段落分けを用いた日本語文章における結束構造の検討

山本和英[†] 増山繁[†] 内藤昭三[††]

本論文は，日本語文章中の複数の文間に存在する結束性を解析することを目的とする．複数文からなる文章の解析における基本的な問題は，文間に存在する結束性を見出すことである．本論文では結束性を構成する要因の中から，「手がかり語」，すなわち接続的語句と，語の類縁性，すなわち出現した語の意味的な類似性の二つに着目した．まず，接続的語句が結束性に与える影響を，段落分けを行うことによって検証した．実際の文章の各文頭に出現する語の傾向に基づいて手がかり語を定義し，各語の統計的特徴と段落長の要素を用いて，計算機によって段落分けを試みた．段落分けの自然さを，原文の再現性およびアンケート結果の2種類の基準により評価した．次に，語彙的要素を用いた段落分けを試みた．シソーラスを使用して語句の類縁性を数値化し，これに基づき文章の結束構造をグラフによりモデル化したものを結束グラフと定義し，段落の設定を評価する関数を，それに基づいて定義した．さらに，前述の手がかり語と語句の類縁性の二つの要素に基づく3種類の方法により，実際に段落分けを試みた．その結果，類縁性のみを考慮した場合と比較して，3種類の方法のいずれにおいても再現率，適合率が向上した．また，この出力結果の自然さを評価するために，再びアンケート調査を実施したところ，作成した評価関数の自然さを支持する結果が得られた．

Cohesion Structure of Japanese Sentences and Paragraphing

KAZUHIDE YAMAMOTO,[†] SHIGERU MASUYAMA[†] and SHOZO NAITO[††]

This paper explores the cohesion in Japanese sentences. When we analyze a text, a problem of how strong each sentence in the text connects each other arises. In this paper, we focus on connectives and lexical cohesion which mainly affect cohesion. First, we inspect how strong connectives affects cohesion using paragraphing. We define cue words by investigating the beginning part of each sentence. Naturalness of paragraphing by cue words and length of sentences is evaluated by coincidence rate to the originals and consequence of a questionnaire. Then, we attempt to paragraph using lexical cohesion. We use thesaurus as the major knowledge-base for elucidating lexical structure in the text. We propose a graph model of cohesion structure of a text, called a cohesion graph, and formalize an evaluation function of paragraphing based on this graph. We paragraph by considering cue words and lexical cohesion using three combined methods. The experiments show that, compared with using lexical cohesion alone, the coincidence rate is improved. The consequence of the questionnaire also indicates that the proposed evaluation function is natural.

1. はじめに

構文規則に従って，品詞を並べても必ずしも意味のある文が成立しないのと同様に，相互に関連のない文を羅列しただけでは意味的にまとまりのある文章は必ずしも成立しない．すなわち，文章中での文の出現順序には，意味的なつながりをつけるための何らかの制約が存在し，逆に，この制約を満たす文の順序により，文章に意味的なまとまりが生じていると考えることができる．本論文の目的は，文章の意味的なまとまり，つまり文章内の結束性を解析することにある．また一般に，段落は文章の意味的なまとまりを示す一つの要素とされている．本論文では，文章の結束性が表出した結果が段落であると考え，実際に文章を段落に分けることによって，文章の結束構造を解明しようと試みた[1)~4)]．

本論文に関連した研究としては，Morris and Hirst によって英語を対象にして語の類縁関係から文章構造を解析した試みがある[5)]．この文献では，シソーラスに *Roget's International Thesaurus* を用いて，語の類縁性に注目した文章の構造解析を行っている．し

† 豊橋技術科学大学知識情報工学系
Department of Knowledge-based Information Engineering, Toyohashi University of Technology
†† NTT ソフトウェア研究所
NTT Software Laboratories

かし，数学的定式化は行っていない，段落を考慮していない，などの点で本研究とは異なる．

また日本語では，山崎が文章の構成単位として「意味的段落」というものを認め，異なり語数などの話題の展開を計る尺度を利用して，実際にその単位を切り出す方法を提案している[6]．本論文を除けば，この研究が筆者らの知る範囲で，日本語文章を段落に分ける試みを行っている唯一の研究であるが，以下のような問題点，および本論文との相違点がある．

