自 然 言 語 処 理 141－24
（2001．1、26）

文節区切りのための品詞列統計情報の調査

小島 丈幸, 乾 伸雄, 小谷 善行
(東京農工大学工学部)

形態素解析済の文を文節に区切る処理は，係り受け解析などの処理のために必要となる．各形態素間に着目した精度のよい文節区切り手法が存在するが，それらは学習データとして文節区切り付きのコーパスが必要である．しかし，形態素区切り付きのコーパスの方が大規模であるため，これを使った手法の開発が期待される．本稿では，形態素解析済コーパスを学習データとした精度のよい文節区切りを目標にした，品詞列コーパスの統計的な調査について報告する．そのために，品詞列がどれくらい文節になりやすいかを表す「文節形成率」を定義した．

その結果，文節形成率が小さくても文節としてふさわしい品詞列があること，文節形成率が高い品詞列においても左右エントロピーにはばらつきがあり，左右エントロピーが高いという基準だけでは適切な文節を抽出できないことがわかった．

Statistical Investigation of Part-Of-Speech Sequences for Bunsetsu Segmentation

Takeyuki Kojima, Nobuo Inui, Yoshiyuki Kotani
(Tokyo University of Agric. and Tech.)

To segment a sequence of words into a sequence of phrases, called Bunsetsu, plays an important role in syntactic analyses. Several researchers used bracketed corpora for extracting Bunsetsu and reported that systems improved performance by learning rules or statistics. However, we expect to use tagged corpora to do it, since bracketd corpora are too small to cope with varieties of sentences. This paper reports some statistical investigations about part-of-speech sequences for Bunsetsu segmentation. We defined "Bunsetsu rate" for each part-of-speech sequence, which indicates the possibility of becoming Bunsetsu.

Experimental results showed that there are some part-of-speech sequences with high plausibility and low Bunsetsu rate. And plausible Bunsetsu candidates cannot be extracted only by using entoropies of left/right sides of them.

1 はじめに

形態素解析済の文を文節に区切る処理は，係り受け解析などの処理のために重要である．本稿では，形態素解析済の文が品詞列コーパスとして与えられているときに，そこから文節区切りに必要な情報を抽出することを目的とする．文あるいは文節を品詞の列と見なし，文節を構成しやすい品詞列が統計的にどういう特徴を持っているのかについて調査した結果を報告する．

文節区切りで成果を上げている手法は，各要素間に着目してそこが文節の切れ目になっているか否かを判定するものである [2, 3]．ここでいう要素とは，形態素や品詞などの文あるいは文節を構成する単位のことである．これらの手法では，文節区切りの付加した学習データが必要である．一方，要素間に着目するのではなく，区切った結果として生じる文節自体の特徴に着目した手法も考えられる．この場合，学習データには必ずしも文節区切りの情報が必要ではない．本稿で

は，この考えに基づき，品詞列の文節になりやすさを表す「文節形成率」を考える．

品詞列コーパスから CFG 規則を獲得する研究 [1] も報告されている．扱う文法体系は異なっているが，品詞列コーパスに内在する情報から構文的なまとまりを発見するという点では類似している．この研究 [1] では，品詞列の左右エントロピーが共に高いものを構文的なまとまりと認定しているが，本稿では同じ基準が文節候補の選択に用いることができるかを検証する．さらに，他に自然言語の特徴を表す有益な情報が得られるかどうか調べる．

文節の候補を見つけるために文節形成率の高い品詞列を得たいが，文節区切りのないデータから直接的に文節形成率を求めることはできない．そこで，逆に文節形成率の高い品詞列の出現環境を調査し，その特徴を捉えることができれば，品詞列の出現状況から文節形成率を推定できることになる．

