Lexical Information for Determining Japanese Unbounded Dependency

Shin-ichiro KAMEI, Kazunori MURAKI and Shin'ichi DOI
Information Technology Research Laboratories, NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
{kamei,k-muraki,doi}@hum.cl.nec.co.jp

Abstract

This paper presents a practical method for a global structure analyzing algorithm of Japanese long sentences with lexical information, a method which we call Lexical Discourse Grammar (LDG). This method assumes that Japanese function words, such as conjunctive particles (postpositions) located at the end of each clause, have modality and suggest global structures of Japanese long sentences in cooperation with modality within predicates or auxiliary verbs. LDG classifies the encapsulating powers of function words into six levels, and modality in predicates into four types. LDG presumes the inter-clausal dependency within Japanese sentences prior to syntactic and semantic analyses, by utilizing the differences of the encapsulating powers each Japanese function word has, and by utilizing modification preference between function words and predicates that reflects consistency of modality in them. In order to confirm the encapsulation power of Japanese function words, we analyzed the speech utterances of a male announcer and found the correlation between a particle's encapsulating power and the pause length inserted after the clause with a conjunctive particle.

1 Introduction

When analyzing long sentences with two or more predicates (i.e. compound and complex sentences), it is difficult to grasp the proper structure of sentences having a large number of possible dependency (modifier-modifee relation) structures. This difficulty is more marked in Japanese than in English, since there are more syntactically ambiguous structures in Japanese. The Japanese language has few syntactic indicators for dividing sentences into phrases or clauses, unlike English with its relative pronouns and subordinate conjunctions. One of the most critical features of Japanese is that the difference between a phrase and a clause is not clear. Even subjects or other obligatory elements of clauses are omitted very often when they are indicated by contexts. In addition, the Japanese language does not have any parts of speech to clearly indicate either the beginning or end of a phrase or a clause. Another critical feature is that the Japanese language is an almost pure Head-final language, i.e., predicates and function words to signify the sentence structure appear at the end of the clause or sentence. This means that it is syntactically possible for all phrases or clauses that can modify predicates to modify all other phrases or clauses that appear in the latter part of long sentences.

These syntactic characteristics of the Japanese language make it difficult to determine the dependency (modification) structure of long sentences. Simple parsing of Japanese long sentences inevitably produces a huge number of possible modification structures. A conventional bottom-up parsing method can reduce ambiguity in modification by local information in the surface structure. However, this inclines toward an improper output, since the locally highest likelihood is sometimes low on the whole.

To overcome this problem, several methods to predict the global structure of long sentences have been proposed. One is a top-down parsing method by matching the input sentence and the domain-specific patterns (Furuse et al., 1992). Improvements made by other researchers enabled this method to parse irregular, incomplete and multiplex patterns, by describing the domain-dependent patterns in the form of grammar (Doi, Muraki, et al., 1993).

Another method employs global structure presumption to divide a sentence into clauses by utilizing general lexical information. It predicts the sentence structure prior to syntactic analysis only by utilizing domain-independent lexical information such as conjunctive particles, parallel expressions, theme transition, etc. (Mizuno et al., 1990; Kurohashi et al., 1992).

Lexical Discourse Grammar (LDG) is one of the approaches with which a global structure of a long sentence is presumed by focusing on function words (Kamei et al, 1986; Doi et al., 1991). LDG assumes that Japanese function words, such as conjunctive particles (postpositions) located at the end of each clause, convey modality, or propositional attitude, and suggest global structures of Japanese long sentences in cooperation with modality in predicates, especially within

auxiliary verbs. LDG can presume the inter-clausal dependency within Japanese sentences prior to syntactic and semantic analyses by utilizing the differences of the encapsulating powers each Japanese function word has, and by utilizing modification preference between function words and predicates that reflects consistency of modality reading or propositional attitude interpretation.

LDG is effective in reducing the syntactic ambiguities, and it has already been applied to a machine translation system. However, it has not clarified the level of the encapsulation powers of Japanese function words or the relation between modality and level. Hence we refined the concept of LDG, particularly the conjunction level of function words, and explain the outline of the refined LDG in this paper. First, we present the encapsulation power of Japanese function words, which are classified into six levels. Second, we state modification preference of Japanese conjunctive particles that reflects modality within them. Finally, we present evidence of the levels of Japanese function words. We think that the levels of clauses produce prosodic information, especially the location and length of pauses, which are influenced by the sentence global structure. We analyzed the speech utterances of a professional news announcer (male) and found a correlation between a particle's encapsulating power and the pause length inserted after the clause with a conjunctive particle.

