
A Discourse Structure Analysis of Technical Japanese Texts and Its Implementation on the WWW*

Jie Chi Yang & Kanji Akahori

Department of Human System Science, Tokyo Institute of Technology, Japan

ABSTRACT

This paper deals with a discourse structure analysis of technical Japanese texts for developing a Japanese writing Computer Assisted Language Learning (CALL) system whose goal is to assist students in learning to write technical Japanese texts. To analyze discourse structures of technical Japanese texts, cohesive expressions are used as cue words. The rules for analyzing texts are based on micro-level and macro-level information, namely cohesive expressions and headlines. A CALL system for helping foreigners to learn to write technical Japanese texts is under development using natural language processing (NLP) techniques. This paper describes a completed part of the work, which is a CALL system that can be used for automatically detecting headlines and cohesive expressions of technical Japanese texts on any World Wide Web (WWW) browser. This approach can be considered as a new means of language learning for the future. Furthermore, a system evaluation is conducted to evaluate the performance of the system. The results of this evaluation show that the system obtained a high degree of accuracy on extraction of cohesive expressions and headlines by using the revised rules set proposed in this study.

1. INTRODUCTION

The aim of this research was to construct a Japanese learning environment for foreign students on the Internet. For students in science and technology universities like the Tokyo Institute of Technology, there is little time for

* The authors would like to thank Dr. Kikuko Nishina (Professor in the International Student Center, Tokyo Institute of Technology) for her kind help with the analysis of technical Japanese texts. The authors would also like to thank Dr. Billy V. Koen (Visiting Professor, CRADLE, Tokyo Institute of Technology) and Ms. Cheong Meng-Mei (Tokyo Institute of Technology) for their suggestions and comments on an earlier version of this paper. This study is funded by the Scientific Research Grant from the Ministry of Education of Japan (Subject No. B(2)10558019).

Correspondence: Jie Chi Yang & Kangi Akahori, Tokyo Institute of Technology, Alcoholic Lab., 'CRADLE', Dept. of Human Science, 2–12, 1, O-okayama, Meguro-ku, Tokyo 152, Japan.
E-mail: yang@cradle.titech.ac.jp.

Manuscript submitted: September, 1999.
Accepted for publication: January, 2000.

enrolling in a regular Japanese language course, which involves spending a lot of time on experiments, studies and research, etc. The Internet environment is provided in almost all laboratories and can become an excellent virtual learning environment if there is a Japanese learning system which can be accessed on the Internet anytime and anywhere. The Internet has stimulated many new approaches to language instruction and learning, and it provides a great opportunity to learn one of the most important skills, writing. This is especially true for students in the science and engineering fields who need to write technical texts.

However, almost all CALL systems are concerned with learning how to improve one's reading and listening skills. Few systems are concerned with writing because of the difficulty of implementing an analysis of sentences typed by students who need to learn to phrase their own sentences freely without following any predefined rules. More and more researchers, therefore, use natural language processing (NLP) techniques to analyze learners' typed sentences (Holland & Kaplan, 1995; Loritz, 1992). Recently, NLP techniques designed for use with CALL have attracted special attention (see, for example, Nagata, 1996; Nerbonne et al., 1998, etc.), as this is expected to help improve writing skills. Yang and Akahori (1997, 1998a) developed a Japanese writing CALL system using NLP techniques which can be used for learning and producing the Japanese passive voice on the World Wide Web (WWW). Comparison of two Web-based CALL systems showed that the method of 'free input' and 'feedback corresponding to learners' typed sentences' is better than the method of 'multiple choice' and 'feedback that only displays the correct answer' (Yang & Akahori, 1999). Furthermore, an evaluation of the learning histories of the subjects who have actually used the system through the Internet shows that the system obtained a high degree of accuracy and instructional effectiveness (Yang & Akahori, 1998a). These results demonstrate the effectiveness of the CALL system for writing, using NLP techniques on the Internet.

Having sufficient vocabulary and grammatical knowledge is important when learning a foreign language. However, although vocabulary and grammatical rules are provided for correct sentence building in a foreign language, this knowledge alone is not enough. Being able to form correct sentences is by no means sufficient when it comes to expressing complex thoughts. The major problem for most foreigners learning Japanese is, apart from the writing system, the building of sentences: that is, knowing the corresponding words, the postfixes signalling the word's function (*de*, *ni*, etc.) and the position of the words (verbs final form). It is of paramount importance to learn how to struc-

ture one's thoughts: i.e., how to make an outline, how to signal the relative importance of a piece of information, and how it relates to the whole. Therefore, in order to write or to comprehend a structured sentence, it is necessary to learn how to associate sentences, in addition to having a good command of vocabulary and grammar. The connection between sentences can be described as a conjunction of adjacent sentences, which is an important criterion for writing a good text as per research in *cohesion or discourse structure* (Chan & T'sou, 1998; Halliday & Hasan, 1976; Kuno, 1978; Mann, 1984; Sugihara, 1994; Zadrozny & Jensen, 1991). Unfortunately, discourse structure is not amenable to single-sentence grammatical analysis, because there are no 'discourse grammars' (Hovy & Scott, 1996).

Since there are few practical CALL systems that use discourse analysis, the goal of this study is to develop such a system for helping learners to write technical Japanese texts. This paper focuses on discourse analysis of technical Japanese texts and its implementation on the WWW. Section 2 describes previous related works on discourse structure analysis and the methods used in this study. The rules for analyzing texts are based on micro-level and macro-level information, namely cohesive expressions and headlines. This CALL system automatically detects headlines and cohesive expressions in technical Japanese texts using NLP techniques. The implementation of the system using NLP techniques is summarized in Section 3, and Section 4 describes a system evaluation that evaluates the performance of the system using the set of rules proposed in this study.

