E-42

大規模コーパスにおける文パターンの分布調査
Distribution of the sentence pattern in a large-scale corpus

高木 優紀江† 林 誠悟† 池田 尚志†
Yukie Takagi Seigo Hayashi Takashi Ikeda

1 はじめに

文節構造解析システム ibukiB を用いて大規模コーパス（毎日新聞記事 ('00) 1年分）を解析し，文節構造と文節列パターン（すなわち文パターン）の出現頻度分布について調査した．

文節機能語については，用言系の各要素，体言系の各要素について意味的に同様と考えられるものは同一の表現に置き換えてみた上で同様の出現頻度分布を調査することも行い，文節構造，文節列パターンの収束率の変化について調べた．これらについて報告する．

2 文節構造解析システム ibukiB

文節構造解析システムの特徴は以下の4点である．

- 文節カテゴリー (15種類) を与える．
- 機能語を機能ごと (体言系では格助詞相当語を中心に4つの構成部，用言系でも4つの構成部) に各要素に分割する．
- 「にかんして」「に関して」などの字面は異なる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．

3 文節構造パターンの出現頻度

'00 年毎日新聞記事1年分 (約140万文のうち括弧等を含む文等を除いた約67万文) を文節構造解析して，種々の統計を行った．

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．結果を表1に示す．

表1 文節カテゴリー統計

文節カテゴリー	説明	数	割合
N	名詞	3342512	62.41%
P1	動詞	1177915	21.99%
SN	形式名詞	156536	2.92%
A	副詞	155487	2.90%
P2	ダ系	138625	2.59%
P4	形容動詞	117094	2.19%
P3	形容詞	113909	2.13%
T	連体詞	63021	1.18%
C	接続詞	50762	0.95%
UN	未知語	27152	0.51%
TO	引用機能語	11539	0.22%
I	感動詞	1380	0.03%
NIL	解析ミス	6	0.00%
UP	句点のみ	1	0.00%

表1より全体の文節 (5355966 文節) の 60%強を名詞で占めていることが分かる．次に続く動詞も 20%を占めている．

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．出現頻度の多かった N:名詞パターンと P1:動詞パターンについて比較した．

表2 機能語部統計

	名詞	動詞
文節数	3342512	1177915
機能語部パターン数	329	7698
90%到達位	11	71
95%到達位	18	224
99%到達位	62	1590
99.9%到達位	173	6521
残りパターン数	156	1177

名詞文節の方が出現頻度は大きいが，パターン数は動詞文節と比べてはるかに少ない．また，名詞上位11パターン，動詞では上位71パターンを見るだけで，機能語部の 90%を把握できる．

要素ごとの統計 各要素ごとにどれぐらいのパターン数が存在するか調査した (表3, 表4)．

表3 名詞要素別統計

	副助詞 (前)	格助詞等	副助詞 (後)	提題助詞
出現頻度数	16131	1751046	698173	484812
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

表4 動詞要素別統計

	使役等	時制等	判断等	接続
出現頻度数	137359	548362	30648	251943
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

表3より，格助詞等の出現頻度の高さ，表4より判断等のパターン数の多さが目に付く．

4 文パターン抽出プログラム

文節構造解析の結果から文パターンを組みあげる文節列作成ツールを作成した．文節列を組み立てる際，様々な抽出条件に対応できるようにルールファイルを作成した．流れを図1に示す．

5 文パターンの抽出

4節のツールを用いて次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そのまま出力．
2. 引用文 (TO) を含む文を除き，用言系の文節では「接続」「係り先情報」「句読点」体言系の文節では「格助詞等」「提題助詞」「句読点」のいづれかを含む場合の文節を出力．
3. 2の条件に加え，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．
4. 3の条件に加え，用言系の文節のみを出力．