2 Lexical Discourse Grammar

2.1 Levels of Conjunctive Particles in Japanese

In Japanese complex or compound sentences, subordinate clauses have several dependency levels relative to the main clause. Conjunctive particles, which are located at the end of clauses and which link them, are classified according to the elements that the clause can contain, or to the correlation between clauses. See the following examples with conjunctive particles *"node"*(ので) and *"nagara"*(ながら) (* is added to meaningless sentences).

1) 彼は彼女が助けたので成功した。
   Kare(He)–wa(TOPIC)
   kanojo(She)–ga(SUBJECT)
   tasuke(Help)–ta(PAST)–node(Because)
   seikou(succeed)–shita(PAST).
   He succeeded because she helped him.
2)* 彼は話したながら帰った。
   Kare(He)–wa(TOPIC)
   hanashi(Talk)–ta(PAST)–nagara(While)
   kaet(Return)–ta(PAST).
   He returned while he was talking.
3)* 彼は彼女が話しながら帰った。
   Kare(He)–wa(TOPIC)
   kanojo(She)–ga(SUBJECT)
   hanashi(Talk)–nagara(While)
   kaet(Return)–ta(PAST).
   He returned while she was talking.
4) {彼が笑いながら尋ねたので}私は答えた。
   {*Kare(He)–ga(SUBJECT)*
   warai(Smile)–nagara(While)
   tazune(Ask)–ta(PAST)–node(Because)}
   watasi(I)–wa(TOPIC)
   kotae(Answer)–ta(PAST).
   I answered as he asked while he was smiling.
5) {彼が尋ねたので}笑いながら私は答えた。
   {*Kare(He)–ga(SUBJECT)*
   tazune(Ask)–ta(PAST)–node(Because)}
   warai(Smile)–nagara(While)
   watasi(I)–wa(TOPIC) kotae(Answer)–ta(PAST).
   I answered while I was smiling as he asked.

A clause with the conjunctive particle to express 'reason' *"node"*(ので) can contain a subjective noun phrase and an auxiliary verb of past tense *"ta"*(た), while a clause with the particle indicating attendant action *"nagara"*(ながら) cannot, as shown in 1)–3). Sentence 4) in comparison with 5) shows that a clause with *"node"*(ので) can subordinate a clause with *"nagara"*(ながら), but the reverse is impossible. In these two sentences, brackets { } show subordinate clauses. Consequently, *"nagara"*(ながら) is ranked at a lower level than *"node"*(ので).

In LDG, conjunction levels of clauses are divided into six classifications according to the elements the clause can contain, as listed in Table 1. These levels construct a hierarchy, i.e., a lower level clause cannot subordinate a higher level one. The levels also represent the encapsulating powers of each Japanese function words located at the end of the clauses. Besides conjunctive particles, Japanese conjunction nouns or relative nouns are also classified and assigned a level. Here, Japanese conjunction nouns, such as "toki" (時 =when) are nouns that can often be used just like conjunctive particles when they are attached at the end of clause. Japanese relative nouns, such as "mae" (前 =before) are another type of nouns that play roles similar to those of conjunctions in English when they are modified by predicative phrases or clauses.

2.2 Modality in Conjunctive Particles and Modification Preferences

The conjunction levels we introduced above reduce the syntactic ambiguities of long sentences. However, in order to select the most reliable structure of sentences, we use another important discourse feature the conjunctive particles have, i.e., modality.

LDG assumes Japanese function words have modality or 'propositional attitudes' and suggest global structures of Japanese long sentences in cooperation with modality within auxiliary verbs. We assume that the same kind of modality in a conjunctive particle and

