COLING 82, J. Horecký (ed.)
North-Holland Publishing Company

LEXICAL PARALLELISM IN TEXT STRUCTURE DETERMINATION AND CONTENT ANALYSIS

Yoshiyuki Sakamoto
Electrotechnical Laboratory
Tsukuba, Japan

Tetsuya Okamoto
University of Electrocommunications
Tokyo, Japan

ABSTRACT

In this paper the problem is discussed about the text structure determination and content analysis by lexical parallelism, or the repetition of lexical items. Intersentential relations are determined through the identical, partly identical or lexico-semantic repetition in Japanese scientific texts. Lexical parallelism ratio and lexical parallelism indicator distance are obtained on computer and by hand. And the application of the characteristics to automatic content analysis is dicsussed.

1. INTRODUCTION

Lexical parallelism, that is, the repetition of lexical items, is an important device for indicating the sentence connections in a text(discouse). The recurrent lexical items, or lexical equivalents need not have the same syntactic function or parts of speech in the two sentences in which they occur. They may be identical in form and in meaning, or they may be related by lexico-semantic relationship, such as synonymy, hyponymy, antonymy. In a special case they may be partly identical both in form and in meaning, as in 超音波 (ultrasonic wave), 音波(sound wave) and 波(sound).

Another device for indicating the sentence connections is a syntactic device, such as substitutes, logical connecters, time and place relaters and structural parallelism [1]. For example, in Japanese substitutes--- これ/この(this), ここ(here), われわれ(we/our), それ(it), time relaters--- 次に (next), 以上の (above mentioned), and logical connecters--- および (and), または (or), 第二に (secondly) belong to this device.

Sevbo studied lexical parallelism in normalized text, where substitutes were replaced by their lexical equivalents and complex sentences were decomposed into successive simple sentences(clauses).

She traced the repetition patterns of lexical items in Subject/Predicate oppossition. She assumes here that the syntactic subject or its dependent, direct or indirect, corresponds to "Subject(old information) of elementary thought" and the syntactic predicate or its dependent to "Predicate(new information) of elementary thought"[2].

In Japanese, sentence components occur in any positions before predicate and old information or topic is placed, as a rule, at/near the beginning of a sentence[3]. In the following discussion we analyze the repetition of lexical items in an unnormalized text without regard to their syntactic functions, parts of speech and topic/comment distinctions, assuming that the lexical equivalents at/near the beginning of the sentences function as the keywords in indicating the sentence connections and the contents of a text.

Nouns do not inflect and most verbs and adjectives have the unchanging stems and inflectional suffixes in Japanese. The important concepts and technical terms (noun, verb or

adjective stems) are written in Kanji (Chinese ideographs) or Katakana(square Japanese syllabary). Katakana is used to transcribe foreign technical terms. Hiragana(Japanese cursive syllabary), on the other hand, is used to write post-positional particles and suffixes, denoting case, topic, mood, tense aspect etc. In view of these facts we define lexical items as a word or phrase in Kanji and Katakana.

We have studied lexical parallelisms in a short tale[4], in technical and scientific texts[5,6], based upon Sevbo's approach. The purpose of the present paper is to obtain the characteristics of lexical parallelism in Japanese technical and scientific texts and to explore the possibilities of utilizing these characteristics for automatic content analysis.

Five text samples are used for experiment and discussion. They are the essays on " Ultrasonic amplification"(Text A), "Brain and automaton"(Text B), "Petrochemical industry"(Text C), "Chemical industry in Japan"(Text D) and "Between organism and inanimate matter"(Text E).

2. LEXICAL PARALLELISM RATIO

The sentence connection of type t in position w is determined between the given j-th sentence Sj and the i-th sentence Si(i < j), if and only if Si is the nearest preceding sentence which contains the lexical item, lexically equivalent to the w-th lexical item from the beginning of the given sentence Sj through the type t repetition(t = 1,2,3; w = 1,2,3,4,5).

The repetitions of type 1,2,3 correspond to the identical, partly identical, lexico-semantic repetitions, respectively.

The lexical equivalents in Sj and Si are called lexical parallelism indicators, and Sj is called a dependent on Si.