1. 「用意」と「準備」のような，同義語の語の置き換えについても全く異なる語として計算してあり，同義語などの類縁性に対する適切な考慮が行われていない．
2. 「語の持つ意味情報の殆どを捨象している」[6]ため，解析の深さに限界がある．
3. 段落の分割基準が主観的であり，不明確である．
4. 「データ数が少なく充分な考察とは言えない」[6]．
5. 計算機を用いて解析していない．

本論文は段落分けそのものが目的ではないが，段落分けが利用可能な応用として，文章作成の支援や推敲の支援がある．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うための基礎データを得ることができる．

2. 手がかり語を用いた段落分け

本章では，文章の結束性の表現手段の一つである「手がかり語」，すなわち接続詞，副詞などの接続的語句が結束性，および段落の設定に与える影響を，段落分けを行うことによって検証する[2]．

2.1 文頭の単語出現調査

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った．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．統計調査は，2種類のテキストに対して，個別に行った．

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予め選び，それらの頻度を調べるという方法で行った．これらの中には，文間に出現して陽に文と文を接続する役割をもつ接続詞，副詞の他にも，(代)名詞，連体詞の中で照応によって文間の接続機能を果たす単語も含めて調査を行っている．この調査の，抽出対象の語の例を**表1**に示す．その他の専門用語などの語については，予備調査の結果，テキストに依存しないで統計的に大きな割合で出現する単語は存在しなかったので，今回の調査対象には含めなかった．

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編である．また調査は，計算機を使用して最長一致法によって対象となる単語を検出した．調査結果の統計データを**表2**に示す．ただし，表2で抽出語の割合とは，調査を行ったすべての文のうちで，文頭に調査対象単語が出現していた文数の割合を示している．

調査結果の品詞別データを**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．この結果は，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端的に示している[7]．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る傾向にあり，段落頭での使用頻度との差は日経サイエンスにおいて顕著である．

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，日経サイエ

表1 調査対象単語の例
Table 1 Examples of extracted words.

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2 調査を行った文章
Table 2 Statistical figures of the investigation.

項目	日経サイエンス	天声人語
調査対象	13編	505編
総段落数	824	3039
総文数	3501	10233
総文字数	212403	416006
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%

ンスでは抽出した語の4割近くに達するという結果を得た．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．この傾向は，日経サイエンス，天声人語という，文章の種類に関係なく見られる．その理由として，連体詞や代名詞の使用による前方の文との間の照応関係により，文間の結合が強くなることが考えられる．

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．このことは，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，抽出した語でどの文章にも出現する可能性のある語のうち，頻度の低い語（本論文では頻度1の語とした）を除いた語すべてを「手がかり語」と定義する．

2.2 計算機による段落分けの試行

以上の調査結果に基づき，計算機によって実際に段落分けを試みた．段落分けアルゴリズムの設計は，次のような方針に従った．

段落分けは，*On-line* 的方法，すなわち順番に入力文を読み込んでいき，文ごとに段落分けを行うかどうかを決定するという，逐次的な方法で行う．段落分けには，手がかり語に関する情報と文長の情報を使用する．文長は，文の情報量を反映していると考えることができるので，段落分け要因の一つとみなし，導入した．

表 3 抽出語の品詞別割合（日経サイエンス）
Table 3 Rate of a part of speech (Nikkei Science).

	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%
そ の 他	6%	5%	6%
指 示 語	38%	32%	40%

表 4 抽出語の品詞別割合（天声人語）
Table 4 Rate of a part of speech (Vox Populi, Vox Dei).

	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%
そ の 他	21%	25%	21%
指 示 語	33%	28%	35%

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．また，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，以下のような段落分けのアルゴリズムを導入した．2.1 節の調査結果から，それぞれの手がかり語に対して，以下の式に従い，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．

$$C_0=100\frac{A}{B} \qquad (1)$$

ただし，

C_0：その手がかり語の原結合度

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

である．手がかり語の抽出対象に含まれない，専門用語等の語については，全文に対して，平均的な結束性を示すと仮定し，それらの語全体を一つの手がかり語のように取り扱って，処理を行った．また，原結合度は2種類の文章それぞれについて別に算出する．