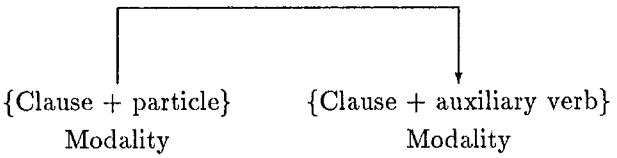
Table 1 Conjunction Level in LDG

levels	Definition of level	Example of function words	
LEV.0	can contain every element	*"to"*(と =that(quotation))	higher ↑
LEV.1	cannot contain sentential particles	*"kara"*(から =because), *"node"*(ので =because), *"keredomo"*(けれども =but)	
LEV.2	cannot contain conclusive modal	*"nara"*(なら =if), *"hoka"*(他 =besides), *"to"*(と =when), *"baai"*(場合 =in case of), *"toki"*(時 =when)	
LEV.3	cannot contain probable modal	*"mae"*(前 =before), *"ato"*(あと =after)	
LEV.4	cannot contain tense expressions	*"totomoni"*(とともに =as), *"tame"*(ため =because) *"todoujini"*(と同時に =at the same time)	↓ lower
LEV.5	cannot contain particle *"ga"*(が)	*"nagara"*(ながら =while), *"tsutsu"*(つつ =while), *"kotonaku"*(ことなく =without)	

a predicate (or an auxiliary verb) correspond to each other. From the parsing viewpoint, this suggests that each conjunctive particle has modification preference with certain predicates or auxiliary verbs.

From the viewpoint of modality, there are four predicate types in Japanese; (1) Auxiliary verbs of the first-type modality (conjecture etc.), (2) Auxiliary verbs of the second-type modality (necessity etc.), (3) Copula, and (4) Plain (present and past tense) forms of Verbs. Here, first-type modality includes conjecture, such as "darou"(だろう) which corresponds to 'may,' 'can,' 'maybe,' and 'possibly' in English auxiliary verbs, adverbs, and adjectives. Second-type modality includes necessity, such as "nakereba-naranai" (なければならない) and "ta-hou-ga-yoi"(たほうがよい) which correspond to 'have to' or 'must,' and 'had better,' 'should,' or 'preferably' in English. The Japanese Copula "da" (だ) or "desu" (です) means definition or speaker's judgment with confidence. Plain forms of verbs are the present or past tense forms of verbs without any modal auxiliary verbs. These forms do not have any modal morpheme, but when they which appear at the end of the sentence and are followed by a period they CAN convey modality, that is, attitudes or intentions of the subject or speaker. Plain forms of verbs in a relative clause which modify a nominal phrase do not have such modality.

LDG assumes that each conjunctive particle has a preference in modifying predicates or auxiliary verbs with consistent modality. There are six levels of modality in conjunctive particles, and there are four types of modality in predicates, as mentioned above. A subordinate clause with modality modifies a consistent modality predicate type. The following figure illustrates the modality consistency between particles and auxiliary verbs in Japanese sentences.



Take the Japanese conjunction noun "toki" (時 = when, if) for example. This word corresponds to either the English conjunction 'when' with neutral reading or 'if' with conjecture modality. When the word "toki" is used as the 'if' reading, this word modifies a clause in which the modality is expressed. In most cases, auxiliary verbs such as "darou" (だろう = may, maybe) or "ta-hou-ga-yoi"(たほうがよい = had better, should, preferably) express the modality of the modifee clause. The Japanese language has some words that indicate or emphasize the fact that the word "toki" is being used as the "if" reading. One of them is the adverb "moshi" (もし) that indicates a supposition reading is applicable. This adverb is never used by itself and always modifies conjunctive forms such as "toki," "nara," "to," and so on, and selects or emphasizes the supposition reading of the conjunctive forms. Another such word is the particle "wa" (は) , which is usually used as a topic marker for a sentence. When "wa" is attached to "toki," that is, in the form of "toki-wa," the supposition reading is enhanced. This tendency is strengthened by the use of comma after the phrase "toki-wa." The phrase "toki-wa" tends to be used to modify phrases with auxiliary verbs of modality.

When this phrase with modality modifies a plain form of a verb with a period at the end of the sentence, the readers recognize that the plain form of the verb contains a kind of modality, such as the subject's or speaker's intention. In other words, modality information of the subordinate clauses is attached to the plain form of the main verb. The following figure illustrates this interpreting mechanism.