Lexical parallelism ratio of type t in position w is defined as follows;

$$\alpha_w^t = (n / N - 1) * 100$$

where n is the number of the determined connections in a text; N-1 is the determinable maximum number of the sentence connections in a text, N being the total number of the sentences in the text; t is type of lexical repetition and w is the position, i.e. the sequence number from the beginning of the sentence.

The experiments were carried out to obtain the characteristics of the lexical parallelism in sample texts on computer and by hand.

In computer experiment lexical items, i.e. the sequence in Kanji or Katakana, were identified and segmented by machine character codes without syntactic and morphological analysis. Then the sentence connections of type 1(identical repetition) are determined in each position and lexical parallelism ratios are obtained(Table 1). On the same samples the optimal sentence connections are determined manually and the lexical parallelism ratios were calculated(Table 2). Except for Text E, the totals of the ratios amount to 72-83%(cf. Table 2) and in computer experiment the ratios of type 1 in the initial position amount to 57-68%(cf. Table 1). And moreover, the initial lexical items(w=1) show the maxima in most samples in Table 1 and by far the highest value in all samples in Table 2, and they decrease with increasing w in Table 2. It is clear from the results that lexical parallelism plays an important role in the intersentential dependency and lexical items at the beginning of the sentences are the most relevant lexical parallelism indicators.

3. LEXICAL PARALLELISM INDICATOR DISTANCE

As an example, intersentential dependency determined manually in Text A, which is the essay on "Ultrasonic amplification" with 123 sentences in four paragraphs, is shown in Table 3 and Figure 1. The lexical parallelism indicator distances are shown as well. Lexical parallelism indicator distance is defined as follows:

$$D_{w,j}^{t} = j - i$$

where D is lexical parallelism indicator distance; t is type of

lexical repetition; w is position of the lexical indicator; i and j are sequence numbers of the governor sentence and dependent sentence respectively.

The distance is supposed to represent the semantic extent of the lexical parallelism indicators, or better the concepts referred by them.

In Figure 1 a diagonal unit distance line indicates the hypothetical situation, where every sentence depends on the immediately preceding sentence. Data show a tendency to distribute near this line in all samples.

Lexical parallelism indicators show the progress of the author's thought in the text in Table 3. Sevbo pointed out the significance of the indicators with large D in indicating the contents of paragraphs and texts. The lexical items with large D are supposed to be the important topics, to which the author of the text returnes after commenting on another topics. In the example the items with large D(D>10) were shown in Figure 2.

These indicators are distributed among paragraphs. For example, the indicator 超音波(ultrasonic wave) extends over 15 sentences(from 9th to 24th) within paragraph 2, which ranges from 2nd to 40th sentence, and the indicator 進行波管 (traveling-wave tube) extends over 22 sentences(100th-122nd) within paragraph 4(85th-123rd) as well. The indicator 進行波増幅 (traveling-wave amplification) covers paragraph 3 completely, ranging from the 41th sentence, or the first sentence of the paragraph, through the 67th sentence to 85th sentence, or the first sentence of the next paragraph. In short, these indicators divide the text into the three paragraphs.

In addition, they reflect appropriately the contents of paragraphs in the sample text, as suggested by the fact that they are partly identical with the following paragraph names: "Introduction"(paragraph 1), "What is the ultrasonic wave?"(paragraph 2), "Microwave and traveling-wave tube"(paragraph 3) and "Ultrasonic wave and traveling-wave amplification"(paragraph 4).

These data suggest that the indicator with large D may be useful as keywords to the contents of a text.

4. CONCLUSION

Lexical parallelism plays an important role in the intersentential dependency, or text structure and lexical items at the beginning of the sentences are the most relevant lexical parallelism indicators.

The initial lexical parallelism indicators with long lexical parallelism indicator distances reflect the contents of paragraphs and may be useful keywords in information retrieval.

The partly identical repetition and lexico-semantic repetition through the lexical items at/near the beginning of the sentence,firstly, intersentential dependency by syntactic device, secondly, the recognition of topic/comment opposition in the sentence, thirdly, and lastly, the application to automatic keyword or key-sentence extraction in content analysis depend on the future researches.

