

手がかり語を用いた日本語文章の 段落分けに関する実証的考察

山本和英 * 増山繁 * 内藤昭三 **

* 豊橋技術科学大学知識情報工学系 ** NTT基礎研究所

本研究では、談話解析に関する一つのトピックである「段落分け」、つまり与えられた日本語文章を、計算機を用いていかに自然に段落分けを行なうかという問題を取り上げ、文章の表層的な情報のみに基づいて段落分けを行なうことを試みた。

まず、実際の文章の各文頭にどのような語が出現しているのかの調査を、科学雑誌と新聞コラムの2種類の文章に対して行なった。次に、その結果から「手がかり語」の定義を行ない、それらの語と文長の要素のみを用いて、計算機によって段落分けを試みた。そして、その出力結果がどの程度自然な文章であるかについて、原文との一致度とアンケートの結果という、二つの評価基準を用いて考察を行なった。

An Experimental Study on Paragraphing Japanese Sentences Using Cue Words

Kazuhide YAMAMOTO * Sigeru MASUYAMA * Shozo NAITO **

* Dept. of Knowledge-based Info. Eng., Toyohashi Univ. of Tech.

** NTT Basic Research Laboratories

In this paper we focus on paragraphing, which is one of the topics in discourse structure analysis. We attempt to paragraph Japanese sentences automatically using only surface information.

First, we investigate what words come in the top of the sentences against articles and essays. Then, based on this investigation, we define cue words and try to paragraph Japanese sentences using these words and length of sentences. Finally, we evaluate naturalness of the results by the following two criteria: the ratio of coincidence of the resulting paragraphing with that of the originals and the consequence of a questionnaire.

1 はじめに

本稿では、日本語文章における段落分け、すなわち、与えられた複数の文を、計算機を用いて自然に段落に分けることを試みる。現在、談話解析の分野においては、談話構造のモデル化とそれに基づく文章の要約処理や、照応・省略解析に関する研究が行なわれているが[1]、日本語文章における計算機を用いた段落分けの研究は、従来知られていない。

自然な段落分けを行なうためには、個々の文を解析して理解することのみでは十分でなく、文と文の相対的な結合の強さなどの文間の関係も解明しなければならぬ。本研究は、特に後者の観点に基づいて実施した。

本研究では、最初の試みとして、文面から読みとれる表層的な要素だけを用いて、段落分けを行なった。具体的には、構文解析や意味解析を行わず、それぞれの文の文頭にある単語と文長の要素だけに基いて与えられた文章を段落分けした。この研究結果によって、文間の結束度などの情報がどの程度表層的に出現しているのかを知ることができる。

今回の研究が利用可能な応用分野として、文章作成の支援や推敲の支援を行なうことを想定している。これらの研究を進めていくことによって、我々が文章を作成する場合に、たとえば、適当な段落分けの位置を計算機が指示することにより文章全体の構造を明確にし、読者にとって読みやすい文章を書くための支援を行なうことができよう。

本稿ではまず、段落についての予備的な考察を2章で行なった後、文頭での単語出現調査について、3章で報告する。そして、4章ではこれらの結果に基づいた、計算機での段落分けの試行について、5章では、その出力結果の文章がどの程度自然であるのかを、原文との一致度とアンケートの結果の2点から述べる。最後に、6章で今後行なうべき課題などについて述べる。

2 「段落」に関する予備的考察

段落とは「小主題を中心とした一まとまりの表現とその区切り」である[11]。段落について分析、あるいは考察することは談話解析を行なう上で重要である。なぜならば、段落内におけ

る文の結合の関係や段落間の構造を把握できなければ、文章全体を把握することは困難であるからである。

実際の文章における段落の区切り方は、内容上のまとまりによる論理的・意味的なもののほか、作者の息の長さ、持続力等による生理的なものや、文章のリズムに対する作者の好みや特殊な創作意図によるもの等がある[11]。このように様々な要素が段落分けには含まれている。

