Treebank-Based Automatic Acquisition of Wide Coverage, Deep Linguistic Resources for Japanese

Masanori OYA
M. Sc.,
School of Computing,
Dublin City University
Supervisor:
Prof. Josef van Genabith

June 20, 2009

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of M.Sc., is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:_____

Masanori Oya

Student ID: 55155243

Date: June 2009

Contents

List of Figures

List of Tables

Abstract

The objective of this thesis is to design, implement and evaluate a methodology for the automatic acquisition of wide-coverage treebank-based deep linguistic resources for Japanese, as part of the GramLab project which focuses on the automatic treebank-based induction of multilingual resources in the framework of Lexical-Functional Grammar (LFG).

After introducing the basic framework of LFG in Chapter 2, I describe the core syntactic and morphological aspects of Japanese in Chapter 3: non-configurationality; the concept of "bunsetsu" or syntactic units and their dependency relationship represented in Directed Acyclic Graphs (DAGs); topicalisation by a particular particle; and frequent use of zero pronouns with or without overt antecedents. Inflecting parts-of-speech and non-inflecting parts-of-speech of Japanese are also described with examples.

In Chapter 4, I provide the linguistic representation of core grammatical features and functions of Japanese in the framework of LFG. I use Directed Acyclic Graphs (DAG) as a framework for the unified representation of surface syntactic, morphological and lexical information in an LFG f-structure.

In Chapters 5 and 6, I describe the automatic annotation algorithm of LFG f-structure functional equations (i.e. labelled dependencies) to the Kyoto Text Corpus version 4.0 (KTC4) and the output of Kurohashi-Nagao Parser (KNP), a dependency parser for Japanese. The original KTC4 and KNP provide unlabelled dependencies only. The method presented in this dissertation also includes zero pronoun identification.

Finally, in Chapter 7 I evaluate the performance of the f-structure annotation algorithm with zero-pronoun identification for KTC4 against a manually-corrected Gold Standard of 500 sentences randomly chosen from KTC4. Using KTC4 treebank trees, currently my method achieves a pred-only dependency f-score of 94.72%. The parsing experiments using KNP output yield a pred-only dependency f-score of 82.38%.

Acknowledgments

First of all, I wish to express my deepest gratitude to my supervisor, Prof. Josef van Genabith, for his insightful suggestions on my work and considerable help for my continuing the study in Dublin, and in my hometown as a remote student. Thanks to his suggestions and help, my ideas in this thesis have been salvaged from the state of vagueness and I am encouraged to study even in a difficult situation. His comments on my draft and in discussion taught me that I still have a lot to learn, in terms of English and of how to organise my idea clearly and logically. I would like to thank Dr. Gareth Jones and Dr. Doug Arnold for their careful review and important suggestions on this thesis.

I would also like to thank all my colleagues: Grzegorz Chrupala, Yuqing Guo, Ines Rehbein, Natalie Schluter, and all the members at the National Centre for Language Technology, School of Computing, Dublin City University. Thanks to them, I enjoyed my academic and social life in Dublin in an environment where I was able to develop my work and also to have some entertaining moments. Special thanks go to Yvette Graham, who offered me help for my life in Dublin and insightful comments on my work.

I am truly grateful to Dr. Michiko Nakano, the supervisor of my Master's thesis at Waseda University, who generously offered me the opportunity to study abroad in Dublin and the environment for my research and educational activity after my coming back to Japan. I would also like to thank my colleagues and friends in Japan: Kazuharu Owada, Norifumi Ueda, Kanji Horiguchi, Yusuke Kondo, Remi Murao, Junko Negishi, Eiichiro Tsutsui, Satoshi Yoshida, Kota Wachi, Akiko Watanabe, and everyone who cared about and helped me.

My parents, Tetsuro Oya and Sanae Oya, are always helping me a lot every day. Thank you very much indeed, and I wish you good health and longevity.

Finally, I would like to express my gratitude to the Science Foundation Ireland for supporting my research with grant 04/IN/I527. I hope I will be able to contribute my time and effort for a better future of Ireland.

Chapter 1

Introduction

The objective of this thesis is to design, implement and evaluate the automatic acquisition of wide-coverage treebank-based deep linguistic resources for Japanese, as part of the GramLab project which focuses on the automatic treebank-based induction of multilingual resources in the framework of Lexical-Functional Grammar (LFG). Automatic induction of deep linguistic resources is an important area of research in NLP, since manual development of linguistic resources is time-consuming and expensive, and coverage tends to be limited. In the GramLab project, automatic treebank-based acquisition of LFG resources is conducted for English, French, German, Chinese and Spanish. The objective of my research is to induce linguistic resources from large-scale Japanese dependency banks, including automatic case-frame induction and long-distance dependency resolution.

1.1 Motivation and Related Work

Treebank-based automatic acquisition of deep linguistic resources has been one of the important topics in the field of NLP. It is expected to overcome the shortcomings of manual development of linguistic resources. Manual development is time-consuming, expensive and often limited in terms of coverage. Automatic acquisition methods should ideally be able to induce linguistic resources that are deep, including not only syntactic properties of given sentences but also semantic properties such as predicate-argument structures and long-distance dependencies (LDDs). Several methods for achieving this goal have been proposed to date, based on different grammatical formalisms like Combinatory Categorial Grammar (CCG) [Steedman, 2000], Head-Driven Phrase Structure Grammar (HPSG) [Pollard and Sag, 1994], and Lexical-Functional Grammar (LFG) [Dalrymple, 2001]. For example,

[Hockenmaier and Steedman, 2002] present an algorithm to translate the Penn Treebank into a CCG-style Treebank. [Miyao and Tsujii, 2005] develop probabilistic models for parsing with HPSG grammars acquired from the Penn-II treebank.

In this context, [Cahill et al., 2002], [Cahill et al., 2003], [Cahill et al., 2004], and [Cahill et al., 2005] develop a method for automatic annotation of LFG f-structure information on the Penn-II Treebank. The goal of the GramLab project is to apply and, where necessary, adapt this methodology to treebanks for different languages (German, French, Spanish, Chinese and Japanese) in order to automatically acquire deep multi-lingual resources from them. In my research as part of the GramLab project I develop a method to automatically annotate f-structure functional equations on the Kyoto University Text Corpus version 4 (KTC4), so that we can acquire f-structure representations for the sentences. Furthermore I applied my method to enrich the output of a Japanese dependency parser (Kurohashi-Nagao Parser, KNP), so that the parser output contains the functional structure equations (f-structure equations) for each sentence. F-structure equations are essential to construct the f-structure representation of each sentence; an f-structure is somewhat more language-independent than the original parser output, and hence this enriched parser output for Japanese text is available for further cross-linguistic research or applications such as machine translation. The method is based on the assumption that non-configurational, free word-order languages, of which Japanese is one example, do not require phrase structure trees as an indispensable level of linguistic representation. Rather, the rich morphological information on each unit in a sentence, along with the unlabelled dependency information between syntactic units in the KTC4, provide us with as much information as phrase-structure trees for more configurational languages.