2. METHODS FOR DISCOURSE STRUCTURE ANALYSIS OF TECHNICAL JAPANESE TEXTS

Many methods concerning the analysis of discourse structure have been proposed in previous related works. Mann and Thompson's (1987, 1988) rhetorical structure theory (RST) is an influential theory of text structure that is being extended to serve as a theoretical basis for computational text planning. RST postulates that a set of about 25 relations suffices to represent the relations that hold within normal English texts. Most relations have a cue word or phrase which informs the listener how to relate the adjacent clauses. RST can be applied to a computational model. There have been attempts at text generation using RST for the implementation of a prototype of the theory (Hovy, 1993; Moore & Paris, 1994).

Cue words are also widely used in the identification of rhetorical relations among portions of a text (Hobbs, 1979; Litman, 1994; Reichman, 1985). Hobbs claims that coherence in conversations and in texts can be partially characterized by a set of coherence relations, which are classified into four categories. Hovy (1993) collected and taxonomized the discourse segment relations; this set of relations contains three taxonomies of approximately 120 relations. Hirschberg and Litman (1993) also summarize the proposed meanings of items classed as cue words in six computational and linguistic treatments.

In most of these earlier works, emphasis was put on the knowledge that is necessary for recognizing discourse structure. The problem of inference based on that knowledge was also emphasized. However, this does not mean that knowledge can be constructed easily from information available on computers. Constructing common knowledge to implement a practical system is often beyond the capabilities of current NLP techniques. Kurohashi and Nagao (1994) proposed an automatic method for detecting discourse structure by checking surface information in text sentences. The information included 'clue expressions', 'occurrence of identical/synonymous words/phrases', and 'similarity between two sentences'. Their results indicate that, in the case of technical Japanese texts, considerable portions of discourse structure can be identified by incorporating the three types of surface information. Cue words often explicitly appear in the surface expressions of technical Japanese texts. Thus, it seems important and necessary to use these explicit cue words to structure one's thoughts in technical Japanese. Foreign learners especially may find it is easier to convey their thoughts using explicit cue words because these can be treated as an indicator of a discourse.

Accordingly, the authors took a similar approach to Kurohashi and Nagao (1994), namely using surface information in texts. The rules for analyzing texts are based on micro-level (cohesive expressions) and macro-level (headlines) information. These are described below.

2.1. The classification of cohesive expressions (cue words)

The selected texts used in this study are 'text sentences of technical Japanese'. This study empirically clarifies the discourse structure of technical Japanese texts by considering the findings of some related works (Hamada et al., 1997; Hinata & Hibiya, 1988; Kitao & Kitao, 1992; Nagara & Chino, 1989; Nishina, 1997; Yamazaki et al., 1992; Yokobayashi & Shimomura, 1988). To examine the discourse structure of technical Japanese texts, the classification of basic expressions by Yamazaki et al. (1992) is adopted in this study. The reason for

this is that their classification covers most of the elements of technical Japanese texts. Based on their findings, the authors have classified cohesive expressions into 15 categories. Each category is classified further into sub-categories (in total, 37). The total number of expressions is 82. All of the expressions are converted into regular expressions to make the rules. In all, 654 distinctions in the regular expressions were extracted from the 15 categories of cohesive expressions. These formed 654 original rules, which are used in the process of analysis.

Table 1. Categories of Cohesive Expressions.

Categories	Sub-category and examples of cohesive expression ¹	Number of rules
1. Comparison	1. Comparison between two substances. <i>A yori B no hou ga . . .</i> (B is . . . than A) 2. Comparison among three or more substances. <i>A, B, C no naka dewa, A ga mottomo . . .</i> (A is the . . . among A, B, and C)	40
2. Contrast	1. Qualitative contrast between two substances. <i>A . . . ni taishite, B . . .</i> (B is . . . while A is . . .) 2. Explanation of different aspects of one substance. <i>A . . . Hanmen, B . . .</i> (A is . . . On the other hand, B is . . .) 3. Explanation of the difference between two substances. <i>A wa B to kotonari . . .</i> (A differs from B . . .)	32
3. Analogy	1. Explanation when two substances belong to one classification. <i>A to B wa tomoni . . .</i> (A and B are both . . .) 2. Stating two substances are identical. <i>A wa B to hitoshii</i> (A is equal to B)	14
4. Cause and reason	1. First describing a cause or reason and then describing the result. <i>A . . . tame B . . .</i> (B . . . because A . . .) 2. First describing the result and then describing the cause or reason. <i>A . . . Korewa B . . . tame dearu</i> (A . . . The reason for this is that B . . .)	19
5. Basis	1. Description of phenomena or relationships based on certain facts or phenomena. <i>A . . . kara B . . . koto ga wakaru</i> (It is obvious that B . . . from A . . .) 2. Expressing judgement using figures, tables and formulas. <i>Zu 1 ni shimesu youni . . .</i> (. . . as shown in Figure 1)	90

Table 1 continues

1. A, B, C, D, E and F indicate a word or a phrase.

Table 1. (*cont.*)

Categories	Sub-category and examples of cohesive expression	Number of rules
6. Composition and enumeration	1. Describing the complete composition of something. <i>A wa B to C kara naru</i> (A consists of B and C) 2. Describing the general composition of something. <i>A wa B to C niyori kouseisareru</i> (A is composed of B and C) 3. Expressing one or the other. <i>A wa B aruiwa C . . .</i> (A is B or C . . .)	131
7. Presentation	1. Expressions used when employing tables, figures and formulas. <i>A . . . wo zu 1 ni shimesu</i> (Figure 1 shows A . . .) 2. Explanation of contents of tables, figures and formulas. <i>Zu 1 wa A . . . wo shimeshiteiru</i> (Figure 1 indicates A . . .)	32
8. Definition	1. Naming. <i>A wo B to yobu</i> (A is called B) 2. Indicating method of definition. <i>A wo (1) shiki de teigisuru</i> (A is defined by the expression (1)) 3. Use of abbreviations. <i>A . . . ika B . . .</i> (A . . . abbreviation B . . .)	38
9. Classification	1. General classification of objects. <i>A dewa, B wo . . . ni bunruishiteiru</i> (In A, B is classified into . . .) 2. Description of classification to which an object belongs. <i>A wa B ni hukumareru</i> (A is included in B) 3. First classifying and then defining an object. <i>A dewa, B wo C to D ni wakeru. C towo E wo ii, D towo F wo iu.</i> (In A, B is classified into C and D. C is called E and D is called F.)	45
10. Hypothesis and conditions	1. Indicating prerequisite conditions. <i>A wa B to kateisuru</i> (A is assumed as B) 2. First stating hypothesis or conditions and then describing conclusions. <i>A . . . to suru toki, B . . .</i> (B . . . when A . . .) 3. Designating the range of hypothesis or conditions. <i>A no jyouken no moto ni . . .</i> (In the condition of A . . .) 4. Designation of limits. <i>A . . . kagiri, B . . .</i> (B . . . as long as A . . .)	62
11. Change of state	1. Description of change from one state to another. <i>A ni yori, B ga C kara D ni naru</i> (B becomes D from C by A) 2. Description of the state after change omitting original state. <i>A ga B naru</i> (A becomes B)	43