日本語の文章については、文献[6]で「段落の感覚というものにはなきにひとしい」とあるように、我々は文章を書く際にあまり段落のことを気にしない傾向がある。しかし、適切な段落分けを行なうことによって、その文章が読みやすく、要点をつかみやすくなるので、文章を段落に分ける作業は、重要な作業なのである[3, 7]。

一般に、書き手の設定した段落と読み手の設定する段落は一致するとは限らない[4, 5]。さらに、段落の分け方の基準は個人によって異なる。例えば、段落に分かれていない文章を、複数の人に段落に分けてもらった場合、それらがすべて一致することはありえない[2]。しかしながら、ほとんどの人が「段落に分けない」位置は存在するはずである。このことから、段落に分けるという作業には、一文内における構文規則（いわゆる文法）ほど強力ではないが、何らかの緩やかな規則が存在すると考えることができる。

3 文頭の単語出現調査

3.1 調査方法

実際の文章で、段落の初めや段落内の文頭に、どのような単語が使用されているのかを知るために、文頭の語の統計調査を行なった。今回調査の対象とした文章は、科学雑誌「日経サイエンス」と朝日新聞の「天声人語」欄の2種類である。統計は別々に集計した。

調査は、すべての文頭について、どの文章にも出現する可能性のある単語を抽出するという方法で行なった。これらの中には、文間に出て文と文を接続する役割をする接続詞、および副詞の他にも、我々が疑似的に、あるいは無意識にこれらの役割を果たさせる単語があると仮定して、広くこれらの単語も含めて調査を行

なっている。

表 1: 調査対象単語の品詞分類

品詞	単語例
接続詞	そして、しかし、一方、そこで
連体詞	この、その、そういう、ある
副詞	たとえば、もし、つまり、なぜ
代名詞	私、彼、われわれ、それ、ここ
名詞	最初、筆者、現在、このよう
(その他)	したがって、とすると、本稿

表 1 に、調査対象単語を [10] に従って品詞に分類したものを示す。ただし、この表で「(その他)」とは、あらかじめ行なった予備調査の結果頻度が高いと予想され、調査対象に含めた語句を指す (例えば、「したがって」「とすると」など)。その他の専門用語などの語は、調査を行なっても統計的に大きな割合で出現する単語は存在しないという仮説に基づき、調査対象には含めていない。

3.2 計算機による調査アルゴリズム

今回はこれらの調査を計算機によって行なった。その際に使用した単語の検出法は、最長一致法である。この方法でマッチングを行なうと、一部の語については誤った認識を行なう可能性がある。しかし、以下の理由によりこれらの誤認は考慮しないとする。

1. マッチングを行なおうとしている語は 3 文字、あるいは 4 文字の語が多く、誤認が入りやすい 1 文字の語はほとんどないため、全体的には誤認は無視できるほどであると思われる。
2. 1 文字の語 (例えば「が」) については、そのままでは誤認を多く含む可能性があるが、マッチングを行なう文字数を増やすことによって (例えば「が、」と 2 文字にすることによって) 極力誤認の混入を防いでいる。

3.3 調査結果

実際に調査した文章は、「日経サイエンス」の記事 13 編、「天声人語」505 日分である。

調査の結果を付録の表 A、B に示す。ただし、付録に掲載した表は調査結果のうち、主なもの (サイエンスは頻度 11 以上、天声人語は頻度 12 以上) である。また、調査結果の統計的数値を以下の表 2 に示す。ただし、この表で抽出語の割合とは、調査を行なったすべての文頭のうち、どの程度の文頭に調査対象単語が出現しているのかを示す。

表 2: 調査を行なった文章

項目	日経サイエンス	天声人語
調査対象	記事 13 編	505 日分
総段落数	824	3,039
総文数	3,501	10,233
総文字数	212,403	416,006
段落当たりの平均文字数	257.8	136.9
その標準偏差	130.6	62.3
抽出語の割合	60.0%	30.2%

これらの集計結果を見てみると、まず、抽出した語の割合が、サイエンスと天声人語で大きく異なることがわかる。これは、サイエンスの方が、文章に論理性が要求されるために、文頭で使用する単語の多くは論理的に説明するための一定範囲の語彙に限定されるのに対し、天声人語では広範な話題を対象としているために、文頭に使用される語彙の範囲が広くなるのではなからうか。

