自然言語処理　33－2
（1982. 10. 22）

語句反復形式による日本語テキストの構造解析

坂本義行
（電子技術総合研究所）

岡本哲也
（電気通信大学）

0．　はじめに

文と文とをつなぐ文間結合要素の研究は，計算機への言語理解に必要なことはすでに幾指摘が成されている．ここでは，テキスト構造の決定あるいは主題分析を形式的に行なう方法として語句反復による方法について論じる．既に日本語の文章について，人間の手による短編小説及び数編の科学文献について分析し，報告を行なった．[1)2)3)] 今回は以下の5編のテキストについて語句反復による文間結合率，文間結合距離等について，計算機と人間の手による結果を比較し，その詳細な分析から計算機による自動解析の可能性，又その特性を利用しての自動主題分析，自動抄録等への応用の可能性を確かめた．

テキストA　超音波増幅
テキストB　脳とオートマトン
テキストC　石油化学工業
テキストD　日本の化学工業
テキストE　生物と無生物

1．　文間結合関係

語彙要素の反復（lexical paral－lelism）は文章（text）中で文と文との接続関係を示す重要な担い手である．反復語句（lexical equiva－lent）は，それが出現した2っの文の間で構文上あるいは品詞が一致する必要はない．形態素的，意味的な一致が可能である．すなはち，2個の語彙関係（lexico－semantic relation）が，同義，類義あるいは対義である場合も存在する．特殊な場合として「超音波」，「音波」，「波」といった，形態素的にも意味的にも部分一致する場合がある．

文間結合関係にはこのほかに，構文関係（syntactic devices）としてQUIRK[4)]によれば，substi－tution, time and place relaters, logical connecters, discourse reference, comparison, ellipsis, structural parallelism がある．

日本語を例にとれば，代用詞に，これ，この，ここ，我々，それ．時間場所関係詞に，次に，以前の，以上の，すぐに，まだ等，論理関係詞に，及び，または，しかし，第二に等がある．このほか，林[5)]の行なった分類もある．しかし今回は，いずれも分析の対象外とした．

次に，主題の展開に関して，セブボ[6)]は，主題が登場人物とともに展開し，文の構文上の主語，述語が主題（topic，古い情報），解説（comment，新しい情報）あるいは心理主語と述語に一致すると仮定できる単純な物語，たとへばトルストイの寓話や童話を資料に，文の規格化（normalization），すなわち，文を等位接続構造を含まない単文に分割した後，その構文的主語，述語（支配，従属）の語句反復を調べ，テキストの展開模式を決定する問題，またこの方法が科学技術論文の自動抄録，自動索引に有効なことを示している．しかし，ここでは構文分析（支配，従属関係）あるいは文の規格化は行なわない．

一般に文は先行文から「既知のもの」を文頭部分に加える[7)]．又日本語では述語を別にすれば，その語順はかなり自由であるとして，文頭近傍の語句に着目して，文間結合関係及びテキストの主題分析を行なった．

さらに日本語の特徴として，名詞は屈折せず，用言は変化しない語幹と活用語尾を有する．又，重要な概念，技術用語は，漢字，カタカナ（外来語），アルファベットで表記

し，一方助詞，助動詞，活用語尾，格指標，法，時制，アスペクト等は平仮名で表記される．すなわち，語彙要素として非平仮名列を対象として実験を行なった．

2．　語句反復による文間結合率

語句反復には，大きく分けて次の3っの型がある

t＝1；　同一語句の反復（完全反復）

t＝2；　語句の構成要素の一部反復（部分反復）

t＝3；　意味的に関連する語句，すなわち同義，類義，内包，外延，対義のいわば，シソーラス的関係のある語彙要素の反復（類義反復を含む）

本研究では，構文分析あるいは文の規格化を行なわず，語句の選択方法として，文頭から5番目まで（w＝1，2，3，4，5）の自立語をインジケータ候補に選んだ．

テキスト中の着目する第j文（SJ；j＝1，2，～，N）のインジケータ候補に対して，これと同一の語句を含む先行文の内，もっとも近い文，第i文Si（j＞i）とが結合関係にあるとし，この語句を語句反復インジケータ（lexical parallelism indicators）と呼び，I_j^{tw} と定義する．