1.2 Methodology: Treebank-Based Automatic Acquisition of Deep Linguistic Resources

This section introduces the method of Treebank-Based Automatic Acquisition of Deep Linguistic Resources, which has been developed in the GramLab Project at DCU. First, Lexical-Functional Grammar, the theoretical framework of the method is briefly introduced. Second, I describe how the English Penn-II Treebank is annotated automatically with f-structure functional equations. The third and last sections describe the application of automatic f-structure functional annotations to parsing with long-distance dependency

(LDD) resolution and subcategorisation frame extraction.

1.2.1 Lexical-Functional Grammar

Lexical-Functional Grammar is a constraint-based theory of linguistic structure which assumes parallel structures, each of which represents a particular level of linguistic information designed to model natural languages. The groupings of words and their hierarchical relationship in terms of phrases are expressed in Constituent Structure (c-structure). This level of linguistic representation is relatively language-specific, and different languages show different settings for constituent structures. More abstract, language-independent levels of linguistic information such as grammatical functions, temporal-modal information, subcategorisation frames, etc. are represented in Functional Structure (f-structure). Therefore two sentences from two different languages with the same meaning are often (but not always) represented by two f-structures which share basically the same structure, regardless of the language-specific differences in surface realisation as represented in the c-structures of these sentences. Formally, the c-structure and f-structure of a sentence are in correspondence through constraints called functional equations annotated to c-structure nodes. C-structure is determined by context-free phrase structure rules annotated with functional equations. Each node in a c-structure projects a piece of f-structure and together the annotated nodes describe the f-structure of the sentence as a whole, establishing the correspondence between c-structure and f-structure.

For example, a context-free phrase structure rule (1) for English is annotated with appropriate functional equations: ↑SUBJ=↓ on the first node labelled NP, and ↑=↓ on the second node labelled VP. The up arrow in an equation refers to the f-structure of the mother node of the node to which the equation is annotated, and the down arrow to the f-structure of the node to which the equation is annotated. The rule as a whole states that the syntactic category S consists of an NP and VP(in this order); and that the NP node contributes the SUBJ in the f-structure of the whole sentence (the f-structure projected from S), and that the f-structure of the VP node is the same as the f-structure of the S node, in other words, that the VP is the head of the S.

(1)

S→	NP	VP
	↑SUBJ =↓	↑=↓

Lexical items constitute the leaves in a phrase structure tree, and they provide the f-structure of the whole sentence with lexical information, and the information provided by lexical items is provided by functional equations on the nodes of context-free phrase structure trees.

What lexical information provides to syntax differs in different languages; for configurational languages, e.g., word order determines the grammatical function of syntactic constituents in a sentence, and functional equations for grammatical functions must be annotated on the nodes of context-free phrase structure trees.

In the case of non-configurational languages, where e.g. morphological features on nouns determine the grammatical functions of given constituents, the specification of grammatical function is provided lexically, viz. as functional equations provided by lexical information. Lexical information plays an important role in LFG, and therefore it can appropriately account for the linguistic phenomena not only in configurational but also in non-configurational languages. Since Japanese is a non-configurational language, LFG is an appropriate framework for representing linguistic information of Japanese. Chapter 3 introduces the core syntactic and morphological features of Japanese, and Chapter 4 describes how the framework of LFG can be used to represent grammatical functions and grammatical features of Japanese.

1.2.2 Automatic F-Structure Annotation on the English Penn-II Treebank

The Penn Treebank provides categorical and configurational information (NP, VP, etc.) in terms of Context-Free Grammar trees for the sentences. This information is exploited for the automatic annotation of functional equations to nodes in CFG trees (developed by [Cahill et al., 2003], [Cahill et al., 2004], [Cahill et al., 2005], inter alia) for English. The algorithm adds functional equations to the original Penn Treebank trees of in the following steps:

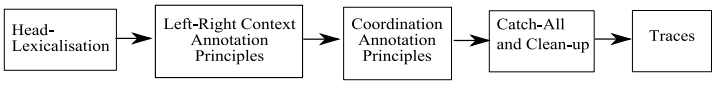

Figure 1.1: Flow Chart of the f-Structure Annotation Algorithm

First the head-lexicalisation module determines the head and the mother categories for each phrase. This step is based on the head-finding rules of [Magerman, 1995] with slight modifications. Next, functional annotations are assigned to each phrasal category based on left-right context annotation principles. These Annotation Principles are based on hand-crafted Left-

Right Annotation Matrices exploiting configurational properties of English. For example, in English, subjects tend to be the rightmost NP sisters for a VP under an S, and in the context-free phrase structure rule in (1), the functional equations are annotated according to the context within the rule, such that if S expands to NP and VP, the NP to the left of VP receives the equation ↑SUBJ=↓, and the VP the equation ↑=↓. The third step covers Annotation Principles for coordinations. This step is distinguished from the second one in order to keep the Left-Right Context Annotation Principles simple and perspicuous. The fourth step is Catch-All and Clean-Up. In this step, inappropriate annotations are fixed, e.g. over-generalisations produced by Left-Right Context Annotation Principles. The outcome of these steps so far is called "proto" f-structures; "proto" because it does not have long-distance dependencies resolved. Long-distance dependency resolution is processed in the next step, the Traces module, in which the information provided by traces and coindexation in the Penn-II Treebank is employed to resolve LDDs in terms of corresponding reentrances at f-structure. The result of this algorithm is sent to a constraint solver, by which the f-structure of each sentence is constructed according to the functional equations annotated through the steps outlined above.

1.2.3 Applying the F-Structure Annotations for Parsing with LDD Resolution

The f-structure annotation algorithm can be applied to several tasks in NLP. First, it can be applied to parsing with LDD resolution. Most probabilistic treebank-based parsers are not able to produce traces and co-indexation, hence LDDs are not represented and resolved in the output. [Cahill, 2004] and [Cahill et al., 2004] deals with this problem by automatically resolving LDDs for parser output on the level of f-structures.