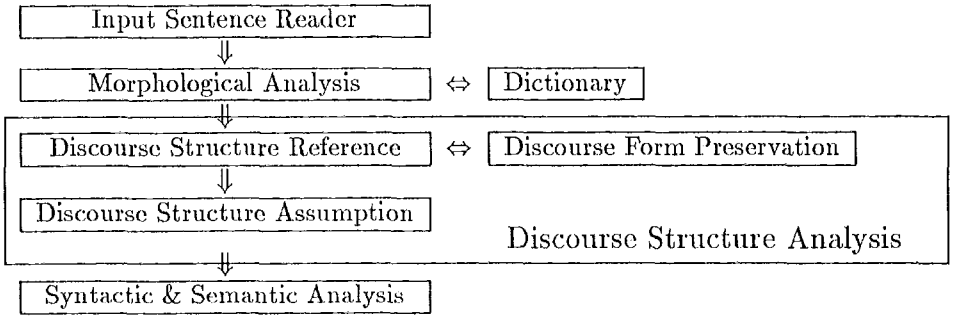
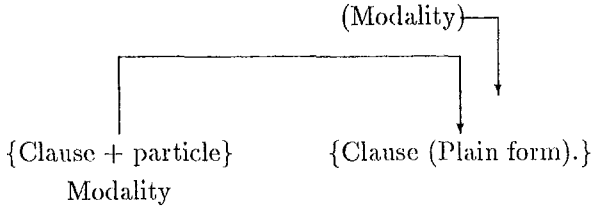
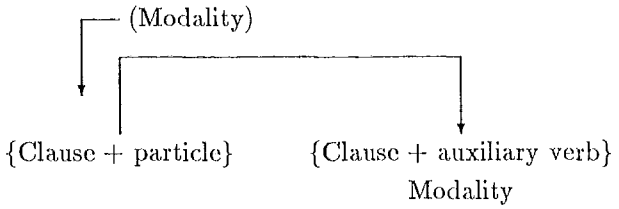


Figure 1 Analysis based on LDG



In contrast, when a subordinate clause does not have modality explicitly and modifies a clause with modality, the readers interpret the subordinate clause as that with a kind of modality such as conjecture. The following figure illustrates this situation.



The modality coincidence described in this section is the base for analyzing Japanese long sentences. The Japanese language has few syntactic indicators to show the segments of sentences, but is rich in semantic indicators which suggest sentence structure. The semantic indicators are the modalities that a wide range of parts of speech have. Conjunctive particles, adverbs, and even plain forms of verbs can have modality in the Japanese language. The modality structure is the key to comprehending Japanese long sentences.

2.3 Japanese Sentence Structure Presumption

We assume that the modality structure can mainly be detected by lexical information. Based on this assumption, LDG presumes the sentence structure before syntactic and semantic analyses on the basis of previously collected lexical information that characterizes the lexical discourse.

Figure 1 shows the configuration of our Japanese long sentence analyzer based on LDG. Input sentences are first analyzed morphologically. The part 'Discourse Structure Analysis' in Fig. 1 then presumes the sentence structure, before syntactic and semantic analysis. Here, 'Discourse' means an inner-sentence congruence in Japanese long sentences that contain two or more predicates.

In order to reduce the huge number of syntactic structures of Japanese long sentences and give priorities to each possible structure, the analyzing method based on LDG uses global modality structure focusing on lexical information.

First, the Discourse Structure Reference module reduces the number of possible syntactic structures, using the level of conjunctive particles described in the previous section. After that, the Discourse Structure Assumption module gives priorities to each possible syntactic structure, using the modification preference based on modality.

3 An Application of LDG

3.1 Pause Control with LDG

The level of conjunctive particles, which indicates the structure of the Japanese long sentences, is the most important feature of LDG. In this section we apply the level to another linguistic phenomenon in order to confirm the validity of this model.

The sentence structure influences a wide range of linguistic phenomena. One example is prosodic information (Dorffner et al., 1990; Iwata et al., 1990; Kaiki et al., 1990; Sakai et al., 1990). If the correct sentence structure is acquired for each input sentence, prosodic information can be accurately calculated. As yet, even the most up-to-date, advanced systems have not achieved the analysis in the deep structure, therefore sentence structure presumption in the surface structure is essential for a robust prosodic control system. LDG meets this requirement since it presumes the sentence structure by means of function words occurring on the surface (Doi, Kamei, et al., 1993). Hereafter, we propose the prosodic control system based on LDG.