この調査結果を品詞別にまとめたものを表 3 と表 4 に示す。ただし、この表にある「指示語」とは、いわゆる「こそあど言葉」であり、品詞とは別の概念であるが、別途に集計を行い、算出した。

品詞別の割合では、接続詞が全体の 4 分の 1 程度しかないことがわかる。このことは、文と文を接続する役割を果たしているのが接続詞だけではないことを示している [5]。一般に、日本語では接続詞は敬遠される傾向にあるという [8, 9]。これは、角のたたない表現を好むということに一因があるらしい [8]。そのため接続詞以外の語を文頭に使用して、接続の役割を代用している場合も多い。また接続詞については、段落内の文頭に比較的多く使用され、段落頭での使用頻度との差はサイエンスにおいて顕著である。

表 3: 抽出語の品詞別割合 (サイエンス)

	全文	段落頭	段落内文頭
接続詞	25%	17%	27%
連体詞	20%	15%	21%
副詞	20%	25%	19%
代名詞	15%	13%	16%
名詞	14%	25%	11%
指示語	38%	32%	40%

表 4: 抽出語の品詞別割合 (天声人語)

	全文	段落頭	段落内文頭
接続詞	23%	21%	24%
連体詞	21%	19%	21%
副詞	15%	17%	14%
代名詞	12%	8%	13%
名詞	8%	10%	7%
指示語	33%	28%	35%

調査結果によると、わずかではあるがサイエンスの方が接続詞を使用する頻度が高い。この理由として考えられるのは、一つには、サイエンスにはより論理的な文章の展開が必要とされることがあり、また5.1で述べるように、サイエンスでは、今回の分析対象に翻訳した文章も含めたために、それらの文章は日本語で論理を組み立てておらず、その結果、比較的接続詞を多用する英語の影響があるのではないと思われる。

また、品詞分類とは別に、指示語について統計をとった結果をみると、天声人語で3分の1、サイエンスでは抽出した語の4割近くを構成しているという結果を得た。この背景には、日本語には関係代名詞がなく、接続語句の使用があまり好まれないことがあると思われる[8]。接続関係を表す語があまり使用されない結果として、これらの指示語を用い、間接的に接続を表現しているのではないと思われる。

段落頭と段落内文頭との比較では、接続詞、連体詞、代名詞は段落内文頭の方が割合が高く、副詞、名詞については段落頭の方が割合が高くなっている。これは、サイエンス、天声人語という文章の種類に関係なく見られる傾向であり、非常に興味深い結果である。連体詞や代名詞に

ついては、前方の文との間に指示・照応関係が存在するために、文間の結合が比較的に強いと思われる。

以上の調査結果から、抽出された語の多くは段落頭よりも段落内の文頭に多く出現することがわかる。つまり、これらの語が文頭に出現した場合には、比較的前の文とつながりやすいことを示している。そこで今回抽出した、どの文章にも出現する可能性のある語のうち、非常に頻度の低い語(頻度1の語)を除いた語すべてを「手がかり語」と定義して、この手がかり語を用いることによって段落分けを行なうことを考えた。

4 計算機による段落分けの試行

4.1 段落分けの方針

以上の調査結果を踏まえて、計算機によって実際に段落分けを試みた。段落分けのアルゴリズムを決定するにあたり、今回は次のような方針で段落分けを行なった。

1. 段落分けは、On-line 的方法で行なう。すなわち、順番に入力文を読み込んでいき、その時点で段落分けを行なうかどうかを決定するという、逐次決定的な方法で行なう。これは、我々が文章を書く場合に、途中まで書いたところで段落に分けるかどうかの判断を行なっていると仮定したためである。
2. 今回の段落分けに使用する情報は、先ほど定義した「手がかり語」に関する情報と、それぞれの文長、および文章の平均文長だけとする。
3. まず、文頭に出現した語によって、前文との基本的な結合の強さを決定する。それを文長の要素によって修正する。
4. 文長の要素は、結合の強さに対して線形に影響すると仮定する。