Parsing can be realised in two different ways, in terms of when functional equations are annotated on the phrase structure trees. One way is to annotate functional equations on parser output trees. The other is to learn a parser whose output already contains functional equations annotated on output trees. With this in mind, [Cahill, 2004] and [Cahill et al., 2004] developed two different parsing architectures; the Pipeline Model implements the first way, and the Integrated Model implements the second way.

In the Pipeline Model, a PCFG-based parser is extracted from the training sections 01-22 of the Penn-II Treebank. Then this parser is used to parse raw text, and the output is annotated with functional equations, then these equations are sent to the constraint solver to generate f-structures of the

sentences.

In the Integrated Model, the treebank trees are annotated with functional equations, then annotated PCFG rules are extracted from the trees. Then these annotated rules are used to parse new text, and we obtain output which is already annotated with functional equations. Again, these functional equations are sent to the constraint solver to generate f-structures.

1.2.4 Applying the F-Structure Annotation Algorithm for Subcategorisation Frame Extraction

The f-structure annotation algorithm can also be applied to subcategorisation frame extraction. Since the f-structures automatically generated by the method explained above provide rich semantic information in the form of predicate-argument structure with LDD resolution, it can be used for automatic subcategorisation frame extraction from a treebank.

A method developed by [van Genabith et al., 1999], [O'Donovan, 2006] determines the local predicate value of each embedded f-structure in the f-structure for each sentence in the treebank, and collects the subcategorisable grammatical functions present at the level of embedding. This method does not predefine the subcategorisation frames before the extraction, but it fully reflects LDDs, and deals with both active and passive voice. The extracted subcategorisation frames are evaluated against COMLEX English Syntax Lexicon ([Grishman et al., 1994]) and the Oxford Advanced Learner's Dictionary.

1.2.5 Multilingual Treebank-Based LFG Grammar Acquisition

The original GramLab method is designed to acquire treebank-based wide-coverage LFG resources for English. One of the interesting research questions to be asked is whether or not this method can be applied to other languages and different treebank encodings and data-structures.

To date, various adaptations of the method have been applied to a number of treebanks for different languages: [O'Donovan et al., 2005] and [Chrupala and van Genabith, 2006] applied the method to the CAST3LB Treebank for Castillian Spanish, [Burke et al., 2004] and [Guo et al., 2007] to the Penn Chinese Treebank (CTB), [Cahill et al., 2003] and [Rehbein and van Genabith, 2007] to the TIGER treebank for German, and [Schluter and van Genabith, 2008] to French. The GramLab project extends the scope and depth of treebank-based wide-coverage automatic acquisition of

deep LFG resources to other languages, such as Arabic, French and Japanese.

1.3 Automatic Acquisition of Deep Linguistic Resources for Japanese

A wide-coverage LFG grammar for Japanese ([Masuichi et al., 2003]) has been developed manually in the ParGram project ([Butt et al., 2002]) along with grammars for a number of other languages. Efforts have been made to apply the grammar in real-world applications such as the Experience Knowledge recycle system ([Yoshioka et al., 2003]) and natural language generation ([Okuma et al., 2006]). However, the Japanese XLE requires a large amount of development time since it has been hand-crafted. Treebank-based automatic PCFG grammar acquisition from a Japanese corpus has been investigated by some researchers ([Noro et al., 2005]), but the results remain language-specific and difficult to apply to further cross-linguistic applications such as machine translation. More appropriate ways to automatically acquire Japanese NLP resources are to focus on the morphology of words which provides us with various kinds of linguistic information, and on dependency relationships among the syntactic units in a sentence.

With this in mind, I use the KTC4 dependency bank ([Kurohashi and Nagao, 1997, Kurohashi and Nagao, 1998]) as the corpus from which wide-coverage LFG resources are acquired. The method I introduce implements the idea that the part-of-speech tags on each morpheme and the unlabelled dependency tags on each syntactic unit in KTC4 provide us with enough information for constructing "proto" f-structures for the texts in the corpus, without employing PCFG-style syntactic trees. This idea is inspired by the difference in the type of syntactic information encoded in the Penn-II treebank and in the Japanese corpus. This difference reflects the language-specific properties of Japanese. Japanese is a non-configurational language which has relatively free-word order and whose grammatical functions associated with phrases are not primarily determined by word order (as in English), but by the morphology of distinguished elements of syntactic phrases, such as case particles for specifying the grammatical function of an NP (e.g., the case particle "-wo" specifies that the noun phrase with this particle is an OBJ of the verb on which this noun phrase is dependent), or verbal inflections for specifying the tense or modal information, and sometimes for the distinction between relative clauses and sentential modifiers.

This language-specific property of Japanese motivates a major difference between the method used in my research and that in Cahill et al. (2002,

2003, 2004), viz., for Japanese f-structure equations are annotated not on the syntactic trees, but directly on the syntactic phrases as represented in KTC4 according to the morphological information within the phrase. In this sense, what must be induced automatically from large corpora of more configurational languages (e.g., English, French, Spanish, among others) has already been implicitly specified in KTC4 in terms of morphological information annotated as part-of-speech tags on each morpheme, and based on these pieces of information together with the unlabelled dependencies provided by KTC4 one can induce LFG grammatical functions and other types of linguistically relevant information.

The flow of information in my method roughly resembles that in Cahill et al. (2002, 2004, 2004a)'s method. First, the original tags annotated on KTC4 are exploited to further annotate appropriate functional equations, and this step corresponds to the Head-Lexicalization and the Left-Right Context Annotation Principles of Cahill et al. (2004, 2004a)'s method. Then, the next step corresponds to Coordination Annotation Principles in their method. F-structure functional equations acquired from KTC4 right after the first step which include coordination must be fixed; in KTC4 the first coordinated phrase is analysed to be dependent on the next coordinated phrase, hence it results in inappropriate f-structure annotations. Finally, further remaining inappropriate f-structure functional equations are fixed, which corresponds to the Catch-All and Clean Up step in Cahill et al.'s algorithm.

1.4 Structure of this Thesis

The following chapters in this thesis introduce, describe and discuss central issues related to automatic acquisition of deep linguistic resources for Japanese.