Table 1 continues

Table 1. (*cont.*)

Categories	Sub-category and examples of cohesive expression	Number of rules
12. Process of change	1. Description of the process of change. A <i>ga</i> B- <i>te</i> <i>iku</i> (A is 'B'-ing)	10
	2. Description of the process in which one change causes another change. A <i>nitsurete</i> . . . (. . . as A)	
13. Change with prerequisites	1. Stating a change under a certain hypothesis or conditions + change. A <i>toki</i> , B <i>wa</i> C <i>ni naru</i> (B becomes C when A)	71
	2. Stating cause or reason for the change. A <i>tame</i> , B <i>wa</i> C <i>youni natta</i> (B came to C because A)	
	3. Expressing the fact that no change has occurred. A . . . <i>ga</i> , B <i>wa</i> <i>henka shinakatta</i> (B has not changed even though A . . .)	
	4. Expressing a change which exceeds a certain limit. A <i>wa</i> B <i>wo koeta</i> (A exceeded B)	
14. Means and methods	1. Using nouns. A <i>ni yotte</i> , B . . . (B . . . by A)	22
	2. Using verbs. A <i>wo shiyoushi</i> , B . . . (B . . . by using A)	
15. Selection	1. Selection. A <i>niwa</i> . . . <i>ga aru ga</i> , B <i>dewa</i> C <i>wo motiita</i> (A included . . . C was used in B)	5
Total	82	654

The result of this classification of the cohesive expressions is shown in Table 1. In each sub-category, one cohesive expression is listed as an example.

There are two patterns of rules: one is for 'simple pattern matching' and the other is for 'discourse analysis'. The former, called rule set A, is written as a regular expression form and the latter, called rule set B, is written as a regular expression combined with the result of morpheme analysis and syntax analysis (see Appendix 1 for details of notation used in this paper).

Figure 1 shows an example of expressions in technical Japanese texts and their corresponding rules. The expressions correspond to the sub-category of 'first describing cause or reason and then describing the result', which belongs to category 4, 'cause and reason'. In this example, the corresponding rules can

<p>Exp. 19_1: ... [tame, (koto) ni yori, node, kekka], ... (in one sentence) (because of; on account of; due to; for; as a result)</p> <p>Exp. 19_2: ... [Kono tame Kore yori Kono kekka], ... (in two sentences) (Because of; On account of; Due to; For; As a result)</p> <p>Rule A19_1: (tame (koto)? Niyori node kekka)</p> <p>Rule B19_1: J+(tame:N20 (koto:N18)? Niyori:Par3 node:Par5 (V-ta)kekka:N2 1), J+.</p> <p>Rule A19_2: (Kono tame Kore yori Kono kekka)</p> <p>Rule B19_2: S. (Kono:MN tame:N20 Kore:N17 yori:Par1 Kono:MN kekka:N21), J+.</p> <p>EX19_1: <i>Gousei ga fusokushita kekka, koyuu shindou suu wo takameru koto ga dekinai.</i> (<i>Due to</i> a lack of rigidity, natural frequency cannot be improved.)</p> <p>EX19_2: <i>3 shiki ni yori suisanki ga dounyusareru. Kono kekka, hannou seiseibutsu no mizushinwasei ga masu.</i> (The hydroxyl is introduced by expression 3. <i>As a result</i>, the reactivity increases due to the affinity of oxygen for hydrogen.)</p>
--

Figure 1. Example of descriptions of rules.

be written in two ways. The rules in rule set A (Rule A19_1 and Rule A19_2) show the simple regular expressions corresponding to both expressions (Exp. 19_1 and Exp. 19_2). The rules in rule set B (Rule B19_1 and Rule B19_2) show the regular expressions including the results of morpheme analysis and syntax analysis, which correspond to both expressions respectively (Exp. 19_1 and Exp. 19_2). These rules are written in a more restrictive form to improve the accuracy of discourse structure analysis. For example, if a sentence is applied to Rule A19_1 by the cue word *kekka* ('as a result'), it is then analyzed by the morpheme analysis and syntax analysis and the result will be matched to Rule B19_1. From the description of Rule B19_1, the past form of the verb should appear before the cue word *kekka*, and parts of speech of the word *kekka* should be N21². A limitation like this could be improved in a more precise analysis.

Besides that, there are more than one example (such as EX19_1 and EX19_2) written in Japanese which correspond to each rule. This 'examples corpus' is used for feedback. When a new text is analyzed, the extracted

2. 'N21' means the 21st noun of the parts-of-speech sub-category. The parts-of-speech sub-category is to ensure a more precise analysis. The details of the notation used in this paper are listed in Appendix 1.

sentences are automatically added to each rule in the 'examples corpus' in order to enlarge the size of the corpus.