2 文節形成率の定義

2.1 品詞列の文節形成率

ここで，すべての品詞列に対して文節形成率を定義する．これは，その品詞列がコーパス中で文節を形成する確率を示すものである．つまり，品詞列 $\boldsymbol{w}$ の文節形成率 $\rho(\boldsymbol{w})$ は，コーパス中に品詞列 $\boldsymbol{w}$ が出現したときにそれが文節となる条件付き確率 $p(\circ|\boldsymbol{w})$ である．

$$\rho(\boldsymbol{w}) \equiv p(\circ|\boldsymbol{w}),$$

ただし，$\circ$ はその品詞列が文節を形成していることを示すものとする．

この値の推定値は，コーパス中における品詞列 $\boldsymbol{w}$ の頻度を $f(\boldsymbol{w})$，文節として出現した $\boldsymbol{w}$ の頻度を $f(\circ,\boldsymbol{w})$ として，

$$\rho(\boldsymbol{w}) \approx \frac{f(\circ,\boldsymbol{w})}{f(\boldsymbol{w})},$$

となる．

当然，文節になりやすいほど文節形成率が高く，最大は必ず文節になる品詞列の場合の 1，最小は絶対文節にならない品詞列の場合の 0 である．

文節形成率は，その品詞列がどれくらい文節になりやすかを示しているが，文節としてのふさわしさをそのまま表したものではない．文節としてふさわしい品詞列でも，たまたま文節をまたいで出現することや他の文節の一部として出現することがあるからである．

2.2 文節形成率に対する累積相対度数

文節形成率の分布を見るために，文節形成率に対する累積相対度数 γ を定義する．この値は品詞列の長さごとに定義されるものとし，文節形成率がある値以下の品詞列の割合を表すものとする．すなわち，次の定義式で表現できる．

$$\gamma_n(r) \equiv \frac{\sum_{|\boldsymbol{w}|=n,\rho(\boldsymbol{w})\leq r} 1}{\sum_{|\boldsymbol{w}|=n} 1}.$$

この式において，分母はコーパスに出現した長さ n の品詞列の異なり数，分子はそのうち文節形成率が r 以下のものの数である．

同様にして，重み付きの累積相対度数 δ を定義する．

$$\delta_n(r) \equiv \frac{\sum_{|\boldsymbol{w}|=n,\rho(\boldsymbol{w})\leq r} f(\boldsymbol{w})}{\sum_{|\boldsymbol{w}|=n} f(\boldsymbol{w})}.$$

この式において，分母はコーパスに出現した長さ n の品詞列ののべ数である．すなわち，ほぼコーパスのサイズに一致する値である．一方，分子は文節形成率が r 以下の長さ n の品詞列の頻度である．

γ,δ は累積相対度数なので，ある文節形成率を持つ品詞列の相対度数はこれらの微分値として得られる．すなわち，文節形成率 r を横軸に，累積相対度数を縦軸にとってグラフにしたとき，文節形成率がその値をとるような品詞列が多いところで傾きが急になる．

2.3 文節形成率に対するカバー率

単純にあるしきい値以上の文節形成率を持つ品詞列を文節の候補と見なした場合に，実際のコーパスに出現する文節をどれくらいカバーしているかを次のカバー率 $q_n(r)$ で定量化する．

$$q_n(r) \equiv \frac{\sum_{|\boldsymbol{w}|\leq n,\rho(\boldsymbol{w})>r} f(\circ,\boldsymbol{w})}{\sum_{\rho(\boldsymbol{w})\leq r} f(\circ,\boldsymbol{w})}.$$

この式において，分母はコーパスに出現したのべ文節数で，分子は長さが n 以下，文節形成率が r より大きい文節ののべ出現数である．

カバー率 $q_n(r)$ は，文節形成率 r に対して単調に減少する．もし，あるしきい値以上の文節形成率を持つ

品詞列を文節と見なすならば，できるだけ少ない文節候補でできるだけコーパスをカバーしなければならないので，カバー率をあまり下げない，できるだけ大きな r を探すことになる．