The structure of this thesis can be represented as follows in a directed acyclic graph:

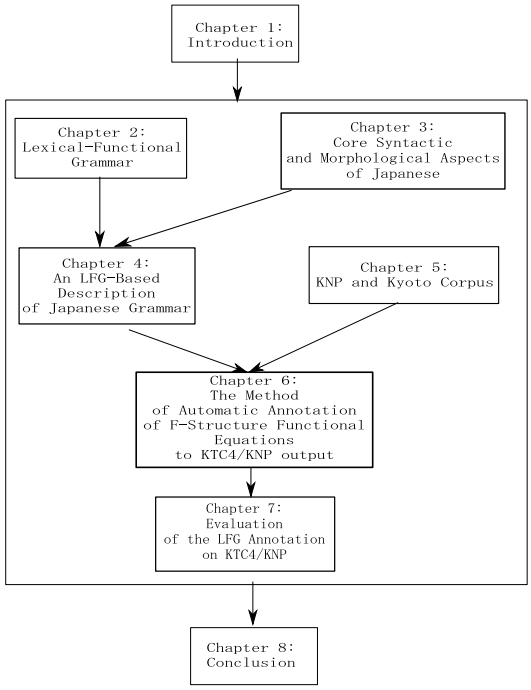

Figure 1.2: Structure of this thesis

Chapter 2 describes the basic framework of LFG with more detail, concentrating on the correspondence between different levels of linguistic representation, functional well-formedness, the subcategorisation frames of verbal predicates, long-distance dependency, control, and anaphora. Chapter 3 provides a general description of core syntactic and morphological aspects of Japanese, which are relevant to the accounts in later chapters. Chapter 4 gives the linguistic representation of core grammatical features and functions of Japanese in the framework of LFG. In this chapter, I propose the idea of using Directed Acyclic Graphs (DAG) as a framework for surface syntactic representation of Japanese. Next, Chapter 5 introduces KNP and the Kyoto Corpus and describes the KNP algorithm for parsing Japanese text, as well as the Kyoto Corpus data-structures and encoding conventions. Chapter 6 introduces the LFG annotation method for the KTC4 dependency bank and KNP parser output. Evaluations of this method are presented and discussed in Chapter 7, and Chapter 8 concludes this thesis.

Chapter 2

Lexical-Functional Grammar

2.1 Introduction

This chapter describes the basic framework of Lexical-Functional Grammar (LFG). This system of grammatical representation was first invented by [Bresnan, 1978], [Bresnan, 1982b], [Bresnan and Kaplan, 1982]. LFG has enjoyed continued popularity and development in applied, theoretical and computational linguistics. The ParGram project ([Butt et al., 2002]) manually develops parallel grammars for a number of languages using a shared set of features in the framework of LFG, using the XLE processing and development environment.

The success of LFG is due to the fact that the framework of LFG enables us to deal with different languages using a common and at the same time flexible representation format; LFG involves several levels of representation for grammatical knowledge about a sentence, and the pieces of information represented at these different levels are integrated through functional descriptions. The three levels of representation in LFG are the following; *constituent structure (c-structure), functional structure (f-structure)* and *argument structure (a-structure)*. There are other levels such as semantic structure or discourse structure, but they are not dealt with in this thesis.

Natural languages show diverse syntactic properties, and describing them only at one particular language-specific level of representation leads us to postulate operations on the representation, which can be ad-hoc and linguistically unmotivated. In addition to language specific properties, LFG can properly capture more abstract language-independent properties of grammatical knowledge at f-structure, the functional level of representation. This level is constructed in a principled way, in a discrete, step-by-step manner, and observes well-formedness conditions. Various linguistic phenomena

such as long-distance dependency, control and anaphora are represented at f-structure.

The following sections show each of these different levels of representation, and describe the correspondence between them. Section 2.2 describes two different levels of representation (c-structure and f-structure), and the correspondence between them. Section 2.3 describes the well-formedness conditions on f-structure. Section 2.4 describes subcategorisation of core arguments by the predicate of a clause. Section 2.5, Section 2.6 and Section 2.7 describe how f-structure representation encodes long-distance dependency, control and anaphora, respectively. The account presented in this chapter is based on the seminal work for LFG by [Bresnan, 1982b] and [Bresnan and Kaplan, 1982], and the summaries of its later developments by [Dalrymple, 2001] and [Falk, 2001].

2.2 C-structure and F-structure

C-structure representations consist of phrase-structure trees which are specified by the phrase structure rules (PS rules) of a context-free grammar, each node of which is annotated with functional equations. Functional equations describe f-structures. The following are examples of PS rules (for English) annotated with functional equations. The up arrow refers to the f-structure of the node which immediately dominates the annotated node, and the down arrow refers to the f-structure of the annotated node itself:

$$\text{(2.1)}\quad \begin{array}{l} \text{S} \rightarrow \end{array} \begin{array}{cc} \text{NP} & \text{VP} \\ (\uparrow\text{SUBJ})=\downarrow & \uparrow=\downarrow \end{array}$$

$$\text{(2.2)}\quad \begin{array}{l} \text{VP} \rightarrow \end{array} \begin{array}{cc} \text{V} & \left(\begin{array}{c} \text{NP} \\ (\uparrow \text{OBJ}) = \downarrow \end{array} \right) \\ \uparrow=\downarrow & \end{array}$$

$$\text{(2.3)}\quad \begin{array}{l} \text{NP} \rightarrow \end{array} \begin{array}{cc} \left(\begin{array}{c} \text{Det} \\ \uparrow=\downarrow \end{array} \right) & \begin{array}{c} \text{N} \\ \uparrow=\downarrow \end{array} \end{array}$$

In the PS rule (2.1), the equation annotated on NP states that the value of the subject feature of the f-structure of S is the f-structure contributed by the NP, and the equation on VP states that the f-structure of VP is the same as the f-structure of the S. In (2.2), the f-structure of V is the same as the f-structure of the VP, and the f-structure of the NP provides the value of the object feature of the f-structure of the VP and the object NP constituent is optional. In (2.3), $\uparrow=\downarrow$ equations specify local heads. (2.3) shows that constituents can be co-heads in LFG.

Lexical items provide lexical information, which is represented in terms of functional annotations as follows:

(2.4) John, N:
(↑ PRED) = 'John'
(↑ NUMBER) = SG
(↑ PERSON) = 3^{rd}

(2.5) studies, V:
(↑ PRED) = 'study<SUBJ, OBJ>'
(↑ SUBJ NUMBER) $=_c$ SG
(↑ SUBJ PERSON) $=_c$ 3^{rd}
(↑ TENSE) = 'PRESENT'

(2.6) languages, N:
(↑ PRED) = 'language'
(↑ NUMBER) = PL

Most lexical equations describe attribute-value pairs. The value of a PRED attribute is called *semantic form*. For example, the first equation in (2.4) states that the f-structure denoted by the upper arrow has an attribute PRED whose value is 'John'. The second states that its number is singular, and the third states that it is 3^{rd} person. The semantic form of the verb "studies" states that this verb must have a SUBJ phrase and an OBJ phrase to form a complete sentence, in other words, the verb is transitive. Semantic forms are similar to subcategorisation frames in other syntactic theories, but also carry important semantic information relevant to argument structure.