2.2. Headlines

There are many text books on good writing, which nearly all contain a lot of material concerning the different kinds of categories or conceptual bricks at the discourse level out of which texts are built (see, for example, Hamada et al., 1997; Hayes, 1989; Hayes & Flower, 1980; Kinoshita, 1981; Shimizu, 1959; Sugihara, 1994;). Hayes and Flower (1980) proposed a model of a writing process derived through protocol analysis. They divided the writer's world into three major parts: the task environment, the writer's long-term memory and the writing process. The writing consists of three major sub-processes called *planning*, *sentence generation* and *revision*. Hayes (1989) wrote: 'In writing, planning includes such activities as determining what writing task needs to be done, setting goals, thinking of things to say, and deciding on an order of subtasks . . . The sentence generation process in writing takes the writing plan and puts it into action . . . In revising, writers examine their text and attempt to improve it by determining how well it fits the rules of good form, e.g., grammar, spelling, and clarity, how well it expresses what they wanted to say, and what it suggests to them that might be better to say.' Hamada et al. (1997) classified the discourse structure of Japanese text sentences into three parts: *introduction*, *main discourse* and *conclusion*. Each part is further divided into several parts to form a framework of technical Japanese text. For example, the 'introduction' consists of 'background explanation', 'bring motivation up' and 'indicating the purpose', etc. This framework can be applied to the writing process of almost all technical Japanese texts. However, it is difficult to detect the text structure by just using this framework because the framework is too extensive and the varieties of different formats used by people for building technical texts too numerous.

Accordingly, instead of a predefined framework, headline is used as macro-level information in this study. There are several reasons why the authors decided to use 'headline' instead. First, a well-chosen headline allows the reader to infer the text structure. Second, different formats of texts can be analyzed independently of the texts' style by using the headline. Third, it is easier to understand when the headline is displayed rather than a tree structure because the headline is a part of the original text.

A combination of headlines and cohesive expressions are employed in the implementation of the present system. Section 3 summarizes the implementation of the system.

System Diagram

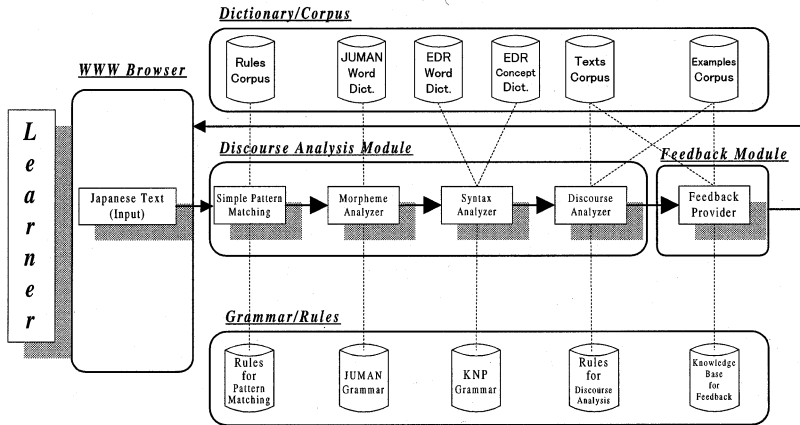


Figure 2. System diagram.

3. IMPLEMENTATION OF THE SYSTEM

As described earlier, the goal of this study is to develop a system for helping learners in the writing of technical Japanese texts. This paper describes a completed part of the work, which is a Web-based CALL system for automatically detecting headlines and cohesive expressions from technical Japanese texts by using NLP techniques.

3.1. Diagram of the system

Figure 2 shows a diagram of the system. The system consists of the interface (i.e., WWW browser), the discourse analysis module, the feedback module, the dictionary/corpus and the grammar/rules.

The discourse analysis module contains 'simple pattern matching', 'morpheme analyzer', 'syntax analyzer', and 'discourse analyzer' components. First, the headlines are extracted and the Japanese texts are divided into sentences using several heuristic rules. Then all the sentences in all texts are matched with all the rules in the 'simple pattern matching' component. The 'rules for pattern matching' component is used during the process of pattern matching. Because of the exclusive character of almost all of the rules, they are written in order of frequency to reduce the running time on the computer.

The frequency of rules is made from the 'rules corpus'. The NLP tools used in the next two components are JUMAN (morpheme analyzer) and KNP (syntax analyzer),³ which were developed by Nagao Laboratory, Kyoto University, Japan. The present system analyzes Japanese text sentences with the morpheme analyzer and syntax analyzer to check the dependency of sentences in the case grammar. Therefore, each cue word in the rules is not only matched against the word itself, but also against the 'parts-of-speech' of the cue word. Only sentences that match the rules written in restrictive form are needed for morpheme analysis and syntax analysis. This takes into consideration the problem of computer running time. The 'rules for discourse analysis' is matched again in restrictive form after the process of syntax analysis. The additional information (parts-of-speech, tense, etc.) is checked to identify the cohesive expressions, especially in the case where one sentence is matched with two or more rules. The electronic dictionary used for syntax analysis (KNP) is EDR.⁴

The feedback module includes the feedback messages provider, knowledge database, texts corpus, examples corpus and a list of all learning histories during the operation of the system. At this stage, a selected text's headlines are shown first to help learners grasp the whole text structure. Secondly, learners can click on the headline of any part of the text that they want to read. Then the original sentences corresponding to the headline are displayed with the extracted cohesive expressions. The cue words in the cohesive expressions are displayed in colour to enable learners to focus on it more easily. Learners can click on any cue words to further find out the cohesive expressions corresponding to the sentences. They can also refer to examples that correspond to the cohesive expressions from the 'examples corpus'.

3.2. Flow of the system

The flow of the system is as follows:

- (1) Learners register themselves to use the system. An ID number is given after registration. The ID number is used to identify the learner because a log of all learning histories is registered during the operation of the system. These registered data are used for learners' feedback.

3. JUMAN and KNP are both free NLP tools which can be found at <ftp://pine.kuee.kyoto-u.ac.jp/pub/>.

4. EDR: Japan Electronic Dictionary Research Institute, Ltd. The EDR is a set of electronic dictionaries which have been developed for advanced natural language processing in a project by the eight biggest Japanese computer makers. The 'word dictionary' and 'concept dictionary' used in this study are two of the EDR set.