なお，$q_n(0)$ は長さ n 以下のすべての文節によるカバー率を表しており，コーパスに出現する文節の最大長を N とすると，$n \geq N$ に対して $q_n(0) = 1$ となる．

2.4 品詞列の左右エントロピー

品詞列の周囲からの独立性を定量化する一つの指標に，次の式で定義される左右エントロピー $H_{\mathrm{L}}(\boldsymbol{w}), H_{\mathrm{R}}(\boldsymbol{w})$ がある[1]．

$$H_{\mathrm{L}}(\boldsymbol{w}) \equiv -\sum_{x} p(x \cdot \boldsymbol{w}|\boldsymbol{w}) \log p(x \cdot \boldsymbol{w}|\boldsymbol{w})$$

$$H_{\mathrm{R}}(\boldsymbol{w}) \equiv -\sum_{x} p(\boldsymbol{w} \cdot x|\boldsymbol{w}) \log p(\boldsymbol{w} \cdot x|\boldsymbol{w}).$$

ただし，ここで x は一つの品詞を表す．もし，文節をなす品詞列が文節をなさない品詞列よりも周囲からの独立性が高いと言えるならば，左右エントロピーが高い品詞列を文節の候補と考えてよいことになる．

3 調査と考察

3.1 コーパスの準備

本調査では，EDR 日本語コーパス [4] の構文情報を元に文節を決定した．EDR 日本語コーパスにおいては，係り受け関係は修飾合成で示されている．そこで，修飾合成で二つの句がまとめあげられていたらその句の境界には文節区切りがある，という簡単な規則で文節区切り付きの品詞列データを作成した．

この文節区切りデータを作る処理は非常に単純であるので，構文構造に交差が存在する文などはうまく文節区切りデータを作れない．これらの数はそれほど多くなく，また文節間の係り受けについては調べないので，これらをはずしても統計情報を大きく変えることはないと判断した．そこで，EDR 日本語コーパスの先頭 10000 文から簡単な規則では変換できない文を除いた 9021 文から文節区切り付き品詞列コーパスを作り，調査に用いた．9021 文ののべ品詞数は 221721，文節数は 81586 である．したがって，平均の文節長は 2.72 である．一文当たりの品詞数は 24.58，文節数は 9.04 である．

品詞は表 1 に示す，EDR 日本語コーパスの粗い品詞分類のものをそのまま用いた．

表 1: 品詞とアルファベットの対応

a:名詞	b:動詞	c:形容詞	d:形容動詞
e:副詞	f:連体詞	g:接続詞	h:接頭語
i:接尾語	j:語尾	k:助詞	l:助動詞
m:感動詞	n:記号	o:数字	x:文頭文末

3.2 文節形成率に対する累積相対度数

文節形成率に対する累積相対度数 γ, δ をそれぞれ図 1,2 に示す．グラフでは，長さ 1 から長さ 5 まで，品詞列の長さごとにプロットしている．

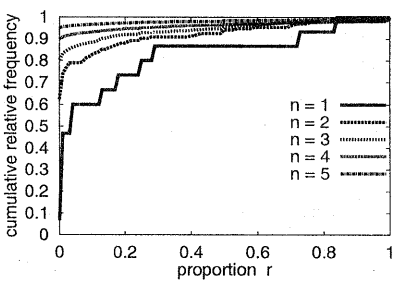

図 1: 累積相対度数 $\gamma_n(r)$

図 1 において縦軸である γ 軸の切片に注目すると，品詞列の長さが長くなるにつれてその値が大きくなっている．この値は，γ が 0 の品詞列の割合，すなわち，絶対に文節を形成しない品詞列の割合を示している．つまり，コーパス中に出現した品詞列のうち一回も文節を形成しなかった品詞列の割合は，品詞列が長いほど増える傾向にある．たとえば，品詞列の長さが 5 のときには，コーパス中に出現した品詞列の 95 % は一回も文節を形成しなかったことを示している．頻度で重み付けした δ でも同様である．