図2にそれぞれの例を示す．

†岐阜大学工学部

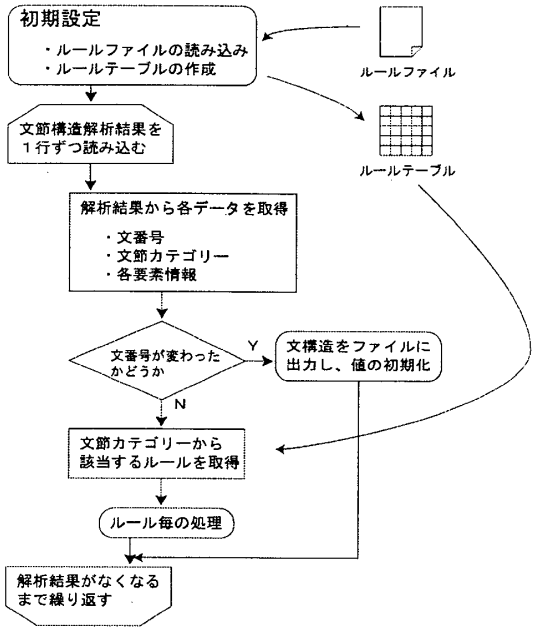

図１　文節列作成の流れ

寺内は、慢性の痛みを訴えていた左ひざの手術を受け、この試合が復帰戦だった。
N(Φ Φ Φ は 連用 、) N(Φ Φ の Φ 連体 Φ)
N(Φ を Φ Φ 連用 Φ) P1(Φ ている/たΦ Φ 連体 Φ)
N(Φ Φ の Φ 連体 Φ) N(Φ を Φ Φ 連用 Φ)
P1(Φ Φ ΦΦ 連用 、) T(Φ Φ Φ Φ 連体 Φ)
N(Φ が Φ Φ 連用 Φ) N(Φ Φ Φ Φ ダ系 Φ)
P2(Φ だった Φ Φ 文末 。)
N(Φ は 、) N(を Φ Φ) P1(Φ 連体 Φ) N(を Φ Φ)
P1(Φ 連用 、) N(が Φ Φ) P2(Φ 文末 。)
N(Φ は 、) P1(Φ 連用 、) N(が Φ Φ) P2(Φ 文末 。)
P1(Φ 連用 、) P2(Φ 文末 。)

図２　パターンの例

6　文パターンの出現頻度

毎日新聞記事（約67万文）に対し，文節構造解析し，5節の抽出条件を適用して文パターンの統計を行った．統計をとる際，以下の項目に注意した．結果は表5に示す．

有効文数　：　抽出条件にあてはまった文数
パターン数　：　文節列のパターン数
収束率　：　どれぐれらい収束したかを表す指数
1-(パターン数)/(有効文数)*100　で表現
頻度１　：　１度しか出現しなかったパターン数
頻度１の割合：パターン数に対しての頻度１のパターンの割合
カバー率　：　出現頻度上位何%のパターンでどれだけをカバーしているか

表５　文パターン統計

有効文数	671032	659273	659183	657519
パターン数	537270	410867	267433	42211
収束率	19.93%	37.68%	59.43%	93.58%
頻度１	508538	375990	232699	31639
頻度１の割合	94.65%	91.51%	87.01%	74.95%
カバー率 50%	201754	81231	3359	4
60%	268858	147158	19247	8
70%	335961	213086	69679	24
80%	403064	279013	135597	119
90%	470167	344940	201515	1423

ほとんどパターンは収束しなかったが，　の条件からは50%を超える収束率だった．特に　の条件では90%を超える収束率だが，用言系のみ（文で言えば述語のみ）出力しているので，この結果から，文節構造解析前の文を想像するのは難しい．このことより文構造を考える上で，一番　が適していると思われる．また，頻度１の割合が，全ての抽出条件において高い割合を示している．

7　表現の標準化

体言系，用言系の各要素に着目し，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」に置換した．この標準化を行った上で，前節と同じ分布調査を試みた．

表６　機能語部統計（標準化）

	名詞	動詞
文節数	3342512	1177915
機能語部パターン数	292	5973
90%到達位	11	56
95%到達位	18	170
99%到達位	57	1097
99.99%到達位	153	4796
残りパターン数	139	1177

名詞の機能語のパターンはあまり収束しなかったが，動詞の機能語のパターン数が約1700減少し，原文のパターン数に比べて約22%も収束した．

表７　文パターン統計（標準化）

有効文数	671032	659273	659183	657519
パターン数	534494	410087	266157	40739
収束率	20.35%	37.80%	59.62%	93.80%
頻度１	505526	375076	231337	30397
頻度１の割合	94.58%	91.46%	86.92%	74.61%
カバー率 50%	198978	80442	3299	4
60%	266082	146369	18652	8
70%	333185	212297	68403	24
80%	400288	278224	134321	114
90%	467391	344150	200239	1275

パターン数と頻度１は，表５と比べては減少しているが，収束率は増加していない．カバー率も同様である．標準化のための置き換えの処理は，まだ十分でなかったかもしれない．更に検討してみるつもりである．

8　おわりに

新聞記事一年分に対して，文節構造と文パターンの出現頻度の分布を調査した．更に同じ意味の表現は同一化するという，標準化を行った上での調査も行った．格助詞等では33パターンで，使役等，時制等，判断等，接続ではそれぞれ26,33,375,123パターンで99%に至ることが分かった．文パターンについては標準化しても，同様の程度のカバー率を得るには至らなかった．標準化については更に検討を進めたい．

参考文献

[1] 一ノ瀬　友紀夫　「文節の内部構造解析と出現頻度統計」，言語処理学会 第８回年次大会,pp.108-111(2002)