テキスト中に見出された語句反復インジケータの総数をnとしたとき，そのテキストの文間結合率（α^{tw}）を以下のようにに定義する．

α^{tw}＝（n／N－1）＊100

ここで，

n：テキスト中で I_j^{tw} が見出された文の総数

N：テキスト中の文の総数

t：反復の型

w：文頭からの語句の順位

3．　語句反復インジケータによる文間結合距離

第j文（Sj）と第i文（Si）が結合関係にあるとき，そのj文からi文を引いた文の数を文間結合距離（D）と定義する．その式を次に示す．

D_j^{tw} ＝ j － i （j＞i）

ここでのt，wは，前の定義と同じ．また，iを支配文（governor sentence）の文番号，jを従属文（dependent sentence）の文番号とする．

この結合距離は語句反復インジケータの及ぶ範囲とも，またその概念の及ぶ範囲とも言えよう．

4．　実験と結果

文間結合率 － 計算機を用いて，形態，構文解析を行なわず，文字種のみに着目して，（計算機内部コードの識別のみによって），インジケータ候補を検出し，これについて I_j^{tw} を探索し，t＝1でかつw＝1，2，3，4，5について，個々独立の文間結合率を算出した結果を表1に示した．

Table 1　Lexical parallelism ratios of type 1 in computer experiment(%)

T＼w	1	2	3	4	5
A	60.4 (75)	61.9 (75)	57.1 (64)	54.2 (58)	56.4 (57)
B	68.2 (71)	64.4 (67)	56.3 (58)	58.4 (59)	57.4 (58)
C	59.4 (41)	45.5 (31)	43.2 (29)	37.5 (24)	32.2 (19)
D	57.2 (71)	61.2 (76)	54.9 (67)	52.5 (60)	56.7 (58)
D	41.1 (37)	53.3 (48)	49.4 (43)	42.1 (35)	50.0 (40)

Note: T - sample texts, w - sequence numbers of indicators, values in() are numbers of determined sentence connections.

表2は人間の手によるものであって，結合の型（t＝1，2，3），語句の順位（w＝1，2，3，4，5）によって，複数個の語句反復イ

ンジケータを有するとき，その中で最適のインジケータを選択し，これらの文について，文間結合率を求めた結果が示されている．

Table 2 Lexical parallelism ratios determined by hand(%)

T \ N-1 \ w		1	2	3	4	5
A	122	60.7 (74)	6.6 (8)	3.2 (4)	0.8 (1)	0.8 (1)
B	103	68.9 (71)	9.7 (10)	1.9 (2)	0.9 (1)	0.9 (1)
C	69	50.7 (35)	8.7 (6)	13.0 (9)	2.9 (2)	0 (0)
D	123	54.9 (67)	13.9 (17)	2.4 (3)	1.6 (2)	0 (0)
E	89	29.2 (26)	5.6 (5)	2.2 (2)	1.1 (1)	0 (0)

Note: N-1 --- the determinable maximum number of intersentential relations.

表１において，テキストＥを除き，ｗ＝１すなわち文頭の語句が最高の結合率（５７％～６８％）を示した．ｗが増加するにしたがって，通常，その結合率は減少する結果が得られた．この結果には多くのノイズを含んでいるが，一般に文頭近傍に語句反復インジケータが置かれており，それは，より文頭の近くに存在することを示している．

表１と表２を比較すると，ｗ＝１での値が非常に近い値を示している．これは，計算機の実験で語句反復インジケータをかなり正確に選択していることを示している．一方，ｗ＝２，３，４，５での値は大きく異なっており，表１ではそこに多くのノイズが含まれている．

表２ではテキストＥを除いて，ｗ＝１で５０％以上の結合率を示し，ｗ＝２以上ではその率が急激にまたほぼ単調に減少している．さらに，各テキストの結合率の合計は，７２％～８３％という高い結合率を示した．各テキストでの語句反復インジケータはその８０％以上が文頭に置