次に，段落分けを行う際に，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，段落を構成する文を順に S_1，$S_2,\cdots,S_n,\cdots$とする．現在，段落分けの判断を行おうとする文 S_n $(n>1)$ において，文 S_n までの段落長 L を，

$$L=\sum_{i=1}^{n-1} l_i \quad (l_i：文\ S_i\ の文長) \qquad (2)$$

と定義する．ただし，段落の第一文$(n=1)$のときは形式的に $L=0$ と定義する．以上の準備に基づき，その文の後で段落を分けるかどうかを判断するための文の結合度 C を，次のように定義した．

$$C=C_0+\alpha_1 L_m-\alpha_2 L \qquad (3)$$

$$=C_0+(\alpha_1-\alpha_2)L_m+\alpha_2(L_m-L) \qquad (4)$$

ここで，L_m：平均段落長，α_1，α_2：定数である．

各文末で算出した結合度が 50 以下ならば，その位置で段落を分ける．以後，この操作を反復し，文章全体の段落分けを行う．

日経サイエンスと天声人語の文章のうちで，2.1 節で手がかり語の調査を行ったのとは異なる文章を対象にして，前述のアルゴリズムを用いて実際に段落分け

を行った．対象は，日経サイエンス6編，天声人語40編である．対象となる文章から段落分けを除去した文章を入力とし，このアルゴリズムにより段落分けを行わせた．言語は，*Common Lisp*(KCL) を用い，*Sun SPARC Station I* 上で実行した．実験に使用するパラメータ α_1 や α_2 の値は，あらかじめ推定を行うことが困難であるため，先に α_1 を仮に固定し，出力段落数が原文の段落数に近くなるように α_2 を決定するという方法を繰り返す，という試行錯誤によりパラメータを決定した．

2.3 原文との比較による自然さの検証

出力された文章の段落分けの自然さの評価基準の一つとして，原文の段落との一致度をあげることができる．比較的良好な結果を得られたいくつかのパラメータ対についての，原文段落分けとの一致した割合を**表5**に示す．

表5では，原文とどの程度一致したかを，再現率，適合率の二つの評価基準で表現している．すなわち，再現率 R_r，適合率 R_p は次式で定義される．

$$R_r=\frac{N_{ab}-1}{N_a-1} \quad (5)$$

$$R_p=\frac{N_{ab}-1}{N_b-1} \quad (6)$$

ただし，

N_{ab}：原文の段落と出力段落とで共通する段落数

N_a：原文の段落数

N_b：抽出した段落数

である．再現率，適合率共に高いほど，原文に対する再現性が高いことを示す．

計算機で出力された結果の再現性が低いことの理由の一つとして，2.1 節での調査において，手がかり語が日経サイエンスでは約6割の文頭にしか出現していないことをあげることができる．このことは，手がかり語情報としては，全体の6割しか使用していないことに対応する．したがって，手がかり語情報だけでは，再現率は最大6割であると考えられる．

表5 原文に対する一致の割合
Table 5 Coincidence rates against the originals.

文章	α_1	α_2	R_r	R_p
日経サイエンス	0.02	0.1	40.3%	36.8%
	0.02	0.15	48.4%	33.5%
	0.03	0.15	44.2%	32.8%
天声人語	0.01	0.1	38.1%	34.6%
	0.1	0.2	37.1%	35.7%
	0.3	0.5	34.1%	30.2%
	1.5	2.0	35.1%	31.8%

2.4 アンケートによる自然さの検証

文章の段落分けの自然さの判断基準は一般には複雑であり，本アルゴリズムによる段落分けの自然さの評価を数値化することは容易ではない．本論文では，段落分けアルゴリズムの評価基準の一つとして，人間がその出力結果を読んで，原文の段落分けと区別ができないかどうかということを採用した．

ここでは，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った．すなわち，被験者に対して，文章の原文を1編以上と，計算機によって段落分けした文章を1編以上含む，合計5編の文章を提示する．5編の記事の内訳は被験者には示さない．そしてこれらの文章のそれぞれについて，原文かどうかを当ててもらう．このアンケートを日経サイエンスと天声人語について行う．ただし，天声人語は提示する合計編数を7編とする．