- (2) The learning page shows a list of technical Japanese texts. Learners can choose any one text by clicking the hyperlink on the list.
- (3) When learners choose one of the texts from the list, the headline of the selected text is analyzed. The result of the extracted headline is displayed on a new page on the left side of the browser.
- (4) Then learners can click on any headline to read the text sentences which correspond to it. The text sentences of the clicked headline are analyzed by the system. The cohesive expressions are extracted by applying the rules. The result of the extracted cohesive expressions is displayed in colour with the text sentences on the top right side of the browser.
- (5) Learners can click on any cue word in a cohesive expression to read the information about it. The result is displayed on the bottom right side of the browser. If one sentence is matched with two or more rules, all the candidates are also shown here and the final decision is then left to the learner. This is designed to motivate learners to pay attention to sentence cohesion.

Figure 3 shows one screen shot of the system (text source: Kurohashi and Nagao, 1994). This example shows steps (3) to (5) of the flow of the system. As shown in this figure, the headlines of the Japanese text are displayed on the left side of the browser. The headlines show the structure of the text. The headline of Section 1 is selected here. On the right side, the original sentences of Section 1 are displayed in the upper part with the cohesive expressions extracted and a link made. For example, when the cue word 'kotoniyori' (in the first line of the third paragraph) is clicked, the matched cohesive expressions are displayed on the lower right side of the browser.

4. SYSTEM EVALUATION

4.1. Methods

A system evaluation is conducted to evaluate the performance of the system. The purpose of the system evaluation is to assess the accuracy of headline extraction and cohesive expression extraction proposed in this study. Also, the results of the system evaluation are used to revise the rules to improve the performance of the system. Therefore, the system evaluation is designed for text analysis in two stages. The original rules are used in the first stage and the revised rules resulting from the first stage are used in the second stage.

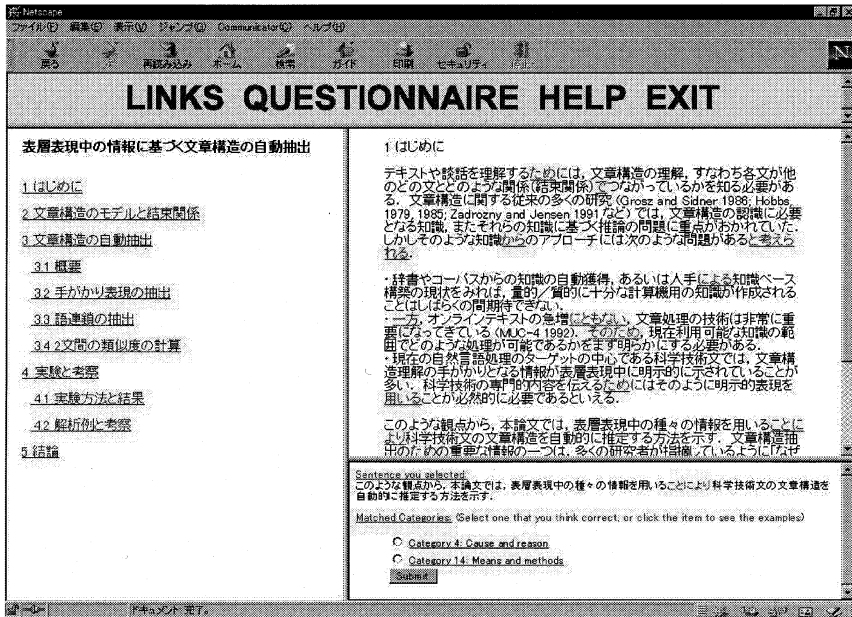


Figure 3. One screen shot of the system.

There are 24 technical Japanese texts used as sources for the analysis. Fifteen texts (a total of 5,738 sentences) are used in the first stage; the other nine texts (a total of 2,420 sentences) are added in the second stage (see Appendix 2 for detailed information about the texts).

The analysis consists of three items in both stages: *headline extraction*, *cohesive expression extraction* and *frequency of the rules*. The flow of the text analysis is as follows: steps (1) to (4) are the first stage; steps (5) to (7) are the second stage.

- (1) Extracting the headlines of the texts.
- (2) Extracting all of the sentences from the texts.
- (3) Matching all sentences to all rules.
- (4) Calculating the frequency of the rules.
- (5) Checking the correctness of the extracted cohesive expressions by hand.
- (6) Revising the rules using the results of steps (4) and (5).
- (7) Repeating the analysis of steps (1) to (4) using the revised rules made in step (6).

Table 2. Accuracy of the Headline Extraction.

Text no.	Number of headlines	Accuracy ratio in Stage 1 ⁵ (%)	Accuracy ratio in Stage 2 (%)
1	11	100.00	100.00
2	10	100.00	100.00
3	19	95.45	100.00
4	18	100.00	100.00
5	13	92.31	100.00
6	15	88.46	100.00
7	16	87.50	100.00
8	13	84.62	100.00
9	11	100.00	100.00
10	7	100.00	100.00
11	10	100.00	100.00
12	10	100.00	100.00
13	28	100.00	100.00
14	21	100.00	100.00
15	10	80.00	80.00
21	12	–	100.00
22	27	–	100.00
23	12	–	100.00
24	14	–	100.00
25	18	–	100.00
26	27	–	100.00
27	11	–	100.00
28	10	–	100.00
29	6	–	100.00
	Average	95.22	99.17

4.2. Results

4.2.1. Headline extraction

Table 2 shows the results of the headline extraction. In Stage 1, the first 15 texts (text numbers 1–15) are used. In Stage 2, the other 9 texts (text numbers 21–29) are added. From the results of Table 2, the accuracy ratio in Stage 1 is 95.22% on average, which could be considered as highly accurate. However, some errors occurred, where sentences were extracted as headlines even though they were not headlines. By looking at the cause of the error, we found that these sentences began mostly with numerical numbers. Therefore, a heuristic rule was added to determine if the sentence is a headline or not when

5. The original rules are used in Stage 1; the revised rules are used in Stage 2.

Table 3. Accuracy of the Cohesive Expression Extraction.