REFERENCE

[1] Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J., A grammar of contemorary English (Longman, London, 1972).

[2] Sevbo, I.N., Struktura svjaznovo teksta i avtomatizatsija (Nayka, M., 1969).

[3] Makino, S., Grammar of repetition (Taisyukan, Tokyo, 1980).

[4] Okamoto, T., Text structure determination and content analysis by lexical parallelism, Proceeding of the Univ. of Electro-Communications, vol.24(1973), no. 1, 177-190.

[5] Okamoto, T., Structure analysis of Japanese text, Mathematical linguistics, No.62(1972), 1-11.

[6] Sakamoto, Y., Okamoto, T., Yatsu, N., Text structure and a model of discourse understanding by lexical parallelism, Proceeding of the 10th annual meeting on information science and technology, (1973), 55-64.

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.

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.

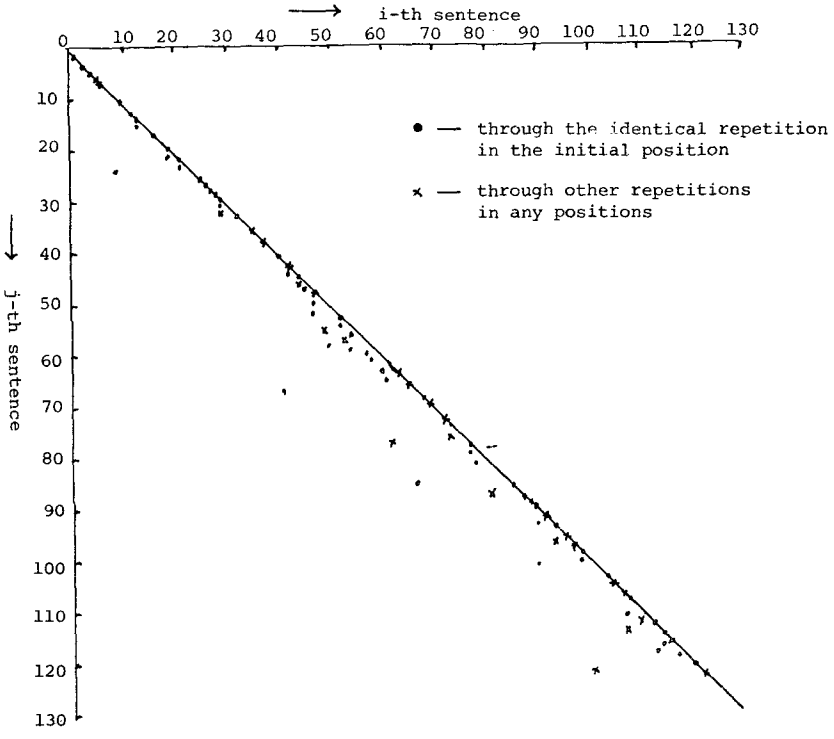

Figure 1 Lexico-semantic intersentential dependency graph in sample text A

Table 3 Lexico-semantic intersetential dependency in sample text A

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 (whistle)	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 (radar)	29	28	1	1	1
供に (together with)	30	29	1	2	1
レーダーは (radar)	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	–	–	–	–
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
Vs	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	100	21	2	1
誕生 (birth)	123	122	1	1	2

Note: 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 respectively; 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 respectively.

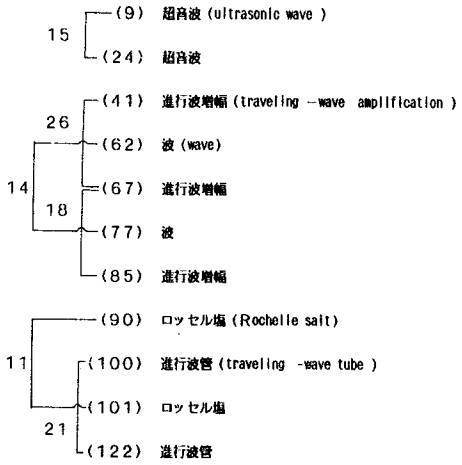

Fig. 2 Distribution of long distance indicator (D>10)

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