以上のような方法によって，実際にアンケートを工学系の大学生，大学院生11人（日経サイエンスは10人）に対して行った．その結果を**表6**に示す．

表6に示すように，天声人語に対する計算機による出力結果は，4割以上が原文と認識された．この結果は，天声人語の原文に対して，原文再認識率が4分の3であることを考慮に入れると，比較的高い数値であると考えることができる．日経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．この理由としては，日経サイエンス1編の段落数が比較的多いために，明らかに不自然と思われる段落の分け方が出現する可能性が高くなることが考えられる．

3. 語の類縁性の導入とそれを追加した段落分け

3.1 語彙的手段による結束性

本章では，前述の手がかり語の要素の他に，語彙的手段による結束性にも着目し，この両者が文章の結束

表6 アンケート調査の結果
Table 6 Recognition rates to the originals at the questionnaire.

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

構造に与える影響を考察する[2].

語彙的手段による結束性は，同一語句を含む，意味の類似性 (similarity) による結束性と，意味の近接性 (contiguity) による結束性に分類できる[8]．前者はさらに，

男の子が立っていた．その少年は泣いていた．

のように，類義語 (synonym) の場合と，

男の子が立っていた．その子供は泣いていた．

のような上位語 (superordinate) や下位語 (hyponym) の場合に分類できる．一方，後者の意味の近接性の例は以下である．

空はとても青い．今日は雲一つない天気である．

前者の要素をとらえるための資料として，シソーラス (thesaurus) が利用できるが，後者をとらえるためには大規模な知識ベースの構築が必要である．本論文では語彙の近接性は扱わず，同一語句，上位・下位語，類義語，対義語のみを考慮し，以後この関係をまとめて語の類縁性と呼ぶ．また，語間の類縁性を測定する基準として，本論文では角川類語新辞典[9]をシソーラスとして使用する．

3.2 結束グラフ

文章中の文間の結束度を表現するために結束グラフを構成する．結束グラフ $G=(V, E, w)$ を次のように定義する．

$V=\{v \mid v$ は文章中の一つの文 s に対応する$\}$

$E=\{(u, v) \mid u$ に対応する文中の語と v に対応する文中の語でシソーラスの中分類，小分類での同一の分類に属する，または同一の語であるものが存在する$\}$

$w: E \to \mathcal{R}$ ($\mathcal{R}$：実数の集合)，$(u, v) \in E$ に対し，下の式(7)で節点 u と v の結束度 $w(u, v)$ を定義する

$$w(u, v)=\exp\{\lambda d(w_1 x_1+w_2 x_2+w_3 x_3)\} \quad (7)$$

d：u と v 間の距離

x_1：u と v での同じ語の組数

x_2：u と v での小分類一致の組数 (x_1 を除く)

x_3：u と v での中分類一致の組数 (x_1, x_2 を除く)

$w_1>w_2>w_3>0$, $\lambda<0$

すなわち，二つの文の語彙的結束度は，前述した3段階の枝の強弱を定数 w_1, w_2, w_3 とし，それぞれに枝の本数を掛けたものの総和として定義している．結束度が文間の距離に関して単調非増加となる性質を持たせるために，指数関数を採用した．

3.3 結束度の評価関数

与えられた文章に対する段落分けの評価関数を，以下のように定義する．

max 評価関数

$$=\alpha \sum_{p_i \in T}\left(\sum_{s_j \neq s_k \in p_i} \frac{w(s_j, s_k)}{|p_i|}-\sum_{s_j \in p_i, s_k \notin p_i} \frac{w(s_j, s_k)}{|p_i|}\right) -\beta \sum_{p_i \in T}\left(\sum_{s_j \in p_i} l(s_j)-\frac{\sum_{p_j \in T}\sum_{s_k \in p_j} l(s_k)}{|T|}\right)^2 \quad (8)$$

ただし，

$T=\{p_1, p_2, \cdots, p_n\}$

：段落 $p_1, p_2, \cdots, p_n$ からなる文章

$p_i=\{s_j, s_{j+1}, \cdots, s_k\}$

：文 $s_j, s_{j+1}, \cdots, s_k$ からなる第 i 段落

$l(s_i)$：文 s_i の文字数

$w(s_i, s_j)$：文 s_i と 文 s_j の語彙的結束度

$|S|$：集合 S の要素数

$\alpha>0$, $\beta>0$：定数

である．

式(8)の二項は，それぞれ文間の語彙的結束度と段落長（段落内の文字数）の要因を定式化したものである．文間の語彙的結束度は，段落内の文間の結束度はプラス要因，段落間の文間の結束度はマイナス要因とした．また，段落長は，バランスの良さをプラス要因とした．

3.4 両要素の併用法

2.2節での式(1)の原結合度 C_0 と，3.3節での評価関数式(8)の両要素を併用するためには，以下の方法が考えられる[3].