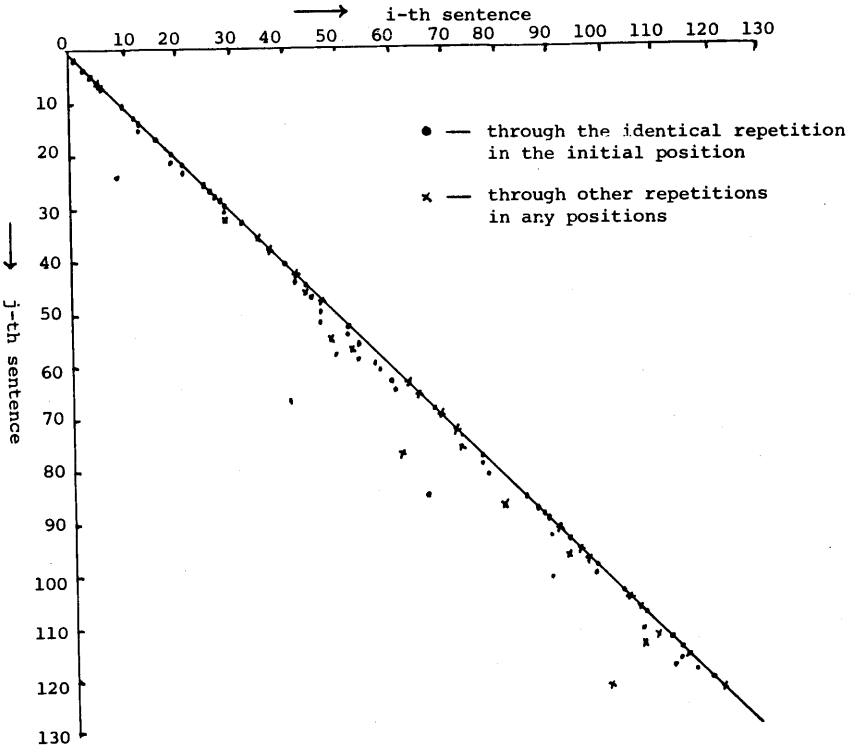

Figure 1 Lexico-semantic intersentential dependency graph in sample text A

Table 3 Lexico-semantic intersentencial dependency in sampe text 1

Indicator	J	I	D	w	t
音の (sound)	1	–	–	–	–
音が (sound)	2	1	1	1	1
超音波は (ultrasonic wave)	3	–	–	–	–
聞 (hear)	4	3	1	1	1
超音波の (ultrasonic wave)	5	4	1	1	1
一は (one)	6	5	1	1	2
第二の (the second)	7	6	1	1	2
この場合 (this case)	8	–	–	–	–
ここでは (here)	9	–	–	–	–
はじめにのべたように (as mentioned before)	10	–	–	–	–
音が (sound)	11	10	1	1	1
今 (now)	12	–	–	–	–
笛から (whisle)	13	12	1	1	1
波長より (wavelength)	14	13	1	1	1
波は (wave)	15	13	2	1	1
このことは (this fact)	16	–	–	–	–
波動的 (wave motion)	17	16	1	1	1
われわれが (we)	18	–	–	–	–
われわれの (our)	19	–	–	–	–
波長が (wavelength)	20	19	1	1	1
われわれの (our)	21	19	2	1	1
目では (eyes)	22	21	1	1	1
音で (sound)	23	21	2	1	1
超音波が (ultrasonic wave)	24	9	15	1	1
諸兄は (you)	25	–	–	–	–
コウモリは (bat)	26	25	1	1	1
コウモリは (bat)	27	26	1	1	1
コウモリの (bat)	28	27	1	1	1
レーダーは (ladar)	29	28	1	1	1
共に (together with)	30	29	1	2	1
レーダーは (ladar)	31	29	2	1	1
音波は (sound wave)	32	29	3	2	2
超音波には (ultrasonic wave)	33	32	1	1	1
例えば (for example)	34	–	–	–	–
途中に (halfway)	35	–	–	–	–
魚群探知機と (fish detector)	36	35	1	1	3
医学の (medicine)	37	–	–	–	–
診断には (diagnosis)	38	37	1	3	1
前おきが (introduction)	39	–	–	–	–
話の (discussion)	40	–	–	–	–
進行波増幅と (traveling –wave amplification)	41	40	1	1	1
第一図の (figure 1)	42	–	–	–	–
振子を (pendulum)	43	42	1	4	1
バネの (spring)	44	42	2	1	1
Iの	45	44	1	1	1
IIの	46	44	2	2	1
Iの	47	45	2	1	1
IIの	48	47	1	2	1
このようなことを (such thing)	49	–	–	–	–
図は (figure)	50	47	3	1	1
このことは (this fact)	51	–	–	–	–
t から	52	47	5	1	1
Iの	53	52	1	1	1
t から	54	52	2	1	1
繰返しを (repetition)	55	49	6	1	2
エネルギーの (energy)	56	54	2	1	1
振幅の (amplitude)	57	53	4	3	1
図を (figure)	58	50	8	1	1
Iの	59	54	5	1	1
エネルギーの (energy)	60	57	3	1	1
図を (figure)	61	58	3	1	1
このことから (this fact)	62	–	–	–	–