Text no.	Number of rules matched	Accuracy ratio in Stage 1 ⁶ (&)	Accuracy ratio in Stage 2 (%)
1	133	56.39	81.20
2	97	71.13	86.60
3	145	71.03	93.10
4	237	60.76	97.89
5	290	69.66	93.79
6	119	72.27	94.96
7	402	72.89	96.52
8	176	70.45	91.48
9	158	71.52	94.30
10	156	69.23	93.59
11	258	65.12	91.09
12	317	82.33	88.64
13	216	76.39	93.06
14	301	70.10	94.02
15	174	74.14	92.53
21	209	–	92.34
22	139	–	86.33
23	204	–	95.10
24	186	–	94.09
25	134	–	95.52
26	255	–	96.47
27	204	–	95.59
28	155	–	90.97
29	208	–	95.67
Average		70.23	92.70

the sentence begins with a numerical number. The result of the headline extraction using the revised rules in Stage 2 gained an exceedingly high accuracy rate of 99.17%.

4.2.2. Cohesive expression extraction

To make the rules for automatically detecting cohesive expressions, 15 texts are analyzed in the first stage. All the sentences of the 15 texts are matched with all the original rules (rule set A) made from the regular expressions. First, the simple pattern matching is executed. After that, all of the sentences that matched the cue words are analyzed by the morpheme analyzer (JUMAN) and syntax analyzer (KNP). The new rules (rule set B), which include the results of

6. The rules in rule set A are used in Stage 1 for simple pattern matching; the rules in rule set B are used in Stage 2 for detecting cohesive expressions combined with morpheme analysis and syntax analysis.

Table 4. Frequency of the Rules.

Frequency	Rule	Number of matches	The rule's category
1	R56	1014	13. Change with prerequisites
2	R52	541	11. Change of state
3	R19	338	4. Cause and reason
4	R31	237	8. Definition
5	R44	221	10. Hypothesis and conditions
6	R8	213	2. Contrast
7	R19_1	198	4. Cause and reason
8	R3_1	179	1. Comparison
9	R50	162	11. Change of state
10	R7_2	150	2. Contrast

morpheme analysis and syntax analysis, are added after judging the correctness of the extracted cohesive expressions by hand. The process of cohesive expression extraction is executed again using the revised rules. The texts used in the second stage are the same ones used in the first stage plus the other 9 texts.

Table 3 shows the results of the cohesive expression extraction. From Table 3, the accuracy in Stage 1 is 70.23% on average. On the other hand, the accuracy in Stage 2 improved to 92.70% on average. This result shows that using the rules combined with morpheme analysis and syntax analysis gained a higher degree of accuracy than only using the rules of simple pattern matching.

4.2.3. Frequency of the rules

After the cohesive expression extraction, the frequency of rules is calculated. Table 4 shows the result of the top 10 frequency of cohesive expressions used in the 24 texts. The result of 'frequency of the rules' is saved to the 'rules corpus'. The order of frequency is taken as the order of the rules to reduce the running time on the computer. The frequency of the rules may differ in different fields; however, it can be considered an indicator of one characteristic of a technical Japanese text.

4.3. Discussion

The results of the system evaluation show that the system gained a higher accuracy in the analysis in Stage 2 after the rules were revised. The headline extraction shows a very high degree of accuracy (99.17%) when using the revised

rules. The cohesive expression extraction also shows a higher degree of accuracy (92.70%) when using the revised rules combined with morpheme analysis and syntax analysis than only using the rules for simple pattern matching.

However, there is still room for improvement on the cohesive expression extraction. New approaches are needed to conquer the limitations of the computer. The authors emphasize the importance of the conception of ‘self-correction’ stated in Yang and Akahori (1999). ‘Self-correction’ means that the correct answer is not given by the system, and instead learners are requested to come up with the correct answer by themselves. Under these circumstances, the system should be designed to provide more information used in the real world to allow learners to come up with the correct answer. In this study, the conception of ‘self-correction’ is adopted. In the case of cohesive expression extraction, when one sentence is matched with two or more rules, all candidates are listed and the final decision is left to the learner. This is designed to motivate learners to pay more attention to the sentence cohesion. Figure 4 shows an example of a sentence that is matched with two rules. It is an enlarged view of the screen shot showing the lower right side in Figure 3. In Figure 4, the upper part shows the sentences with the results of cohesive expression extraction, whereas the lower part shows the result when learners clicked the cue word ‘*kotoniyori*’. The cue word in this sentence is matched with two rules and the revised rules resulting from the morpheme analysis and syntax analysis. The candidates of the two categories that correspond to the two rules are listed. Learners are requested to select the correct answer by referring to the examples made from the ‘examples corpus’.

The rules are revised using the result of the analysis of the first stage. The result of the analysis in the second stage shows that the revised rules are useful. However, the process of revision of the rules by hand is a really serious matter.

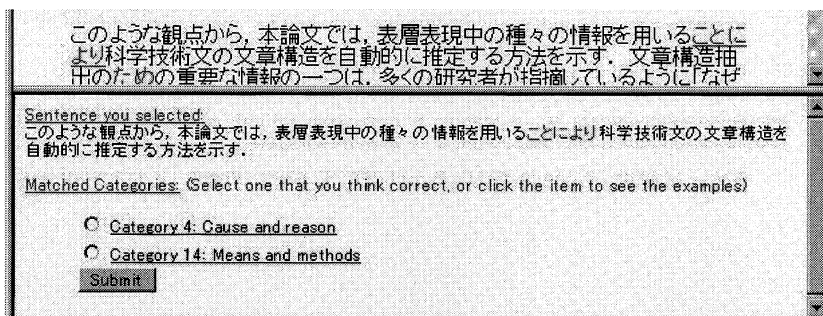


Figure 4. One example of a sentence matched with two rules.