1. 一つの数値に集約する方法（集約法）
   原結合度と評価関数式の両者に対して加算，あるいは乗算等の演算を施して一つの数値に集約する方法である．
2. 原結合度を先に考慮する方法（閾値法）
   まず原結合度に着目して，その数値がある上限閾値以上の個所では段落をつなぎ，下限閾値以下の個所では段落を分ける．残りの個所に関しては，評価関数式を考慮して段落分けを行う．
3. 集約法＋閾値法
   前述の集約法と閾値法は，別の部分で統合を行っているために，同時に実現することも可能である．これを同時に行うこの方法は，原結合度の値が両端に近い場合は閾値法を採用して，中間の部分は集約法を採用することによって実現

される．

文頭に現れる手がかり語の情報は，類縁性などによる文章の結束構造を，明示的に表現したものとみなすことができる．この観点からは，手がかり語の情報は語の類縁性の情報の一部であるともみなすことができるので，評価関数式を先に考慮する方法は除外した．

3.5 システムの構成

上記のモデルに基づき，実際に段落分けを行うシステムを作成した．システムはすべて，*Common Lisp* (KCL) を用い，*Sun SPARC Station I* 上で作成した．システムは三つの部分から構成されている．

1. 形態素解析部

   この部分では，語尾変化処理と，文節数最少法による形態素解析を行う．本実験ではシソーラスに掲載されている語を切り出すことを目的としているので，簡易の形態素解析である．辞書には，シソーラスと同じ角川類語新辞典[9]に固有名詞の辞書を付加したものを採用した．

2. 評価関数値の算出

   形態素解析の終了後，すべての文の組合せについて，前述した計算式で評価関数値を計算する☆．実験では，3種類の枝の重みの比 $w_1 : w_2 : w_3$ を，10 : 8 : 3 としている．

3. 段落分け

   Off-line による方法，つまり文章全体がすでに入力されてから段落分けを行っている．具体的な段落分けは，以下のとおりである．

   (a) 初期設定として，1段落1文とする．閾値法では，まず原結合度と閾値に基づき，段落に分けない個所と分ける個所を決定する．

   (b) ある位置で隣接する段落を併合したとして，その時の評価関数の減少量が最も小さい位置（あるいは最も増加する位置）を段落のつなぎ目とする．

   (c) この操作を繰り返し行い，目的関数がそれ以上改善されなくなった段落の分け方を最も自然な段落分けの候補とする．

段落分けの近似方法については，順次併合する方法と，順次分割する方法が考えられる．本論文では，予備実験で相対的に結果の良かった前者を採用した．後者が相対的に悪かった理由は，初期状態から数段落を構成するまでの段落決定において，後者では段落長の要因が大き過ぎるためと考えられる．

3.6 実験結果

実験は，朝日新聞「天声人語」50編を用いて，

1. 類縁性のみを考慮した方法
2. 集約法
3. 閾値法
4. 集約法＋閾値法

の4種類について行った．実験結果を**図1**，**図2**に示す．ただし閾値法では，原結合度 80 以上の場合に段落をつなぎ，原結合度 20 以下の場合に段落に分けるとして実験を行った．また，集約法による実験では，集約演算に乗算を用いた[4]．

図1，図2の縦軸の再現率，適合率は，それぞれ式(5)，(6)に従う．また，横軸の α は，(8)式において β を 0.001 に固定した時の α である．両図によると，集約法，閾値法共に，従来の語の類縁性のみを考慮した方法よりも再現率，適合率共に向上している．これは，前後の2文間が語彙的に強い（あるいは弱い）

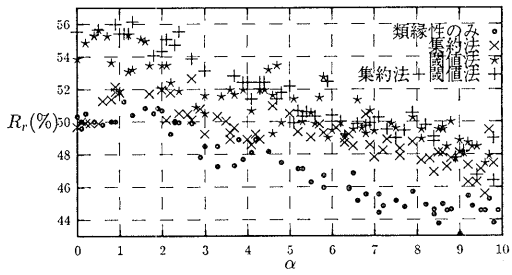


図1 α と再現率との関係
Fig. 1 α vs. rate of recall.