Figure 2.1 presents a phrase structure tree generated by the annotated PS rules (2.1) and (2.2), and the lexical items (2.4), (2.5) and (2.6):

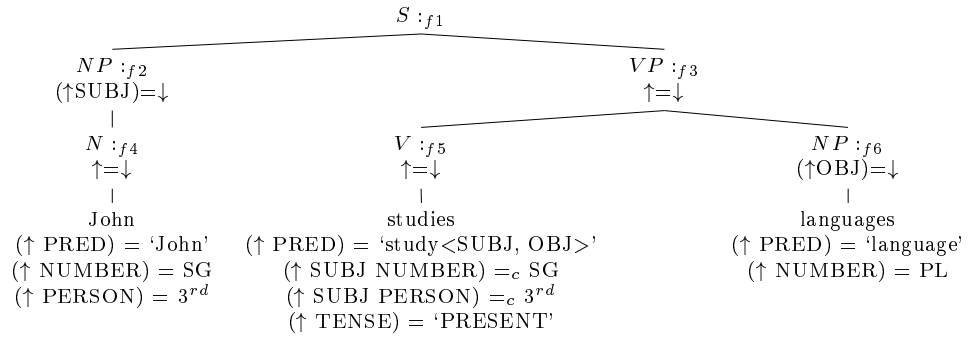

Figure 2.1: C-structure for the sentence "John studies languages."

The subscript on each node of the tree represents a functional variable which refers to the f-structure corresponding to the node. The upper and down arrows on the functional equations are instantiated by these functional variables so that the local f-structures (including the f-structure corresponding to the root node S) are described through equations.

F-structures are finite, hierarchical functions. Mathematically, such functions (i.e., f-structures) can be described using equations. A functional equation denotes a fragment of f-structure; in general, an equation

(2.7) $F(B) = X$

states that the f-structure F has the value X for the attribute B, and this is graphically represented in attribute-value matrix format as f-structure (2.8):

(2.8) $_{F}\left[\begin{array}{cc} \mathrm{B} & \mathrm{X} \end{array}\right]$

A value can be another function; in this case, the f-structure is nested:

(2.9) $F(B) = F'$

(2.10) $_{F}\left[\mathrm{B}\ {}_{F'}[\,]\right]$

Two functions merge when they are equal. With equations (2.7), (2.9) and (2.11), we have the f-structure (2.12):

(2.11) $F' = F''$

(2.12) $_{F}\left[\mathrm{B}\ {}_{F',F''}\left[\ \right]\right]$

The f-structure X or Y can have attributes and values for these attributes, and the values can be another f-structure, in which case the resulting f-structure nests recursively (equations (2.7), (2.9), (2.11), (2.13) and (2.14)):

(2.13) $F'(C) = d$

(2.14) $F''(E) = f$

(2.15) $_{F}\left[\mathrm{B}\ {}_{F',F''}\left[\begin{array}{cc} \mathrm{C} & \mathrm{d} \\ \mathrm{E} & \mathrm{f} \end{array}\right]\right]$

The notation used in LFG differs a little from that used in mathematics to describe functions (i.e., f-structures). Instead of $F(B) = X$, in LFG we write $(F \ \ B) = X$.

Therefore, the LFG functional equation (2.16) means that the f-structure f1 has the value f2 for the attribute SUBJ:

(2.16) $(f1\ SUBJ) = f2$

This is graphically represented in attribute-value matrix format below:

(2.17) $$\underset{f1}{\left[\text{SUBJ}\ \underset{f2}{\left[\ \ \right]}\right]}$$

Two functions merge when they are equal. When we have equation (2.18) in addition to (2.16), then these functions merge and we have a fragment f-structure (2.19):

(2.18) $f2 = f4$

(2.19) $$\underset{f1}{\left[\text{SUBJ}\ \underset{f2,\ f4}{\left[\ \ \right]}\right]}$$

In Figure 2.1, instantiating the $\uparrow$ and $\downarrow$ metavariables to the local f-structure identifier f_i for the leftmost NP branch of the tree, we obtain:

(2.20)
$(f1\ SUBJ) = f2$
$f2 = f4$
$(f4\ PRED) = \text{`}John'$
$(f4\ NUMBER) = SG$
$(f4\ PERSON) = 3^{rd}$

The set of equations in (2.20) describes the f-structure fragment in (2.21):

(2.21) $$\underset{f1}{\left[\begin{array}{ll}\text{SUBJ} & \underset{f2,f4}{\left[\begin{array}{ll}\text{PRED} & \text{`John'} \\ \text{NUMBER} & \text{SG} \\ \text{PERSON} & 3^{rd}\end{array}\right]}\end{array}\right]}$$

From the root and the right branch of the tree in Figure 2.1 we obtain the set of equations in (2.22) and the f-structure in (2.23):

(2.22) $f1 = f3$
$f3 = f5$
$(f3\ OBJ) = f6$
$(f5\ PRED) = \text{`}study\text{<}SUBJ, OBJ\text{>}'$
$(f5\ SUBJ\ NUMBER) =_c SG$
$(f5\ SUBJ\ PERSON) =_c 3^{rd}$
$(f5\ TENSE) = \text{`}PRESENT'$
$(f6\ PRED) = \text{`}language'$
$(f6\ NUMBER) = PL$

(2.23)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{NUMBER} & \text{SG} \\ \text{PERSON} & 3^{rd} \end{bmatrix} \\ \text{OBJ}_{f6} & \begin{bmatrix} \text{PRED} & \text{`language'} \\ \text{NUMBER} & \text{PL} \end{bmatrix} \\ \text{PRED} & \text{`study<SUBJ, OBJ>'} \\ \text{TENSE} & \text{PRESENT} \end{bmatrix}_{f1,\ f3,\ f5}$$

Finally at the root node, the f-structures (2.21) and (2.23) are merged and we obtain the f-structure for the entire sentence:

(2.24)
$$\begin{bmatrix} \text{SUBJ}_{f2,f4} & \begin{bmatrix} \text{PRED} & \text{`John'} \\ \text{NUMBER} & \text{SG} \\ \text{PERSON} & 3^{rd} \end{bmatrix} \\ \text{OBJ}_{f6} & \begin{bmatrix} \text{PRED} & \text{`language'} \\ \text{NUMBER} & \text{PL} \end{bmatrix} \\ \text{PRED} & \text{`study<SUBJ, OBJ>'} \\ \text{TENSE} & \text{PRESENT} \end{bmatrix}_{f1,\ f3,\ f5}$$