これらの方針に基づいて、計算機による段落分けのアルゴリズムを実現した。

4.2 結合度の設定

まず、段落分けのための予備調査として行なった前章の結果から、それぞれの手がかり語に対して段落内での結合性の強さを表す数値（これを原結合度と定義する）を割り当てた。すなわち、原結合度とは、0から100までの範囲の数値であり、この数値が大きいほど前の文とつながる可能性が高いことを示す。この値の算出は、各手がかり語に対し、それらがどのくらいの割合で段落内文頭に出現したかによって行なう。すなわち、 $\text{原結合度} = A \div B \times 100$

（ただし、A：その手がかり語の段落内の文頭での出現数、B：その手がかり語の全出現数）と算出する。手がかり語として抽出していない語に対しては、前述の調査での、段落頭の出現個数と段落内文頭の出現個数から同様の計算を行ない、すべて一律の値として処理を行なう。原結合度は、2種類の文章それぞれについて算出する。

次に、段落分けを行なう際に、段落の長さも考慮に入れるため、ある文までのその段落長を、段落の初めから、その文の前の文までの文字数と定義した。つまり、段落頭の文の段落長は0、次の文の段落長は最初の文の長さ、というように定義した。

そして、段落を分けるかどうかを決定する結合度の計算を、次のように定義した。

$$\text{結合度} = \text{原結合度} + C_1 \times \text{平均段落長} - C_2 \times \text{現在の段落長} \quad (1)$$

ここで、 C_1 および C_2 は、どの程度段落の平均段落長および現在の段落長が影響するかというパラメータである。この式はまた、次のように表現することもできる。

$$\begin{aligned} \text{結合度} &= \text{原結合度} \\ &+ (C_1 - C_2) \times \text{平均段落長} \\ &+ C_2 \times (\text{平均段落長} - \text{現在の段落長}) \end{aligned} \quad (2)$$

つまり、結合度に影響する要素は、平均段落長の要素と、現在の段落長が平均段落長に比べてどの程度の長さなのかという要素の二つであ

る。この計算式は、我々が文章を書いていく過程で、現在の段落の長さが短い場合段落を続けようと試み、ある程度以上の長さになると、今度は段落に分けようとする意識が働くことと仮定して、この仮定を最も単純な数式で表現したものである。

そして、算出した結合度が50以下の時、その位置で段落を分ける。以後、この作業を繰り返してゆき、すべての文章を段落に分けていく。

4.3 計算機での出力

サイエンスと天声人語の文章のうちで、3章で手がかり語の調査を行っていない文章を対象にして、上で述べたアルゴリズムを用いて実際に段落分けを行なった。対象になった文章は、サイエンス6編、天声人語40日分である。

検証の方法は、まず、対象となる文章から段落分けを除去した文章を、計算機に入力する。そして計算機がこのアルゴリズムに基づいて段落分けを行ない、文章を出力するというものである。使用した計算機はSun4/60であり、プログラム言語はCommon Lisp (KCL)を用いた。

実験に使用するパラメータ C_1 や C_2 の値は、あらかじめ推定を行うことが非常に困難である。そこで、試行錯誤を繰り返して決定した。具体的には、まず様々なパラメータを設定して段落分けの出力を行ない、原文の総段落数と出力される段落数ができるだけ接近するように、二つのパラメータの設定を行なう。次にそれらの中から、原文との一致ができるだけ多くなるような数値に決定する方法をとった。

5 文章の自然さに関する考察

5.1 原文との一致による自然さの検証

出力された文章において、どの程度自然な段落分けがなされているかを示す一つの目安は、原文の段落とどの程度一致しているかを比較することである。そこで、その比較の結果を表5に示す。

表 5: 原文に対する一致の割合

C_1	C_2	総段落数	出力数	一致数
サイエンス				
0.02	0.1	285	312	115
0.02	0.15	285	411	138
0.03	0.15	285	384	126
天声人語				
0.01	0.1	202	222	77
0.1	0.2	202	210	75
0.3	0.5	202	228	69
1.5	2.0	202	223	71