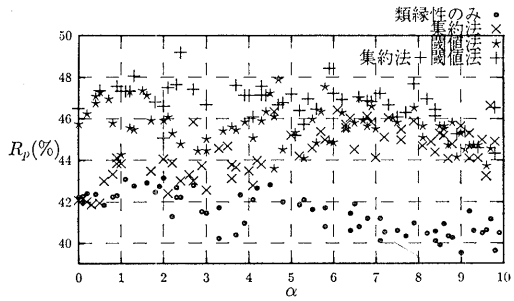


図2 α と適合率との関係
Fig. 2 α vs. rate of precision.

☆ 文間の結束度の計算は文数に対して二乗に比例した計算量，記憶量となるが，実際のテキストでは比較的少ない文数でくくられており，十分実用可能であると考えられる．また，ある距離以上の文間の結束度はすべて0と近似した修正を行えば，長いテキストでも計算量，記憶量共に軽減が可能であり，さらに実用的になる．

結束性を持った関係になっていても，手がかり語によってその結束性の強弱を変化させることができることを示している．

本論文で使用している段落分けはヒューリスティックな方法のため，その妥当性を検証する．図1，あるいは図2より，最も妥当であると考えられるパラメータを設定して段落分けを行った出力結果が，すべての段落分けの評価値のうちどの程度の順位になっているかを検証した結果を**図3**に示す．図3で，横軸は（近似解の全評価値に対する順位／すべての段落分けの組合せ）を対数軸で，縦軸はテキストの全文数を，図中の菱形1個は1テキストに対する値を，意味する．実験に使用した天声人語の最少文数は 14 文であった．すべての段落分けの組合せ数は，文数に対して指数的に増加する．24文以上のテキストに対しては，計算時間上評価できなかった．

図3は，本近似アルゴリズムが，全組合せの中の評価値が最高値の段落分けを必ずしも出力するとは限らないが，上位の段落分けを出力しており，近似アルゴリズムの妥当性を示している．

4. 段落分けの自然さの検証

本章では，段落分けの自然さの評価を，原文の再現性という基準だけで行うのは不十分であることを明らかにするとともに，段落分けのすべての可能性の中での原文段落分けの評価値順位，アンケート調査による人間の感覚との相関という二基準によって評価関数の妥当性の検証を行う．

4.1 原文との比較による評価関数の妥当性

原文の段落分けは自然であると仮定すれば，段落分けのすべての可能性の中での原文の段落分けの評価値順位を調べることによって，評価関数の妥当性を検証することができる．そこで，この検証結果を**図4**に示す．原文の評価値は比較的上位にあるので，本論文で定式化した評価関数は，この基準の下で妥当であると考えられる．

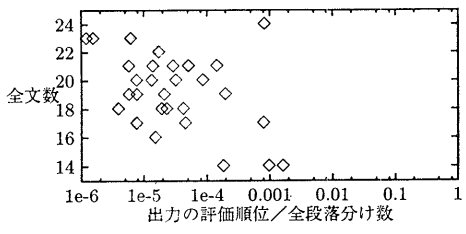


図 3 出力文章の評価値順位
Fig. 3 Ranking of the heuristic outputs in the evaluation.

4.2 アンケート調査

段落分けの評価関数と人間の感覚がどの程度一致するかを検証するため，アンケート調査を行った．アンケートはまず，文単位に切った文章と，その文章の段落分けとして尤もらしい候補（原文の段落分けを含む）を五つ提示する．その上で，

1. 提示した五つの候補を自然な順に並べる
2. 提示した候補の中から原文を選択する
3. （提示した五つの候補に関係なく）最も自然な段落分けを指摘する
4. 絶対に段落に分けない位置を（任意個数）指摘する