Indicator	J	I	D	w	t
エネルギーは (energy)	63	60	3	1	1
位相速度の (phase velocity)	64	63	1	1	2
Iの	65	61	4	1	1
振幅は (amplitude)	66	65	1	2	1
進行波増幅の (traveling –wave amplification)	67	41	26	1	1
このことを (this fact)	68	–	–	–	–
電波は (radio wave)	69	68	1	1	1
波を (wave)	70	69	1	1	2
電気信号は (electric signal)	71	–	–	–	–
電界の (electric field)	72	–	–	–	–
電子の (electron)	73	72	1	5	1
電子の (electron)	74	73	1	1	1
これは (this)	75	–	–	–	–
外側の (outside)	76	73	3	2	3
波の (wave)	77	62	14	3	1
Ve	78	77	1	1	1
(1)	79	77	2	1	1
これは (this)	80	–	–	–	–
(2)	81	78	3	1	1
これが (this)	82	–	–	–	–
マイクロウェーブ通信に (microwave communication)	83	–	–	–	–
物理学の (physics)	84	–	–	–	–
進行波増幅を (traveling –wave amplification)	85	67	18	1	1
二つの (two)	86	85	1	1	1
エネルギー源として (energy source)	87	81	6	1	2
電気系と (electric system)	88	87	1	1	1
圧電現象と (piezoelectric phenomenon)	89	88	1	1	1
圧電現象を (piezoelectric phenomenon)	90	89	1	1	1
諸兄は (you)	91	–	–	–	–
ピックアップは (pick-up)	92	91	1	1	2
圧電結晶には (piezoelectric crystal)	93	90	3	1	1
これは (this)	94	93	1	1	1
逆効果を (reverse effect)	95	–	–	–	–
レシーバには (receiver)	96	95	1	1	2
圧電結晶では (piezoelectric crystal)	97	93	4	3	1
圧電結晶中を (in piezoelectric crystal)	98	97	1	1	2
電界は (electric field)	99	98	1	1	1
圧電結晶で (piezoelectric crystal)	100	98	2	1	1
ロッセル塩や (Rochelle salt)	101	90	11	1	1
このようなときに (such time)	102	–	–	–	–
圧電半導体 (piezoelectric semiconductor)	103	–	–	–	–
圧電半導体は (piezoelectric semiconductor)	104	103	1	1	1
Cdsは	105	104	1	2	1
それは (it)	106	–	–	–	–
Cds結晶に (Cds crystal)	107	106	1	1	2
増幅器は (amplifier)	108	107	1	1	1
第三図に (figure 3)	109	–	–	–	–
装置に (equipment)	110	–	–	–	–
光を (light)	111	107	4	1	1
加速電圧を (accelerated voltage)	112	110	1	1	2
減衰量が (attenuation quantity)	113	112	1	1	1
結晶に (crystal)	114	107	7	1	2
電圧の (voltage)	115	114	1	1	1
出力を (output)	116	115	1	1	1
結晶に (crystal)	117	114	3	1	1
減衰が (attenuation)	118	113	5	1	1
加速電圧を (accelerated voltage)	119	117	2	1	1
以上の (above mentioned)	120	–	–	–	–
増幅器は (amplifier)	121	120	1	1	1
進行波管が (traveling –wave tube)	122	101	21	2	1
誕生 (birth)	123	122	1	1	2

No te: 1) English equivalents are shown in ( ); 2) underlined Hiragana sequences are postpositional particles, denoting topic, case, contrast, etc; 3) hyphen means that J-th sentence was not connected with any preceeding sentence by lexical equivalence.

Symbols: I, J -- sequence numbers of the dependent sentence and governor sentence respectiveiy; D -- lexical parallelism indicator distance; w -- sequence number of the lexical indicator from the beginning sentence; t -- type of lexical repetition, 1, 2, 3 - identical, partial, lexico-semantic respectiveiy.

かれている状況をも示している．一方，w＝5以上のところに語句反復インジケータがあることはきわめてまれである．

この結果，文頭近傍に古い情報（主題），文の後部に新しい情報（解説）を与える構造をよく示していると言えよう．特に日本語においては文法的な主語でなく，先行文中で述べられた既知の情報を以下の文での文頭に置く構造が語順の自由度からも許されるところに特徴がある．

文間結合距離

テキストAについて，人間の手で選択を行なった語句反復インジケータの表を表3に，その文間結合関係のグラフを図1で示した．なお，テキストAは，「超音波増幅」という解説文で4節，123文からなっている．

図1では，理想的な結合距離としてほぼ対角線（距離1の線）で示している．これは，テキストの先頭の文を除いてすべての文がそれぞれ連接する次の文を支配していることを意味する．このテキストではほぼこの線の近傍に分布している．

表3では語句反復のインジケータが著者の主題の展開にそって出現していることを示している．既にセブボが指摘しているように，重要なインジケータはテキスト，節といった範囲を示すような大きな距離（D）を持っていることを指摘している．Dの大きな語句反復インジケータは，その著者が別の主題（topic）について解説してのち，再び戻ってくる主題を表わすとも考えられる．