結果の考察

- 出力された数値が低いことの一つの理由として、前述した「手がかり語」調査において、手がかり語がサイエンスにおいて約6割の文頭にしか出現していないので、全体の6割の情報量しか使用していないと考えられる。この情報を最大限に活用しても6割の一致しかみられないのではない。
- 長い文章であると、ある部分で段落の分け方が、原文と異なった場合、その影響が以後の文に受け継がれて、その結果、あまり一致の割合が上昇しないのではない。
- サイエンスの文章については、日本人の執筆した文章の他に、翻訳されて掲載されているものがある。今回検証を行なった文章の著者は、日本人5編、外国人1編であるが、「手がかり語」の調査を行なった文章の著者は、日本人7編、外国人6編と、約半分が翻訳された文章である。翻訳文の場合、外国語での段落に対する感覚がそのまま翻訳される可能性があり、これらの関係は明らかではない。

5.2 アンケートによる自然さの検証

5.2.1 「自然さ」の定義

どの程度自然な段落分けができたのかを判断するには、原文との一致をみるだけでは不十分である。なぜならば、原文と一致させること、あるいは原文の段落を推定することが本研究の目的ではなく、自然な段落分けを行なうことが

目的であるからである。

ある文章が自然かどうかの判断基準は非常にあいまいであり、今回のアルゴリズムによって作成した文章が自然かどうか、あるいはどの程度自然かを数値化することは容易ではない。そこで今回の研究では、人間がその文章を読んで、この文章が原文なのか、あるいは機械的に段落分けを行なった文章なのか、区別ができない程度の段落分けを行なったときに、その文章が自然であると考えられる。

5.2.2 主観性の除去

ある文章が自然であるということを、上のように定義した場合、人間に判断してもらわなければならない。そこで、今回のアルゴリズムで出力した文章が自然かどうかのアンケートを行なうのであるが、出力結果のみを提示して単に「この文章が原文かどうか区別できますか？」という質問事項であるとする、とても主観的なものになってしまう。なぜならば、そのように質問することによって、被験者は我々が望む回答を予想でき（すなわち、原文と区別できない）、その結果として我々が暗黙のうちに回答を誘導してしまうおそれがある。よってこのような質問では客観的な測定ができない。

そこで本研究ではこの点に十分配慮をして、主観性を取り除いて人間に判断してもらうために、次のような方法でアンケートを行なった。

被験者に対して、文章の原文を1編以上と、コンピュータによって段落分けした文章を1編以上を含む、合計5編の記事を提示する。5編の記事の内訳は被験者には示さない。そしてこれらの記事のそれぞれについて、原文かどうかを当ててもらう。このアンケートをサイエンスと天声人語について行なう。ただし、天声人語は提示する合計編数を7編とする。

この調査であれば、暗黙のうちに筆者らの期待する回答（すなわち、コンピュータの設定したアルゴリズムが原文と比較して区別できないということ）に誘導する恐れは少ない。そのため、どの程度自然な文章かを測る尺度として、客観性のある数値になると予想される。

5.2.3 アンケート調査の結果

以上のような方法によって、実際にアンケートを被験者11人（サイエンスは10人）に対して行なった。その結果を表6に示す。

表 6: アンケート調査の結果

	提示のべ文章	原文認識率
サイエンス		
原文	27編	67%
出力文章	23編	26%
天声人語		
原文	34編	76%
出力文章	43編	44%

この表にあるように、計算機による出力結果は、天声人語で4割以上が原文と認識された。この結果を、4分の3の原文しか原文として認識できないことと併せて考えると、比較的高い数値であろうと思われる。

サイエンスについては、天声人語よりも低い認識率となった。これは、一編の段落数が比較的多いために、段落分けが不自然な部分が多くなり、その結果として天声人語よりも低い認識率となったと予想される。

また、原文に対する認識率が7割前後であることも注目される。これは、個人によって段落分けの基準が同一ではない、つまり段落分けには個人差が存在することを実証している。

6 おわりに

本研究では、表層的な要素だけで、どの程度自然な段落分けができるかを、可能な限り実証的に考察した。その結果、文頭に出現する語の傾向や、文頭の語と文長の要素だけを用いただけの段落分けが、天声人語で4割以上の自然さがあることなどが判明した。