という四つの設問に答えてもらう．提示文章には，朝日新聞「天声人語」3編を用いた．被験者は工学系の大学生・大学院生であり，21名から回答を得た．

4.3 調査結果

評価関数の順位と人間の感覚による順位の相関を評価するために，スピアマンの順位相関係数（*Spearman's Rank Correlation Coefficient*）[10] を用いた．この相関係数値を，−10（逆の相関がある）〜+10（相関がある）に正規化したものを**図5**の縦軸に示す．図5の横軸は，被験者が最も自然であると回答した段落分

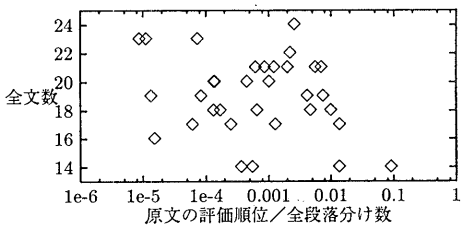


図 4 原文の評価値順位
Fig. 4 Ranking of the originals in the evaluation.

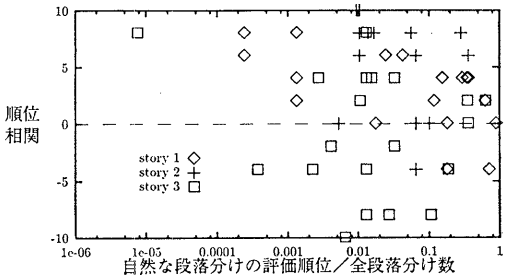


図 5 順位相関係数の分布
Fig. 5 Rank correlation coefficient.

けの評価関数値の，段落分けのすべての可能性の中での順位を示している．

各文章ごとの順位相関係数の平均と，原文を正しく指摘した人数を**表7**に示す．

また，評価関数値の最も良い段落分けの妥当性を評価するために，アンケート調査から，評価関数値の最も良い段落分けの，各段落分けの位置dに対して，以下の条件を満たすアンケート調査結果を集計した．

- 段落分けの位置dが「最も自然であると回答した段落分け」(「適当な位置」と呼ぶ) に含まれる．
- 段落分けの位置dが「絶対に段落に分けないと回答した段落分け位置」(「不適当な位置」と呼ぶ) に含まれる．

これを文章別に集計したものを**表8**に示す．(ただし，表8で各欄は (n_1/n_2, n_1: 適当な位置であると指摘した人数，n_2: 不適当な位置であると指摘した人数) を示す．段落の位置 ($a, b, \cdots$) は文章によって異なることに注意．)

4.4 考　　察

- 表7は，原文の段落分けであることの認識率が非常に低いことを示している．このことは，読者にとって著者による段落分けが最も自然な段落分けであるとは限らないことを意味する．これは，原文の段落分けだけが自然な段落分けではないこと，自然な段落分けには自由度があることを示している．
- 表7より，順位相関係数は，文章によっては必ずしも正の値をとらない．この原因としては，アンケートで提示した五つの選択肢が，いずれも尤もらしい候補であったことが考えられる．すべての段落分けの可能性の中から全くランダムに選んだ五つの選択肢の順位づけと，評価関数との間で順位相関をとれば，相関係数値は，調査の最高値 (5.43) 程度，あるいはそれ以上になると予想される．このことから，提示した評価関数と人間の感覚との間の相関度は高いと考えられる．
- 表8より，最も評価値の高かった段落分けの位置の多くは，アンケート調査で「適当な位置」と指摘されていることから，計算機による出力結果の妥当性を支持している．一方，表中の一か所 (文章 2-a) は，ほとんどの被験者が「不適当な位置」と回答している段落位置にもかかわらず，原文ではこの位置で段落を分けている．このことから，著者と読者の段落に対する感覚が必ずしも一致しないことがわかる．

表7 順位相関係数の平均および原文指摘者数
Table 7 Means of R.C.C. and the number of persons to indicate the originals.

	文章1	文章2	文章3
順位相関係数	3.24	5.43	−0.57
原文指摘数 (21名中)	7	0	1

表8 各段落位置の適当，不適当 (21名中)
Table 8 Appropriateness of paragraph positions (out of 21 persons).