逆に r が 0.9 のときの γ の値は，品詞列の長さにあ

[1] 対数は本来底が 2 のものであるが，本稿では大小関係だけが重要なので，便宜上自然対数を用いている．

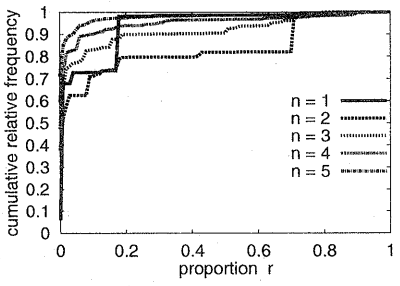

図 2: 重み付き累積相対度数 $\delta_n(r)$

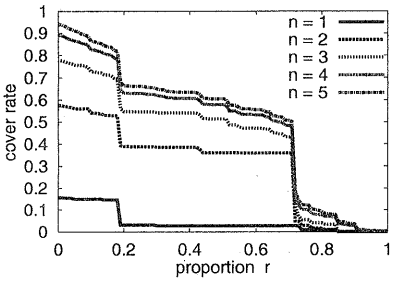

図 3: カバー率 $q_n(r)$

まり依存せず 0.98 くらいであるので，r が 0.9 を超えるような品詞列の割合は 2 % 程度と言える．つまり，文節を形成する可能性が 90 % であるような品詞列は全体の 2 % 程度しか存在しないことを表している．

δ に関しては，全体的に γ よりも小さい値で移行している．これは，文節形成率が低いところでは，個々の品詞列の頻度が小さいことを示している．つまり，文節を形成せずにたまたま並んでしまったような品詞列はなかなか出にくいので頻度が小さいが，文節を形成することの多い品詞列は，文節として何度も出現するので頻度が高くなっていると考えられる．

3.3 文節形成率に対するカバー率

あるしきい値 r 以上の文節形成率を持つ品詞列を文節と見なたしときに，コーパスをどれくらいカバーできるかの目安となるカバー率 $q_n(r)$ を図 3 に示す．調べる品詞列の最大長 n が 1 から 5 までのものを示した．

図 3 を見ると，文節形成率が 0.2 と 0.7 辺りでカバー率が大きく変化していることがわかる．それらがそれぞれ長さが 1 の文節と長さが 2 の文節に対応していることもわかる．具体的には,「名詞」と「名詞 助詞」である．名詞は文節形成率が 0.17 と小さいが，頻度が非常に多いのでカバー率を考えたときには無視できなくなる．

カバー率が大きく変化する前の 0.7 をしきい値に設定した場合，$n = 5$ であってもカバー率は 0.5 である．すなわち，ある文節形成率以上の品詞列を文節候補と見なす手法では，全体の半分の文節しかカバーできないことになる．

図 3 を見ると $q_5(0)$ は 0.94 になっている．これはコーパスに長さが 6 以上文節が 6 % ほど出現したことを表している．今回コーパスを作成する手法は非常に単純で，動詞を中心とした文節で長いものが現れやすくなっていたので，その影響で長さが 6 以上の文節の数も多くなった．

3.4 文節形成率に対する頻度

文節形成率と頻度の関係を調べるために，コーパスに出現した品詞列それぞれを文節形成率-頻度空間にプロットしたものを図 4 から 図 7 に示す．品詞列の長さが 1 から 5 のものに対応している．

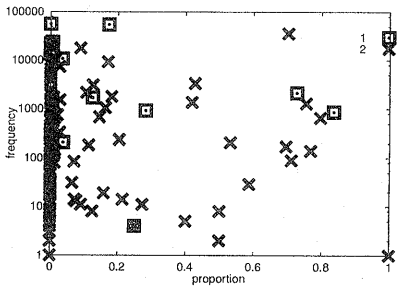

図 4: 文節形成率に対する頻度 (長さ 1,2)