5．自動抄録等への応用

テキストAのに於ける文間結合距離Dが10以上の語句反復インジケータとその距離Dを示したものが図2である．これらのインジケータは節内で分布している．

インジケータ「超音波」は，2番目の文から40番目の文を範囲とする第2パラグラフ内で9番目の文から24番目の文の15文にまたがっており，「進行波管」は，85文から123文の第4パラグラフ内，100～122文の22文にまたがっている．更に，「進行波増幅」は，第3パラグラフを完全にカバーしており，そのパラグラフの先頭の41文から67文の26文にと，同時に67文から次のパラグラフの先頭の85文の18文にまたがっている．

つまり，これらの3個のインジケータがテキストを3っのパラグラフに分割しているともいえる．

さらに，これらのインジケータは，このテキストの内容を適格に表わしているもいる．なぜならば，各インジケータはそのパラグラフ名と部分一致した語句からなっている．すなわち，第1パラグラフ名が「概説」（第1文），第2パラグラフ名が「超音波とは」，第3パラグラフ名が「マイクロ波と進行波管」，第4パラグラフ名が「超音波と進行波管」からなっている．

この結果，Dが大きな語句反復イ

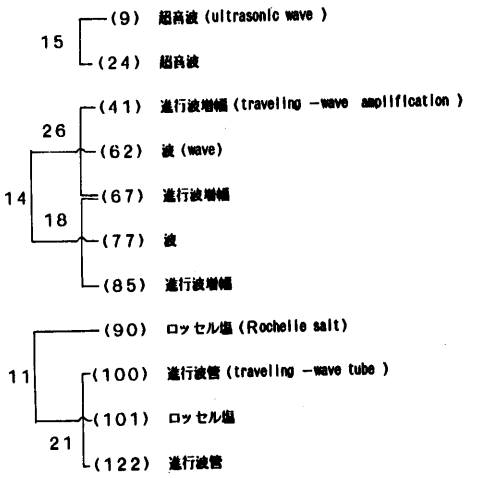

Fig. 2 Distribution of long distance indicator (D>10)

Note : numbers in ( ) correspond to the sequence numbers of the sentences, the numbers on the lines to the distances.

ンジケータは，自動主題分析，自動抄録等でのkeyword として，役立つであろう．

６．　むすびに

１）　文間関係において，語句反復は構文関係以上に重要な役割を担っている．

２）　文頭（ｗ＝１）の語句が語句反復インジケータの候補となる確率が最も高い．このことは，日本語文では，文の先頭で先行文からの既知の情報を受け取る傾向が特に顕著である．

３）　大きな結合距離を持つ語句反復インジケータは，そのテキストあるいはパラグラフの内容を表している．又，テキスト及びパラグラフの冒頭あるいは末尾の文の語句反復インジケータは，大きなＤをもつ傾向がある．

今後の問題点

１）部分一致，意味的な一致の判定法及び一個の文で複数個の語句反復インジケータが見出されたとき最適なインジケータの選択法が計算機による自動解析の問題点である．

２）文間結合関係において，構文的結合関係は重要な役割を担っているが，そこにはメタ言語的な関係と目的言語的な（内容を表わす）関係を表わす２種類の言語が存在し，それが形式的に表面に現れていない点から，分析が非常に困難である

３）自動主題分析，自動抄録におけるkeyword 及びkey sentenceの抽出には，Ｄの他に頻度が問題となる．

４）今後，日本語のテキスト，辞書，文法が質量共に完備されると，定量的な測定が可能となる．

参考文献

１）岡本哲也，「日本語テキストの構造分析」，計量国語学，ｐ１～１１，第６２号，１９７２．

２）岡本哲也，「語句反復形式によるテキストの構造と主題の決定」，電気通信大学学報，ｐ１７７～１９０，２４巻，１号，１９７３．

３）坂本義行，岡本哲也，谷津直和，「テキスト構造の理解モデル－語句の反復形式による分析－」，第１０回情報科学研究集会論文集，ｐ５５～６４，１９７３．

４）Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., “A Grammar of Contemporary English”, Longman, London, １９７２．

５）林四郎，「文章おける始発，承前，転換について」，計量国語学，３９号～１８，４１号１～１７，４２号１～１７，４３号／４４号９～２５，１９６７～６８．

６）Sevbo, I. N., ”Struktura svjaznovo teksta i automaizatsija”, Nayka, M., １９６９．

７）牧野成一，「くりかえしの文法」大修館書店，東京，１９８０．