

## 日本語技術文における係り受け関係の複雑な文の抽出

小池, 康弘  
九州大学システム情報科学研究院知能システム学部門

菅沼, 明  
九州大学システム情報科学研究院知能システム学部門

谷口, 倫一郎  
九州大学システム情報科学研究院知能システム学部門

<https://hdl.handle.net/2324/5714>

---

出版情報：電子情報通信学会ソサイエティ大会講演論文集。1998年。情報・システム，pp.27-27，1998-09-07。電子情報通信学会

バージョン：

権利関係：

# D-5-1 日本語技術文における係り受け関係の複雑な文の抽出

## Extraction Methods for Sentences Including Complex Dependency

in Japanese Technical Documents

小池 康弘      菅沼 明      谷口 倫一郎

Yasuhiro Koike   Akira Suganuma   Rin'ichiro Taniguchi

九州大学 大学院システム情報科学研究科

Graduate School of Information Science and Electrical Engineering, Kyushu University

### 1 はじめに

既存の校正・推敲支援システムは、表層的な情報を提示するものが多いようである。そこで、表層的な情報のみでなく構文情報などを考慮した情報を提示することを考えた。

文献[1]で、分かりやすい文章についてアンケートを取った結果が紹介されていた。それによると、「係り受けが分かりやすい文」という項目が上位に位置している。このことから、分かりにくい文の原因として「係り受け関係が複雑である」を挙げることができる。本研究では、係り受け関係の複雑な文を抽出することにより、分かりにくい文という情報の提示を目標とした。

文の係り受け関係が複雑になってしまう原因として、(1)前置修飾節が非常に長い、(2)一つの語に多くの修飾語に係る、(3)修飾語の順序が悪い、の3つの構造が考えられる。これらの構造を含む文を、抽出する対象とした。

### 2 係り受け関係の複雑な文の抽出

係り受け関係の複雑な文を抽出するにあたり、文の係り受け構造を視覚的に表現する「係り受けブロック図」と、係り受け関係の複雑さを数値化した「ペナルティ」を導入した。

#### 2.1 係り受けブロック図

係り受けブロック図とは、一種の構文解析木であり、文の文節単位での係り受け関係を表す図である。係り受けブロック図の例を図1に示す。図中の1つ1つの箱を「ブロック」と呼ぶ。ブロックが横に並んでいる時、左のブロックが右のブロックに係っていることを表す。

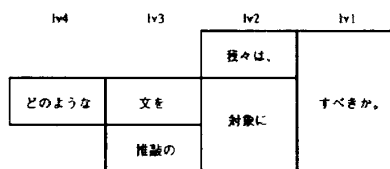


図1: 係り受けブロック図の例

係り受け関係が複雑で読みにくい文は、係り受けブロック図が極端に横や縦に長くなったり、入り組んだブロック図になったりする。そのため、係り受けブロック図によって、視覚的に文の係り受け関係の複雑さを表現することができる。

#### 2.2 ペナルティ

ペナルティとは、係り受けブロック図の複雑さを数値化したものである。このペナルティの計算ルールは、先に挙げた3つの構造を含む文において、高い値を算出するように設定した。つまり、ペナルティの値が大きい文は、文の係り受け構造が複雑な文であるとみなすことができる。

ブロックAにn個のブロック $B_1, \dots, B_n$ に係っている場合、ブロックAのペナルティを次の式で定義する。なお $P_A, P_{B_k}$ はブロックA,  $B_k$ のペナルティ、 $L_A$ はブロックAのレベル、 $f_1(x), f_2(x), f_3(x)$ はそれぞれの抽出対象に対する重みづけ関数(増加関数)である。

$$P_A = f_2(n) \sum_{k=1}^n (P_{B_k} * f_3(k)) + f_1(L_A)$$

この式により、ブロック図の左端のブロックから順に、ブロック毎のペナルティを求める。そして右端のブロックのペナルティの和を、文のペナルティとする。

#### 2.3 抽出方法

入力文のペナルティと設定された閾値を比較し、ペナルティが閾値を超える文は、係り受け関係が複雑な文の候補として抽出・提示する。

#### 2.4 実験

人間の感覚を調査するために、50文を任意に選び出し、大学生を対象にアンケートを行なった。1文当たり約20人に評価してもらった。このアンケートで、「一度、文頭から文末まで目を通しただけで、文の修飾・被修飾関係の構造が理解できた」と評価した人の割合を、「文の読みやすさ」の指標として扱った。この指標が、60%未満である文を、係り受け関係が複雑な文として抽出することを目標とした。

次に、人手によって文の係り受けを解析し、ペナルティを計算した。ペナルティの3つの重み付け関数は、

$$f_1(x) = x, f_2(x) = x, f_3(x) = (x - 1) * 0.4 + 1$$

とした。その結果得られたペナルティと読みやすさの指標とを比較した。文の読みやすさとペナルティの関係を図2に示す。

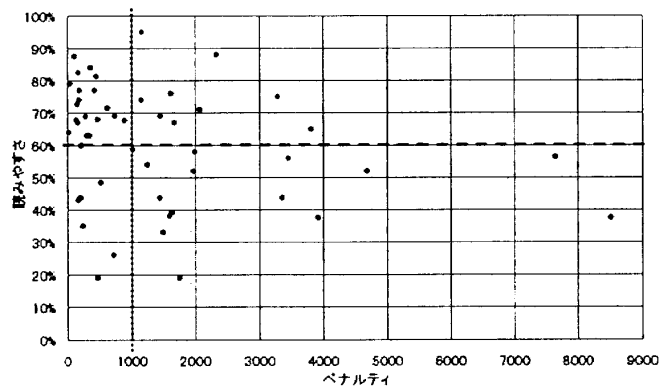


図2: 文の読みやすさとペナルティの関係

このグラフから、閾値を1000点とした。この時、再現率は71%、適合率は63%となる。

### 3 おわりに

本稿では、「係り受けブロック図」と「ペナルティ」によって、係り受け関係の複雑な文を抽出する方法について述べた。この方法によって実際の文を評価したところ、分かりにくい文をある程度指摘できることがわかった。今後は、計算機による実装を行なっていく予定である。

#### 参考文献

- [1] 尾川正二:原稿の書き方, 講談社現代新書 (1976)
- [2] 本田勝一:日本語の作文技術, 朝日新聞社 (1982)