どのグラフを見ても，大多数の品詞列が文節形成率の低いところに集まっていることがわかる．これは図 1 からも導かれる．頻度が低いところでは，文節形成率に

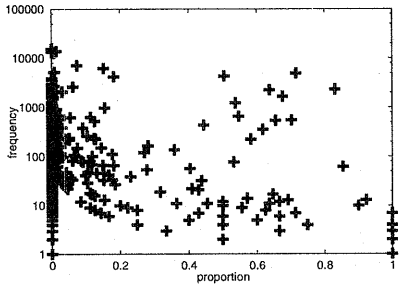

図 5: 文節形成率に対する頻度 (長さ 3)

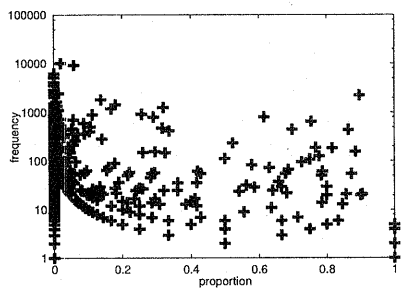

図 6: 文節形成率に対する頻度 (長さ 4)

あまり関係なくばらついている.

頻度が 100 を超えるような品詞列では，文節形成率が 0.5 付近のものは少ない．大多数は文節形成率が 0 近辺にあり，残りのほとんどの文節形成率は 0.6 を超えている．しかも，この傾向は品詞列が長くなるほど強くなっている．これは，頻度が高い，すなわち信頼性が高い品詞列では，文節形成率が 0.2 より小さいものか文節形成率が 0.6 より大きいものとに大別され，中間の文節形成率を持つものが少ないことを表している．

また，これは当然のことではあるが，品詞列の出現頻度は，その長さが大きくなるにつれ全体的に減少する傾向にある．

3.5 高い文節形成率を持つ品詞列

3.4 節では，文節形成率と頻度の関係を全体的に述べた．ここでは，高い文節形成率を持つ品詞列についてその長さごとに述べる．

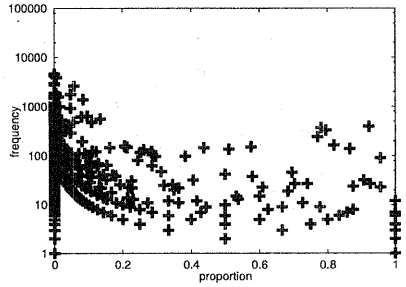

図 7: 文節形成率に対する頻度 (長さ 5)

文節形成率が 0.6 以上，頻度が 100 以上の品詞列を抽出したところ，長さが 5 以下では 30 種類あった．これら 30 品詞列を，左エントロピー-右エントロピー空間にプロットしたものを図 8 に示す．なお，スペースの都合上，図中では各品詞をアルファベット一文字で表している．品詞とアルファベットの対応は表 1 に基づく．

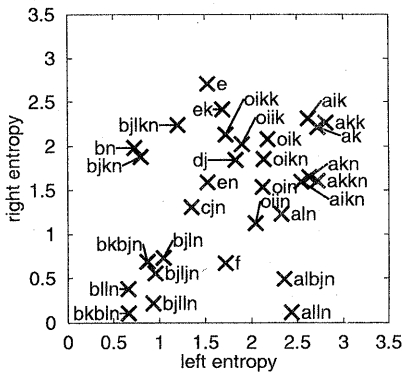

図 8: 文節らしい品詞列の左右エントロピー

図 8 を見ると，文節形成率の高い品詞列，すなわち文節になりやすい品詞列においても文節ごとに左右のエントロピーの値にばらつきがある．また，一つの文節の左右のエントロピーが大きく異なっているものもある．単純に左右エントロピーが共に高いものを文節候補としてしまうとうまくいかないことが予想される．

先頭の品詞と左側エントロピーの間には明確な関係がある．すなわち，左側エントロピーが低いほうから

順に，動詞 (b), 副詞 (e), 数字 (o), 名詞 (a) が先頭の品詞列が集まっている．文節をなす名詞の前には，その文節と同じ用言に係る連用句やその名詞に係ってくる連体句など，様々な品詞が来やすいので名詞の左側エントロピーが高くなっていると考えられる．