It requires a great deal of effort and time to finish this work. Therefore, the method of revising rules automatically is expected.

5. CONCLUSION

In this paper, the authors deal with the discourse structure analysis of technical Japanese texts and its implementation in a CALL system on the WWW. The cohesive expressions are classified into 15 categories, where 37 sub-categories and 654 rules are made. The rules are used for analyzing headlines and cohesive expressions in technical Japanese texts. The system has been developed using NLP techniques. The present system has the following functions: the ability to detect automatically headlines and cohesive expressions, to display this information, and to request learners to come up with the correct answer themselves. It can also be accessed on the WWW browser. Furthermore, the result of the system evaluation shows that the system obtained a high degree of accuracy on the headline extraction and the cohesive expression extraction using the revised rules set.

The limitations of current NLP techniques have caused some obstacles in the development of NLP-based CALL systems, but this state-of-the-art technology has had a more positive impact on CALL research. As described in Nerbonne et al. (1998): 'Although, indeed, linguistics has not yet been able to encode the entire complexity of natural language, this does not imply that NLP cannot be useful to CALL.' This paper shows the advantages of NLP, with one advantage being its ability to improve the performance of a CALL system. However, more research on humanities is needed. We believe that the success of CALL systems not only relies on new technology but also on the humanities.

The authors plan to improve the present system by employing new rules, enhancing the interface and increasing the size of the 'text corpus'. The authors are also planning to conduct an experimental study to examine the effectiveness of the system.

REFERENCES

- Chan, S.W.K. & T'sou, B.K. (1998) 'Analyzing discourse structure using lexical cohesion: A connectionist tool', in *Proceedings of the 1998 IEEE International Joint Conference on Neural Networks*, pp.657-62.

- Dahlgren, K. (1996) 'Discourse coherence and segmentation', in E.H. Hovy & D.R. Scott (eds) *Computational and Conversational Discourse: Burning Issues—An Interdisciplinary Account*. Berlin: Springer-Verlag, pp.111–38.
- Hajicova, E. (1996) 'The information structure of the sentence and the coherence of discourse', in E.H. Hovy & D.R. Scott (eds) *Computational and Conversational Discourse: Burning Issues—An Interdisciplinary Account*. Berlin: Springer-Verlag, pp.97–107.
- Halliday, M.A.K. & Hasan, R. (1976) *Cohesion in English*. London and New York: Longman.
- Hamada, M., Hirao, T. & Yui, K. (1997) *Paper Workbook for University/Foreign Student*. Tokyo: Kuroshio. [in Japanese]
- Hayes, J.R. (1989) 'Writing as problem solving', in J.R Hayes (eds) *The Complete Problem Solver*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hayes, J.R. & Flower, L.S. (1980) 'Identifying the organization of writing processes', in L. Gregg & E. Steinberg (eds) *Cognitive Processes in Writing: An Interdisciplinary Approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Hinata, S. & Hibiya, J. (1988) in S. Nagara (ed.) *Discourse Structure*, series 16. Tokyo: Aratake. [in Japanese]
- Hirschberg, J. & Litman, D. (1993) 'Empirical studies on the disambiguation of cue phrases', *Computational Linguistics* 19 (3): 501–30.
- Hobbs, J.R. (1979) 'Coherence and coreference', *Cognitive Science* 3: 67–90.
- Holland, V.M. & Kaplan, J.D. (1995) 'Natural language processing techniques in computer assisted language learning: Status and instructional issues', *Instructional Science* 23: 351–80.
- Hovy, E.H. (1993) 'Automated discourse generation using discourse structure relations', *Artificial Intelligence* 63: 341–85.
- Hovy, E.H. & Scott, D.R. (eds) (1996) *Computational and Conversational Discourse: Burning Issues—An Interdisciplinary Account*. Berlin: Springer-Verlag.
- Kinoshita, K. (1981) *The Techniques of Writing for Scientists*. Tokyo: Chuoukouronsya. [in Japanese]
- Kitao, S.K. & Kitao, K. (1992) *Understanding English Paragraphs: Improving Reading and Writing Skills*. Tokyo: Eichosha.
- Kuno, S. (1978) *The Grammar of Discourse*. Tokyo: Daisiyukan. [in Japanese]
- Kurohashi, S. & Nagao, M. (1994) 'Automatic detection of discourse structure by checking surface information in sentences', *Journal of Natural Language Processing* 1 (1): 3–20. [in Japanese]
- Lam, F.S. & Pennington, M.C. (1995) 'The computer vs. the pen: A comparative study of word processing in a Hong Kong secondary classroom', *Computer Assisted Language Learning* 8 (1): 75–92.
- Lin, C.C., Niwa, Y. & Narita, S. (1997) 'Logical structure analysis of book document images using contents information', in *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, pp.1048–54.
- Litman, D.J. (1994) 'Classifying cue phrases in text and speech using machine learning', in *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp.806–13.
- Loritz, D. (1992) 'Generalized transition network parsing for language study: The GPARS system for English, Russian, Japanese and Chinese', *CALICO Journal* 10 (1): 5–22.
- Mann, W.C. (1984) 'Discourse structures for text generation', in *Proceedings of the Tenth COLING*, pp.367–75.