Figure 2.2 depicts the correspondence (commonly represented in terms of pointed arrows from c-structure nodes into f-structure) between the c-structure and the f-structure for the sentence "John studies languages":

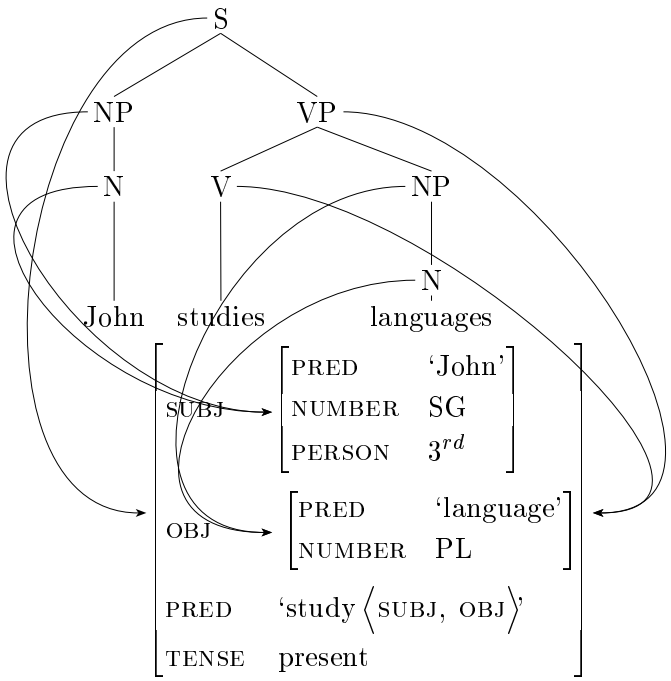

Figure 2.2: Correspondence between C- and F-structures

2.3 Functional Well-Formedness

F-structures are subject to three well-formedness conditions: uniqueness, coherence and completeness.

2.3.1 Uniqueness Condition

The Uniqueness condition ensures that every grammatical feature of each grammatical unit has a unique value. The sentence below is ungrammatical due to the conflict between the value of the NUMBER attribute in the lexical entry of "a" and in "languages":

(2.25) *John studies a languages.

(2.26) a, DET:
($\uparrow$ DET) = 'a'
($\uparrow$ NUMBER) = SG

(2.27) languages, N:

(↑ PRED) = 'language'
(↑ NUMBER) = PL

SG is the value of the NUMBER attribute of the determiner 'a'. This value cannot merge with the PL value for the number attribute of 'languages'. As a result, these lexical entries combined by rule (2.3) yield an ill-formed f-structure with two values for one attribute, violating the uniqueness condition:

(2.28)
$$* \begin{bmatrix} \textsc{det} & \text{'a'} \\ \textsc{pred} & \text{'language'} \\ \textsc{number} & \text{SG} \\ \textsc{number} & \text{PL} \end{bmatrix}$$

2.3.2 Coherence Condition

The Coherence condition requires that f-structure should be non-redundant: every core grammatical function (SUBJ, OBJ, OBJ2, OBL2 for English) present at a particular level of f-structure must be subcategorised for by the local predicate (i.e., the predicate requires these functions) at that particular level of the f-structure:

(2.29) *John studies Russian English.

The example above is ungrammatical because the second noun "English" is not subcategorised for by the verb "studies". The ill-formed f-structure for this sentence would look like (2.30):

(2.30)
$$* \begin{bmatrix} \textsc{subj} & \begin{bmatrix} \textsc{pred} & \text{'John'} \\ \textsc{number} & \text{SG} \\ \textsc{person} & 3^{rd} \end{bmatrix} \\ \textsc{obj} & \begin{bmatrix} \textsc{pred} & \text{'Russian'} \\ \textsc{person} & 3^{rd} \end{bmatrix} \\ \textsc{obj2} & \begin{bmatrix} \textsc{pred} & \text{'English'} \\ \textsc{person} & 3^{rd} \end{bmatrix} \\ \textsc{pred} & \text{'study} \left\langle \text{SUBJ, OBJ} \right\rangle\text{'} \\ \textsc{tense} & \text{PRESENT} \end{bmatrix}$$

2.3.3 Completeness Condition

An f-structure satisfies the Completeness condition if it contains values for all the grammatical functions that are subcategorised for by its local predicates:

(2.31) *Studies languages.

The example above is ungrammatical because it has no subject though it is not in the imperative mood. The ill-formed f-structure for it would be like (2.32):

(2.32)
$$*\begin{bmatrix} \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'language'} \\ \text{NUMBER} & \text{PL} \end{bmatrix} \\ \text{PRED} & \text{'study} \left\langle \text{SUBJ, OBJ} \right\rangle\text{'} \\ \text{TENSE} & \text{present} \end{bmatrix}$$

2.3.4 Constraint Equations

The lexical entry for "studies" has specifications for the number and the person attributes of the subject NP it agrees with, but they are represented by *constraint equations* (the subscript 'c' on an equation marks its status). Constraint equations do not assign any value to the attribute: they specify conditions to be met if there are values for the attribute in question in other fragments of f-structure. In the case of "studies", the equation "(↑ SUBJ NUMBER) $=_c$ SG" specifies the condition that the value of the attribute NUMBER of the subject must be SG. This condition is met if the clause whose main predicate is "studies" has a subject AND the subject has the value SG for the attribute NUMBER.

(2.33) studies, V:
(↑ PRED) = 'study<SUBJ, OBJ>'
(↑ SUBJ NUMBER) $=_c$ SG
(↑ SUBJ PERSON) $=_c$ 3^{rd}
(↑ TENSE) = 'PRESENT'

Unlike English, there are so-called 'pro-drop' languages in which syntactic subjects are optional. Example (2.34) is from Russian:

(2.34) Govorit
"He/She/It speaks."

The lexical entry for "govorit" does not have conditional equations, and it lexically introduces a subject predicate 'pro':

(2.35) govorit, V:
($\uparrow$ PRED) = 'govorit <SUBJ>; speaks'
($\uparrow$ SUBJ PRED) = 'pro'
($\uparrow$ SUBJ NUMBER) = SG
($\uparrow$ SUBJ PERSON) = 3^{rd}
($\uparrow$ TENSE) = 'PRESENT'

For pro-drop languages, the lexical entries for verbs define the features of their subjects, and in many languages different inflectional forms reflect the difference of relevant features. In a communicative situation the identity of the referent of 'pro' is determined in the given context in which the sentence is spoken or written, and the subject's person and number features help readers or listeners to determine the referent of the pro. To give another example, "goverim" means "we speak" and (2.37) is the example lexical entry with subject number plural and person 1st for the pro-drop subject:

(2.36) Govorim.
"We speak."