The presumption function for sentence structure (lexical discourse) by LDG is applied to pre-processing

Table 2 Pause length data

Conjunction Levels	With a comma		Without a comma	
	Number of cases (with/without a pause)	Average pause length [msec]	Number of cases (with/without a pause)	Average pause length [msec]
LEV.0(*"to"*,*"tte"*,etc.)	0	–	7 (1/ 6)	7.9
LEV.1(*"ga"*,*"kara"*,*"node"*,etc.)	11 (11/ 0)	461.6	3 (3/ 0)	563.3
LEV.2(*"ba"*,*"to"*,*"tara"*,etc.)	11 (11/ 0)	437.0	13 (11/ 2)	243.8
LEV.3(*"ato"*,etc.)	2 (2/ 0)	277.5	1 (0/ 1)	0.0
LEV.4(*"tame"*,*"temo"*,*"hodo"*,etc.)	5 (5/ 0)	421.5	8 (6/ 2)	201.6
LEV.5(*"tsutsu"*,*"nagara"*,*"zuni"*,etc.)	4 (4/ 0)	293.8	5 (2/ 3)	120.0
Verbs in adverbial form	40 (40/ 0)	468.8	6 (6/ 0)	252.1
Verbs in adverbial form + *"te"*(て)	14 (12/ 2)	331.8	30 (15/15)	127.4
Adjectives in adverbial form	3 (3/ 0)	542.5	19 (8/11)	89.2
Adjectives in adverbial form + *"te"*	1 (1/ 0)	410.0	6 (4/ 2)	56.7
Predicative auxiliary verb *"da"*(だ)	4 (4/ 0)	603.1	3 (3/ 0)	359.2

ahead of speech synthesis, in a text-to-speech system. It can presume the global sentence structure through lexical information without any analysis in the deep structure. It is also possible to consider the pause length inserted after each clause in relation to the lexical information in LDG. In other words, pauses are more frequently inserted after the clause of the higher conjunction levels than those of the lower levels. Consequently, the pause length and location can be more efficiently controlled with the LDG conjunction levels.

To develop a text-to-speech conversion system with LDG, it is necessary to prepare the LDG conjunction level information of a large number of conjunct equivalents such as conjunctive particles. Statistical data should also be collected from human speech and reading, in regard to the correlations between pause length and the LDG conjunction levels. This substantial data is added to the lexical information to be used for speech synthesis in cooperation with pronunciation and accent.

3.2 Data Analysis

To confirm the correlation between the conjunction level and the pause length, we have analyzed speech data spoken by a professional news announcer (male), reading newspapers and magazines at a regular speed. We extracted conjunctive particles and verbs, auxiliary verbs and adjectives in adverbial form from the speech data, and classified these words by the LDG conjunction level. The average pause length for each level was calculated for two separate cases; words preceding a comma and words without a comma. See Figure 2 and Table 2.

For words without a comma (marked with white bars in Fig. 2), the result shows that the higher the conjunction level is, the longer the average pause length is (except for LEV.0, which is a particle for quotation). This tendency basically does not depend on whether or not a comma exists after the words. However, for words with a comma (marked with black bars in Fig. 2) pause length of Lev. 3 is shorter than that of Lev. 4. We suppose that the reason for this phenomenon is that a comma adds modality to the words and lengthens the pauses, as described in the previous section. Taking the comma effect into consideration, we can conclude that there is a solid correlation between the LDG conjunction level and the pause length.

LEV.0 (*"to"*(と) and *"tte"*(って) functioning in a similar way as quotation marks), however, requires a careful observation. This conjunction level, the highest rank, can contain every element, even an independent sentence. In this case, the relation between the conjunctive particle and its preceding clause is so weak that a pause tends to be inserted BEFORE the conjunctive particle, not after it. Therefore, in the present data, a pause was inserted after the particle in only one case out of seven.

