情報とした．

次の文Aは、①②③により，以下のようにパターン化される．

A．トレーナーを着た松本被告が法廷に入る．

| 項目 | 名詞 | 動詞 |
|---|---|---|
| 文節数 | 6429487 | 1618924 |
| 機能語部パターン数 | 337 | 9304 |
| 90%到達位 | 9 | 73 |
| 95%到達位 | 14 | 233 |
| 98%到達位 | 31 | 827 |
| 99%到達位 | 51 | 1732 |
| 99.5%到達位 | 78 | 3249 |
| 99.9%到達位 | 155 | 7677 |
| 残りパターン数 | 182 | 1626 |
| 頻度 5 | 3 | 282 |
| 頻度 4 | 8 | 398 |
| 頻度 3 | 10 | 642 |
| 頻度 2 | 13 | 1221 |
| 頻度 1 | 12 | 4422 |

表 3: 機能語部統計

① N(Φ; を; Φ; Φ; 連用; Φ)P1(Φ; た; Φ; Φ; 連体; Φ)
N(Φ; が; Φ; Φ; 連用; Φ)N(Φ; に; Φ; Φ; 連用; Φ)
P1(Φ; Φ; Φ; Φ; 文末;. )

② N(が; Φ; Φ)N(に; Φ; Φ)P1(Φ; 文末;. )

③ P1(Φ; Φ; Φ; Φ; 文末;. )

## 5 文節列パターンの出現頻度統計

5節の規則およびツールを用いて，毎日新聞記事('99)10 万文に対して文節列パターンの統計をおこなった．表 4 に示す．有効文数は実際に文節列を抽出した文数である．

収束率には以下の式を与えた．

$$1 - \{(\text{パターン数})/(\text{有効文数})\} \times 100$$

この結果をみると，①の条件では引用文も含んでいるため文構造はほとんど収束しなかった．しかし，②の条件ではパターン数も随分収束し，収束率も50%を超えた．③の条件では用言文節のみを考えたため収束はさらに進み 80%を超えた．しかし，②や③の場合でも頻度 1 のパターンは多数存在していることもわかった．

| | ① | ② | ③ |
|---|---|---|---|
| 有効文数 | 98170 | 74451 | 74246 |
| パターン数 | 88613 | 33287 | 14368 |
| 収束率 | 9.74% | 55.29% | 80.65% |
| 頻度 1 | 86427 | 29124 | 11633 |
| 頻度 1 の割合 | 97.53% | 87.49% | 80.96% |

表 4: 文節列パターン統計 (10 万文)

次に，1 年分ではどのような結果になるのか調べた．表 5 に示す．

| | ① | ② | ③ |
|---|---|---|---|
| 有効文数 | 651580 | 496983 | 495553 |
| パターン数 | 558548 | 173790 | 66506 |
| 収束率 | 14.28% | 65.03% | 86.58% |
| 頻度 1 | 541344 | 149157 | 53181 |
| 頻度 1 の割合 | 96.92% | 85.83% | 79.96% |

表 5: 文節列パターン統計 ( 1 年分)

10 万文時と比べ約 6-7 倍の文数に対してパターン統計をおこなったが収束は期待していたほどは進まなかった．文パターンは多様に存在することを再認識する結果となった．

文構造を考える上で，最も適当であると考えられる②の条件時のカバー率を表 6 に示す．カバー率とは，出現頻度の多いパターンの上位何解で全体のどれだけをカバーできるかを示したものである．

| カバー率 | パターン数 |
|---|---|
| 50% | 1130 |
| 60% | 5740 |
| 70% | 24633 |
| 80% | 74394 |
| 90% | 124092 |

表 6: 文節列パターンカバー率

50%をカバーするには 1130 パターンを考えればよいことが分かった．しかし，カバー率 90%を達成するには 124092 ものパターンを考えなければいけないという結果になった．

## 6 おわりに

大量のテキストについて文節構造解析をおこない，日本語文に表われる文節パターンの出現頻度統計と，文節列パターンの出現頻度統計をおこなった．

接続規則による形態素解析では，規則を完全には記述できないために一般に膨大な数の機能語パターンを許すことになってしまうが，現実には現れるパターンの数は限られたものであることが分かる．名詞系では 337 パターンであるし，全体の機能語部でみても頻度順の上位 6729 パターンで実際に現れる 99.9%をカバーしている．このことは，機能語部については 6729 パターンについて考慮しておけば，1 文 10 文節とすれば，100 文のうち 99 文についての機能語部分に関する処理は間違わないはずであるということを意味する．機械翻訳などの応用において考慮に値する事実である．

文節列パターンは機能語部のパターンのようには高いカバー率を得ることは出来なかったが，文節列の抽出のしかたによってはわずか 1130 パターンで 50%もの文をカバーできることが分かった．

カバー率を上げるためには，文節機能語の標準形への変換 (言い換え) について検討することが課題である．

文節列パターンは文の骨格構造を表現するものであるが，同じ文節列パターンでも構文解析の形が同じになるとは限らない．各文節の意味的条件によって異なる形に構文解析される．たとえば，②の抽出条件の場合，

- 彼らはそのまま国境を越え、オーストリアへの大量脱出が実現した。
  $((N_1$は $N_2$を $P1_1.)(N_3$が $P1_2.))$.
  $N_1$の $N_3$: 彼らのオーストリアへの大量脱出
- カメラは甲子園のグラウンドを写し、最後のナレーションが重ねられる。
  $(N_1$は $(N_2$を $P1_1$, $N_3$が $P1_2.))$
  $N_1$に $P1_2$: カメラにナレーションが重ねられる

同じ文節列パターンに対して構文解析の形を調べ上げておき、その対応条件を分析しておくことで新しいデータ駆動型の構文解析の方法が考えられるかもしれない。