(2.37) govorim, V:
($\uparrow$ PRED) = 'govorim <SUBJ>; speak'
($\uparrow$ SUBJ PRED) = 'pro'
($\uparrow$ SUBJ NUMBER) = PL
($\uparrow$ SUBJ PERSON) = 1^{st}
($\uparrow$ TENSE) = 'PRESENT'

Japanese is also a pro-drop language, hence the lexical entries for verbs specify that the PRED value of the subject is 'pro'. However, verbal inflections of Japanese verbs do not mark subject features such as person and number. The same form can be used for subjects with different number and person features:

(2.38) Hanashimasu
hanas-i-mas-u
speak-conn.base-Aux-decl.base
"I/We/You/They speak" or "He/She/It speaks"

(2.39) ($\uparrow$ PRED) = 'hanas- <SUBJ>; speak'
($\uparrow$ SUBJ PRED) = 'pro'

This underspecification of verbal predicates' arguments in Japanese causes ambiguity which requires additional effort for automatic f-structure annotation for Japanese. This issue will be dealt with later in Chapter 6.

2.4 Subcategorisation and Argument Structure

The well-formedness conditions are related to subcategorisation: the Completeness condition requires the f-structure of a sentence to contain all the grammatical functions which are subcategorised for by the local predicates, and the Coherence condition requires that the f-structure should not contain any grammatical functions not subcategorised for by the predicates.

Subcategorisation for grammatical functions by a predicate is considered to show a certain regularity, since a given verb tends to be used with phrases with certain grammatical functions. A given verb needs a certain number of *syntactic arguments* in order to form a grammatical sentence. This regularity is the basis of the categorisation of verbs into transitive, intransitive or ditransitive, etc. It is possible to categorise verbs into further subcategories, according to the semantic type of each argument that a given verb can take ([Levin, 1993]).

In the framework of LFG, the subcategorisation relationship between a predicate and its arguments is represented in terms of a 'semantic form'. Consider, for example, the lexical entry for the verb "studies" (2.33), repeated below in (2.40):

(2.40) studies, V:
(↑ PRED) = 'study<SUBJ, OBJ>'
(↑ SUBJ NUMBER) $=_c$ SG
(↑ SUBJ PERSON) $=_c$ 3^{rd}
(↑ TENSE) = 'PRESENT'

The equation "(↑ PRED) = 'study<SUBJ, OBJ>"' specifies that the verb 'studies' should take a subject argument and an object argument. Since this verb takes two syntactic arguments, this verb is categorised as a transitive verb.

Along with the regularity in terms of which grammatical functions are required for an f-structure to be complete and coherent, there is another type of regularity in terms of which semantic property a given syntactic argument with a particular grammatical function must have. The regularity between the semantics of an argument and its grammatical function is accounted for by *Lexical Mapping Theory* (LMT) [Bresnan and Kanerva, 1989].

2.5 Long-Distance Dependencies

Long-distance dependencies (LDD) refer to the relationship between a filler and a gap within a sentence. A filler is an extraposed phrase, such as a WH-phrase in (2.41a), or a topicalised phrase in (2.41b), and the gap is the place where the fronted phrase would appear in an unmarked sentence:

(2.41)

a. *Which book* do you think I gave her?
b. *That book*, I don't think you will like.

"Long-distance" dependencies can span any number of words (or clause boundaries) between a filler and its gap. A filler has two different grammatical functions; one is the grammatical function which is given to the phrase locally, and the other is that which is assigned to its gap. In (2.41a), the phrase *Which book* is given the grammatical function FOCUS since it appears at the beginning of a WH-question, and also OBJ since the gap represents the 'missing' object NP following the verb *gave*. [Falk, 2001] claims that English has two structural positions for fillers: the specifier position of a CP for wh phrases, and a position adjoined to an IP for topicalised phrases. A discourse function DF (TOPIC or FOCUS) is assigned to the phrase at the filler position via PS rules given below:

(2.42)

a. $\text{CP} \rightarrow$ XP: $(\uparrow\text{DF})=\downarrow$, $(\downarrow\text{PRON})=_c \text{WH}$ C': $\uparrow = \downarrow$

b. $\text{IP} \rightarrow$ XP: $(\uparrow\text{DF})=\downarrow$, $(\downarrow\text{PRON}) \neq \text{WH}$ IP: $\uparrow = \downarrow$