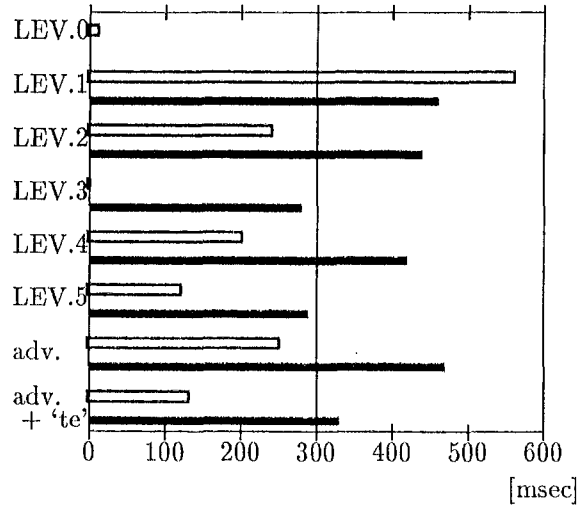


Figure 2 Average Pause Length Diagram

In Table 1, no level is assigned to two of the most

frequent groups: verbs and auxiliary verbs in adverbial form, and verbs and auxiliary verbs in the same form with a conjunctive particle "*te*"(て). These groups are difficult to allocate to a single level, as they are used in expressing many factors such as parataxis, cause, means, attendant circumstances, and because they vary semantically and syntactically. However, in reference to the pause length data, the adverbial verbs in the former group might fall into LEV.1 or LEV.2, while those in the latter group with "*te*"(て) might fall into LEV.4 or LEV.5. Conventionally, these two groups are often treated as one "adverbial form", although many functional differences have been pointed out between them. Our data supports the difference with respect to the pause length. There are identical tendencies between two adverbial forms of adjectives: ("*-ku*"(～ク) and "*-kute*"(～クテ)) and two adverbial forms of pseudo adjectives: ("*-ni*"(～ニ) and "*-de*"(～デ)).

4 Concluding Remarks

We have proposed a practical method for a global structure analyzing algorithm of Japanese long sentences with lexical information (Lexical Discourse Grammar: LDG). This model assumes that Japanese conjunctive particles convey modality, and modality structure can basically be detected by lexical information. We assign a 'conjunction level' to each conjunctive particle and reduce the number of possible syntactic structures of Japanese long sentences. In addition, we assume that all conjunctive particles have a modification preference according to their modality. This preference assigns priorities to the possible structures of the sentences.

We applied LDG to a prosodic information control method in a Japanese text-to-speech conversion system to confirm the conjunction level experimentally. This method controls pause location and length in speech synthesis with the conjunction level in LDG, using only lexical information with no need for syntactic analysis. Even so, it can tune the pause length more finely than methods without sentence structure presumption. Analyzing speech data, we confirmed a correlation between the level of a function word and the length of a pause inserted after that word. We are now in the process of developing a speech synthesis system with this method, by defining the default pause length for each conjunction level. In future research, LDG will also be applied to other prosodic information (rhythm and intonation).

There can be little doubt that LDG will be more effective when two or more conjunct equivalents of different levels appear in one sentence, since the LDG conjunction levels are closely related to the inter-clausal dependency. Unfortunately there were few such cases in the data used in this paper. In future work, we will collect such data to proove this hypothesis, in so doing will refine our method to improveits ability to analyze long Japanese sentences.

References

[1] S. Doi, K. Muraki and S. Kamei. 1991. Lexical Discourse Grammar and its Application for Decision of Global Dependency (II). IEICE Technical Report, NLC91-29(PRU91-64). (in Japanese)

[2] S. Doi, K. Muraki, S. Kamei, and K. Yamabana. 1993. Long Sentence Analysis by Domain-Specific Pattern Grammar. In *Proceedings of EACL93.*

[3] S. Doi, S. Kamei, K. Muraki, Y. Mitome, and K. Iwata. 1993. Prosodic Information Control by Lexical Discourse Grammar. In *Proceedings SIG-SLUD of the JSAI*, SIG-SLUD-9301-4. (in Japanese)

[4] G. Dorffner, E. Buchberger and M. Kommenda. 1990. Integrating Stress and Intonation into a Concept-to-Speech System. In *Proceedings of COLING90*, 1990

[5] O. Furuse and H. Iida. 1992. An Example-Based Method for Transfer-Driven machine translation. In *Proceedings of TMI'92*, pp.139-150.

[6] K. Iwata, Y. Mitome and T. Watanabe. 1990. Pause Rule for Japanese Text-to-speech Conversion Using Pause Insertion Probability. In *Proceedings of ICSLP*, 2, pp.837-840.

[7] N. Kaiki and Y. Sagisaka. 1990. Analysis of Pause Duration based on Local Phrase Structure. IEICE Technical Report, SP91-130. (in Japanese)

[8] S. Kamei and K. Muraki. 1986. Proposal of Lexical Discourse Grammar. IEICE Technical Report, NLC86-7. (in Japanese)

[9] S. Kurohashi and M. Nagao. 1992. Dynamic Programming Method for Analyzing Conjunctive Structures in Japanese. In *Proceedings of COLING92*, pp.170-176.

[10] J. Mizuno and J. Nakagaki. 1990. A Study for three Structures of Japanese Sentence. IPSJ SIG Notes, 90-NL-76-4. (in Japanese)

[11] S. Sakai and K. Muraki. 1990. From Interlingua to Speech : Generating Prosodic Information from Conceptual Representation. In *Proceedings of ICASSP90*, S6a.10.