- Mann, W.C. & Thompson, S.A. (1987) 'Rhetorical structure theory: Description and construction of text structures', in G. Kempen (ed.) *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*. Dordrecht: Martinus Nijhoff Publishers, pp.85–95.
- Mann, W.C. & Thompson, S.A. (1988) 'Rhetorical structure theory: Toward a functional theory of text organization', 8 (3): 243–81.
- Maruyama, H. & Watanabe, H. (1987) 'A discourse analysis technique for a natural language interface system', in *Proceedings of the Eleventh Annual International Computer Software & Applications Conference*, pp.578–85.
- Moore, J.D. & Paris, C. (1994) 'Planning text for advisory dialogues: Capturing intentional and rhetorical information', *Computational Linguistics* 19 (4): 651–94.
- Nagara, S. & Chino, N. (1989) in S. Nagara (ed.) *Style*, series 9. Tokyo: Aratake. [in Japanese]
- Nagata, N. (1996) 'Computer vs. workbook instruction in second language acquisition', *CALICO Journal* 14 (1): 53–75.
- Nerbonne, J., Dokter, D. & Smit, P. (1998) 'Morphological processing and computer-assisted language learning', *CALL Journal* 11 (5): 543–59.
- Nishina, K. (1997) *The Fundamental Research for the Development of Learning Systems of Technical Japanese*, doctoral dissertation. Tokyo: Tokyo Institute of Technology. [in Japanese]
- Passonneau, R.J. & Litman, D.J. (1996) 'Empirical analysis of three dimensions of spoken discourse: Segmentation, coherence, and linguistic devices', in E.H. Hovy & D.R. Scott (eds) *Computational and Conversational Discourse: Burning Issues—An Interdisciplinary Account*. Berlin: Springer-Verlag, pp.161–94.
- Reichman, R. (1985) *Getting Computers to Talk Like You and Me: Discourse Context, Focus, and Semantics*. The MIT press, Bradford Books.
- Shimizu, I. (1959) *The Manner of Japanese Writing*. Tokyo: Iwamani. [in Japanese]
- Sugihara, K. (1994) *The Manner of English Writing for Scientists*. Tokyo: Chuoukouronsya. [in Japanese]
- Yamazaki, N., Tomita, Y., Hirabayashi, Y. & Hatano, Y. (1992) *Handbook of Technical Japanese*. Tokyo: Soutakusha. [in Japanese]
- Yang, J.C. & Akahori, K. (1997) 'Development of computer assisted language learning system for Japanese writing using natural language processing techniques: A study on passive voice', in *Proceedings of the Eighth World Conference on Artificial Intelligence in Education (AI-ED97)*, pp.263–70.
- Yang, J.C. & Akahori, K. (1998a) 'Error analysis in Japanese writing and its implementation in a computer assisted language learning system on the World Wide Web', *CALICO Journal* 15 (1–3): 47–66.
- Yang, J.C. & Akahori, K. (1998b) 'A discourse structure analysis of technical Japanese text', in *Proceedings of the Sixth International Conference on Computers in Education (ICCE98)*. Beijing, China, pp.643–50.
- Yang, J.C. & Akahori, K. (1999) 'An evaluation of Japanese CALL systems on the WWW comparing a freely input approach with multiple selection', *CALL Journal* 12 (1): 59–79.
- Yokobayashi, H. & Shimomura, A. (1988) in S. Nagara (ed.) *Connectives*, series 6. Tokyo: Aratake. [in Japanese]
- Zadrozny, W. & Jensen, K. (1991) 'Semantics of paragraphs', *Computational Linguistics* 17 (2).

APPENDIX 1

NOTATION USED IN THIS PAPER

Type I: regular expressions

(X)*: Match X 0 or more times (X is a regular expression)

(X)+: Match X 1 or more times (X is a regular expression)

(X)?: Match X 1 or 0 times (X is a regular expression)

X|Y: Match X or Y (X, Y are regular expressions)

^: Match the prefix of the sentence

\$: Match the suffix of the sentence

[A-Z]: Match any word which belongs to A to Z (A, Z are any words)

[^A-Z]: Match any word which does not belong to A to Z (A, Z are any words)

[ABC]: Match any word which belongs to ABC (A, B, C are any words)

[^ABC]: Match any word which does not belong to ABC (A, B, C are any words)

.: Match any word

Type II: notation used for system analysis

K: *kanji*

Kana: *hiragana* or *katakana*

J: *kanji* or *kana* of Japanese

P: phrase

S: sentence

H: headline

C: cohesive expressions

R: rules

EX: examples

Type III: parts of speech (part of the list)

Par: a postpositional particle of Japanese

Par1: Particle-Case particle-General

Par2: Particle-Case particle-Quotation

Par3: Particle-Case particle-Copula

Par5: Particle-Connected particle

Par6: Particle-Adverb used

V: verb

V-ta: past form of verb

V-te: te-form of verb

N: noun

N2: Noun-General

N4: Noun-Proper noun-General

- N10: Noun-Numeral
 N12: Noun-Suffix-General
 N17: Noun-Pronoun-General
 N18: Noun-Non independence-General
 N19: Noun-Non independence-Stem of auxiliary verb
 N20: Noun-Non independence-Adverb possible
 N21: Noun-Adverb possible

Adj: adjective

Adv: adverb

Auxv: auxiliary verb

Conj: conjunction

Pre: prefix

Suf: suffix

MN: word of modifying noun, connection to the indeclinable parts of speech that can stand as the subject of a sentence

APPENDIX 2

SOURCE OF TEXTS USED IN THE SYSTEM EVALUATION

Table 5. Source of Texts Used in the System Evaluation (in Stages 1 & 2).

Text no.	Field of text	Number of sentences
1	Physics	226
2	Physics	120
3	Architecture	244
4	Artificial Intelligence	473
5	Artificial Intelligence	400
6	Information Science	434
7	Artificial Intelligence	326
8	Artificial Intelligence	255
9	Natural Language Processing	212
10	Artificial Intelligence	276
11	Artificial Intelligence	322
12	Cognitive Science	494
13	Cognitive Science	471
14	Cognitive Science	427
15	Cognitive Science	1058

Table 6. Source of Texts Used in the System Evaluation (in Stage 2).

Text no.	Field of text	Number of sentences
21	Artificial Intelligence	337
22	Educational Technology	312
23	Educational Technology	234
24	Educational Technology	289
25	Network	228
26	Information Security	305
27	Image/Pattern Recognition	199
28	Image/Pattern Recognition	230
29	Educational Technology	286