この「自然さ」のアンケートについては、主観性を取り除くことができた反面、質問項目が「原文かどうか」というものであったので、「おかしい段落の分け方ではないのだが、原文ではなさそうだ」という場合に、この質問では正しく評価されない。そこで、この質問方法にも改良の余地があるので、今後検討していきたい。

最後に、第1章でも述べたように、文内の、あるいは文間の意味的な要素を考慮しなければ本当に自然な段落分けが行なえないことは明らかである。これから、照応や省略などの意味的な要素も考慮して、自然な段落分けを行なうことを試みる予定である。そして、それらの成果と「手がかり語」の要素を合わせて、どの程度自然な段落分けができるか、研究を進めていきたい。

参考文献

- [1] Grosz and Sidner: Attention, Intentions, and the Structure of Discourse, Computational Linguistics, Vol.12, No.3, pp.175-204(1986)
- [2] 福本淳一：筆者の主張に基づく日本語文章の構造化、NL78-15 (1990)
- [3] 春日、桜井：日本語の表現と構造、pp.55-68、双文社 (1979)
- [4] 寺村、佐久間、杉戸、半澤編：ケーススタディ日本語の文章・談話、桜楓社 (1990)
- [5] 永野賢：文章論総説、朝倉書店 (1986)
- [6] 外山滋比古：日本語の個性、pp.17-29、中公新書 433 (1976)
- [7] 木下是雄：理科系の作文技術、pp.58-74、中公新書 624 (1981)
- [8] 大出晃：日本語と論理、pp.109-133、講談社現代新書 (1965)
- [9] 金田一春彦：日本語新版 (下)、pp.140-146、岩波新書 3 (新赤版) (1988)
- [10] 新明解国語辞典 第三版、三省堂 (1981)
- [11] 国語学会編：国語学大辞典、pp.592-593、東京堂出版 (1980)

表 B: 天声人語の調査結果

語句	段落頭	段落内	原結合度
(	0	22	100
あの	0	13	100
そして	3	50	94
だから	3	38	93
これ	12	86	88
とくに	2	15	88
それ	22	146	87
その	36	229	86
なぜ	5	28	85
つまり	4	22	85
こんな	2	11	85
前	2	11	85
われわれ	2	10	83
そこ	3	14	82
だが	22	91	81
そういう	7	29	81
そこで	6	22	79
が、	12	43	78
ただ	7	25	78
しかも	5	18	78
ところが	6	20	77
ここ	3	9	75
それに	15	42	74
また	9	26	74
当時	8	23	74
しかし	46	125	73
たとえば	16	42	72
私	5	13	72
「	199	434	69
いま	23	49	68
ある	12	25	68
そんな	14	28	67
そう	7	14	67
この	78	146	65
彼	11	18	62
もし	8	13	62
こんど	13	20	61
こう	21	30	59
第	13	19	59
まず	5	7	58
それでも	7	9	56
もう	14	13	48
きょう	10	9	47
最近	15	11	42
もともと	16	10	38

表 A: 日経サイエンスの調査結果

語句	段落頭	段落内	原結合度
そして	0	65	100
すなわち	0	38	100
なぜ	0	18	100
という	0	15	100
たとえば	1	51	98
したがって	2	48	96
つまり	2	36	95
それ	2	42	95
もし	1	21	95
彼	1	15	94
もちろん	1	13	93
また	7	85	92
その	9	94	91
これ	20	146	88
ある	2	14	88
しかし	24	130	84
この	48	242	83
さらに	9	31	78
そこで	8	28	78
こう	7	24	77
ここ	9	28	76
一方	8	23	74
ところが	6	17	74
前	3	8	73
このよう	20	51	72
これまで	4	10	71
「	7	16	70
すでに	5	11	69
最初	5	9	64
第	6	10	63
現在	10	16	62
いま	5	8	62
次に	8	12	60
私	19	25	57
まず	9	11	55
最近	9	8	47
ところで	9	8	47
もう	6	5	45
以上	10	2	17
さて	26	4	13