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 discussed.

1. INTRODUCTION

Lexical parallelism, that is, the repetition of lexical items, is an important device for indicating the sentence connections in a text (discourse). 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 (ultrasonic wave), (sound wave) and (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-- (this), (here), (we/our), (it), time relaters-- (next), (above mentioned), and logical connecters-- (and), (or), (secondly) belong to this device.

Isebo 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 opposition. 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), “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 \( S_j \) and the \( i \)-th sentence \( S_i \) (\( 1 < j \)), if and only if \( S_i \) 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 \( S_j \) 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, lexicosemantic repetitions, respectively.

The lexical equivalents in \( S_j \) and \( S_i \) are called lexical parallelism indicators, and \( S_j \) is called a dependent on \( S_i \).

Lexical parallelism ratio of type \( t \) in position \( w \) is defined as follows:

\[
\alpha^t_w = \left( \frac{n}{N-1} \right) \times 100
\]

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:

\[
D = j - i
\]

where \( D \) is lexical parallelism indicator distance; \( t \) is type of
LEXICAL PARALLELISM IN TEXT STRUCTURE DETERMINATION

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 1. Sebastian 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 returns after commenting on another topic. 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 "ultrasonic wave" extends over 15 sentences (from 9th to 24th) within paragraph 2, which ranges from 2nd to 40th sentence, and the indicator "traveling-wave tube" extends over 22 sentences (100th-122nd) within paragraph 4 (85th-123rd) as well. The indicator "traveling-wave amplification" covers paragraph 3 completely, ranging from the 41st 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, these indicators reflect 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 lexicosemantic 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 contemporary English (Longman, London, 1972).
[2] Sebastian, 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(%)

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

Note: T - sample texts, w - sequence numbers of indicators, values in() maximum number of intersentential connections.

Table 2 Lexical parallelism ratios determined by hand(%)

| A  | 60.7 | 6.6 | 3.2 | 0.8 | 0.8 |
|---|---|---|---|---|---|
|   | (74) | (8) | (4) | (1) | (1) |
| B  | 68.9 | 4.7 | 1.9 | 0.9 | 0.9 |
|   | (71) | (10) | (2) | (1) | (1) |
| C  | 50.7 | 8.7 | 13.0 | 2.9 | 0 |
|   | (35) | (6) | (9) | (2) | (0) |
| D  | 54.9 | 13.9 | 2.4 | 1.6 | 0 |
|   | (67) | (17) | (3) | (2) | (0) |
| E  | 29.2 | 5.6 | 2.2 | 1.1 | 0 |
|   | (26) | (5) | (2) | (1) | (0) |

Note: N-I --- the determinable numbers of determined sentence connections.

Figure 1 Lexico-semantic intersentential dependency graph in sample text A
## Lexical Parallelism in Text Structure Determination

### Table 3: Lexico-semantic intersetential dependency in sample text A

| Indicator | J | D | t |
|-----------|---|---|---|
| (sound)   | 1 | - | - |
| (ultra~wave) | 2 | 1 | 1 |
| (ultrasonic) | 3 | - | - |
| (ultrasonic wave) | 4 | 1 | 1 |
| (ultrasonic wave) | 5 | 4 | 1 |
| (ultrasonic wave) | 6 | 5 | 1 |
| (ultrasonic wave) | 7 | 6 | 1 |
| (ultrasonic wave) | 8 | - | - |
| (ultrasound) | 9 | - | - |
| (ultrasound) | 10 | - | - |
| (ultrasound) | 11 | 10 | 1 |
| (ultrasound) | 12 | - | - |
| (ultrasound) | 13 | 12 | 1 |
| (ultrasound) | 14 | 13 | 1 |
| (ultrasound) | 15 | 13 | 2 |
| (ultrasound) | 16 | - | - |
| (ultrasound) | 17 | 16 | 1 |
| (ultrasound) | 18 | - | - |
| (ultrasound) | 19 | - | - |
| (ultrasound) | 20 | 19 | 1 |
| (ultrasound) | 21 | 19 | 2 |
| (ultrasound) | 22 | 19 | 3 |
| (ultrasound) | 23 | 20 | 1 |
| (ultrasound) | 24 | 20 | 2 |
| (ultrasound) | 25 | - | - |
| (ultrasound) | 26 | 25 | 1 |
| (ultrasound) | 27 | 25 | 2 |
| (ultrasound) | 28 | 25 | 3 |
| (ultrasound) | 29 | 26 | 1 |
| (ultrasound) | 30 | 26 | 2 |
| (ultrasound) | 31 | 26 | 3 |
| (ultrasound) | 32 | 26 | 4 |
| (ultrasound) | 33 | 27 | 1 |
| (ultrasound) | 34 | - | - |
| (ultrasound) | 35 | - | - |
| (ultrasound) | 36 | 35 | 1 |
| (ultrasound) | 37 | - | - |
| (ultrasound) | 38 | 37 | 1 |
| (ultrasound) | 39 | - | - |
| (ultrasound) | 40 | - | - |
| (ultrasound) | 41 | 40 | 1 |
| (ultrasound) | 42 | - | - |
| (ultrasound) | 43 | 42 | 1 |
| (ultrasound) | 44 | 42 | 2 |
| (ultrasound) | 45 | 42 | 3 |
| (ultrasound) | 46 | 43 | 1 |
| (ultrasound) | 47 | 43 | 2 |
| (ultrasound) | 48 | 43 | 3 |
| (ultrasound) | 49 | - | - |
| (ultrasound) | 50 | 47 | 1 |
| (ultrasound) | 51 | - | - |
| (ultrasound) | 52 | 47 | 1 |
| (ultrasound) | 53 | 47 | 2 |
| (ultrasound) | 54 | 47 | 3 |
| (ultrasound) | 55 | 48 | 1 |
| (ultrasound) | 56 | 48 | 2 |
| (ultrasound) | 57 | 48 | 3 |
| (ultrasound) | 58 | 50 | 1 |
| (ultrasound) | 59 | 50 | 2 |
| (ultrasound) | 60 | 50 | 3 |
| (ultrasound) | 61 | 50 | 4 |
| (ultrasound) | 62 | - | - |

**Note:** 1) English equivalents are shown in ( ). 2) Dotted lines show intersetential dependency. 3) Partial parallelism indicator: 1/1, 2/1, 3/1, 4/1, 5/1, 6/1, 7/1, 8/1, 9/1, 10/1, 11/1, 12/1, 13/1, 14/1, 15/1, 16/1, 17/1, 18/1, 19/1, 20/1, 21/1, 22/1, 23/1, 24/1, 25/1, 26/1, 27/1, 28/1, 29/1, 30/1, 31/1, 32/1, 33/1, 34/1, 35/1, 36/1, 37/1, 38/1, 39/1, 40/1, 41/1, 42/1, 43/1, 44/1, 45/1, 46/1, 47/1, 48/1, 49/1, 50/1, 51/1, 52/1, 53/1, 54/1, 55/1, 56/1, 57/1, 58/1, 59/1, 60/1, 61/1, 62/1.
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.