段落位置	a	b	c	d	e	f
文章1	9/2	12/2	21/0			
文章2	0/16	20/0	12/1	17/0	19/0	
文章3	3/6	16/3	3/4	17/1	8/0	9/2

5. おわりに

本論文では，日本語文章中の結束性を解析することを目的にして，「手がかり語」，すなわち接続的語句と，語の類縁性，すなわち出現した語の意味的な類似性の二つの要素に着目し，計算機を用いて段落分けを行うことによってこれらの要素の影響を考察した．その結果，次の2点が明らかになった．

1. 日本語文章の結束性には手がかり語と語の類縁性が共に影響を与えているが，特に，明示的，意識的な前者の要素よりも後者の方がより大きな影響を与えている．
2. 日本語文章には，自然と感じる段落分けが存在する．ただし，これは唯一ではなく，また，個人差などの要素によってある程度変動する．

今後の課題としては次の2点をあげることができる．

- 語彙の近接性の考慮

  前述したように，3章の実験では語彙の類縁性のみを考慮して実験を行っており，語彙の近接性は取り扱っていない．しかし，語彙的な手段による結束性を考慮するためには語彙の近接性の考慮も必要である．しかし，これを取り扱うためには，大規模知識ベースなどの情報が必要であり，その開発が待たれる．
- 照応・省略の考慮

  本論文では，結束性表示の要素として，手がかり語と語彙の類縁性の二つを考慮した．これらの他に結束性を示している要素として，指示語の使用による照応，あるいは省略がある[8]．これらの要素につい

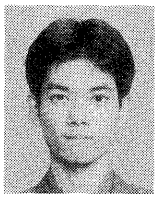
ても今後，検討していきたい．

謝辞 本研究で，シソーラスに使用した「角川類語新辞典」を機械可読辞書の形で提供いただき，その使用許可をいただいた(株)角川書店に深謝する．

参考文献

1) 山本和英，増山 繁，内藤昭三：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，情報処理学会研究会資料，NL 84-9 (1991)．
2) 山本和英，増山 繁，内藤昭三：語の類縁性を用いた日本語文章の段落分けの試み，第 45 回情報処理学会全国大会論文集，6G-8 (1992)．
3) 山本和英，増山 繁，内藤昭三：手がかり語及び語の類縁性を併用した段落分け，情報処理学会研究会資料，NL 92-6 (1992)．
4) 山本和英，増山 繁，内藤昭三：段落分けに関わる諸要素の評価について，第 46 回情報処理学会全国大会論文集，7B-8 (1993)．
5) Morris, J. and Hirst, G.: Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text, *Computational Linguistics*, Vol. 17, No. 1, pp. 21-48 (1991).
6) 山崎 誠：文章の話題の展開を計る尺度，計量国語学，Vol. 13, No. 8, pp. 346-360 (1983)．
7) 永野 賢：文章論総説，朝倉書店 (1986)．
8) 池上嘉彦：テクストとテクストの構造，国立国語研究所 (編)，談話の研究と教育 I，pp. 7-42，大蔵省印刷局 (1983)．
9) 大野 晋，浜西正人：角川類語新辞典，角川書店 (1981)．
10) 竹内 啓 (編)：統計学辞典，東洋経済新報社 (1989)．

(平成5年2月23日受付)
(平成6年7月14日採録)



山本 和英（学生会員）

1969 年生．1991 年豊橋技術科学大学知識情報工学課程卒業．1993年同大学院修士課程修了．現在同大学院博士後期課程システム情報工学専攻在学中．自然言語処理，特に談話処理の研究に従事．言語処理学会，ACL（計算言語学会）各会員．

増山 繁（正会員）

1952 年生．1977 年京都大学数理工学科卒業．1983年同博士後期課程修了．1982年日本学術振興会奨励研究員．1984年京都大学工学部助手．1989 年豊橋技術科学大学知識情報工学系講師．1990年同助教授，現在に至る．グラフ・ネットワークのアルゴリズム，組合せ最適化，並列アルゴリズム，自然言語処理等の研究に従事．工学博士．電子情報通信学会，ACL（計算言語学会），言語処理学会，日本オペレーションズ・リサーチ学会等会員．

内藤 昭三（正会員）

1955 年生．1979 年京都大学工学部数理工学専攻修士課程修了．同年 NTT 入社．現在 NTT ソフトウェア研究所広域コンピューティング研究部所属．自然言語処理，要求仕様獲得の研究開発に従事．電子情報通信学会，人工知能学会，言語処理学会，ACL（計算言語学会），計量国語学会各会員．