一方，末尾の品詞は記号 (n) か助詞 (k) がほとんどで，助詞が末尾にくる品詞列は左右エントロピーが共に高いところに固まっている．

3.5.1 体言が中心となっている文節

図 8 で左右エントロピーが共に高い三つ (aik, ak, akk) は，名詞の後に格助詞や連用助詞が接続してできた連用句である．助詞が二つある品詞列 (akk) は,「には」など格助詞と係助詞が接続したもの,「だけが」など副助詞と格助詞または係助詞または連体助詞が接続したものがある．

これら三つの品詞列の後に記号 (n) が接続したものは，やや右側エントロピーが低下する．図 8 では先の三つの品詞列のやや下に位置している．このときの末尾の記号 (n) は読点である．

また，先の三つの品詞列の名詞 (a) の部分を数字と接尾語 (oi) に置き換えた一群も，左右エントロピーがやや低いところに現れている．

図 8 の右下にある二つの品詞列 (akbjn, alln) は「～である.」「～だった.」「～でした.」などであり，最後の記号は句点である．文末にくる文節は後ろに文頭か文末しかこないので，右側エントロピーは低くなっている．

3.5.2 動詞が中心となっている文節

図 8 で左側に位置している二つの品詞列 (bjkn, bn) は，それぞれ「すれば,」「しつつ,」のように接続助詞の後に読点がきたもの,「提携,」「決め,」など，前者の接続助詞が省略されたものになっている．これらのやや右に位置する品詞列 (bjlkn) は,「～であったが,」などの接続助詞と読点で終わる文節と「～ますか.」などの終助詞と句点で終わる文節が混ざっている．出現する環境が異なる文節が混ざっているので，その分左右エントロピーも増加していると考えられる．

図 8 の左下にある品詞列のうち，bkbjn は「～ている.」「～てしまう.」などの様態を表す文末表現で，bjljn は「～される.」「～られた.」などの文末表現である．これらの語尾が助動詞に変わったものがさらに左下に位置していて，bkbln は「～ていた.」「～てきた.」など，bjlln は「～された.」「～られた.」などになっている．

bjln は「～した.」「あります.」のように句点で終わるものと「開かれ,」「損なわず,」のように読点で終わるものが混ざっていた．blln も「開かれた.」「やめました.」のように句点で終わるものと「集められず,」「訪れたそうで,」のように読点で終わるものが混ざっていた．

4 おわりに

本稿では，品詞列コーパスから文節区切りのための情報を抽出する調査を行った．品詞列の文節になりやすさを示す文節形成率を定義した．

その結果，品詞列は文節形成率が 0.2 より小さいか 0.7 より大きいものがほとんどであることがわかった．また，単純に文節形成率が大きいものだけを文節候補とすることはできないこと，左右エントロピーが大きいものだけを文節候補とすることはできないことがわかった．

今後は，文節形成率を用いて文節区切りを行う手法や文節区切りなしコーパスから文節形成率を推定する手法を考察する必要がある．

参考文献

[1] 森信介, 長尾眞. タグ付きコーパスからの統語規則の獲得. 情報処理学会論文誌, Vol. 37, No. 9, pp. 1688–1696, 1996.

[2] 村田真樹, 内元清貴, 馬青, 井佐原均. 排反な規則を用いた文節まとめあげ. 情報処理学会論文誌, Vol. 37, No. 6, pp. 1234–1244, 1996.

[3] 張玉潔, 尾関和彦. 分類木を用いた日本語の自動文節分割. 情報処理学会自然言語処理研究会, Vol. 121, No. 1, pp. 1–8, 1997.

[4] 日本電子化辞書研究所. EDR 電子化辞書 1.5 版 使用説明書, 1996.