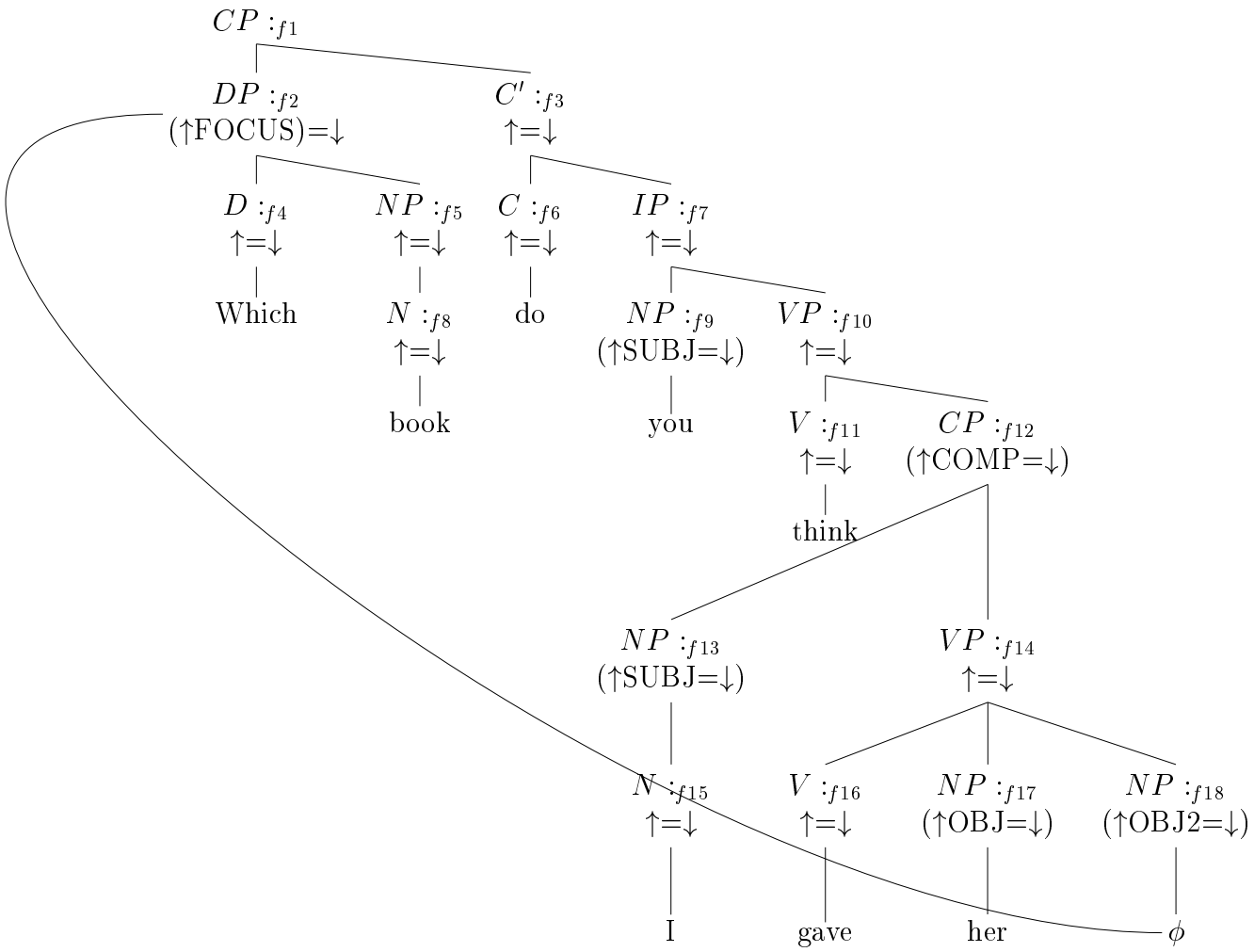

Figure 2.3: C-structure for the sentence "Which book do you think I gave her" (based on [Falk, 2001])

The LDD-resolved f-structure representation for (2.41) is Figure 2.4. The line connecting FOCUS and the OBJ2 of the verb 'gave' represents a reentrancy in the f-structure and encodes the LDD relationship between the filler (FOCUS) and the gap (OBJ2):

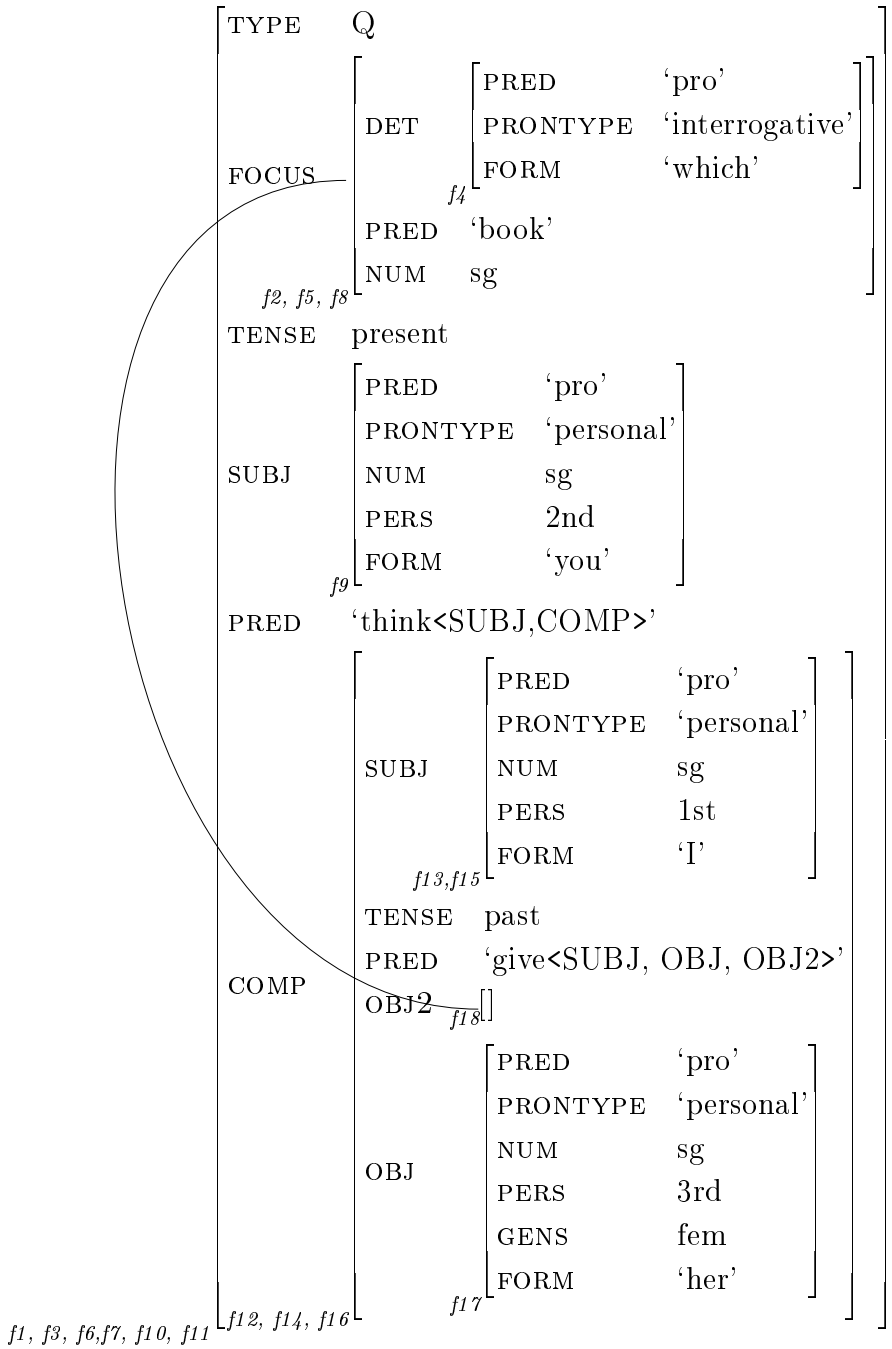

Figure 2.4: F-structure for (2.41a)

2.5.1 Functional Uncertainty

The name "Long-Distance" dependencies comes from the fact that there can be any number of words (or clause boundaries) between a filler and its gap. This is illustrated by the following sentences:

(2.43)

a. Which book did you give her?
b. Which book do you think I gave her?
c. Which book do you think Ken believes I gave her?
d. Which book do you think Sarah denies that Ken believes I gave her?
e. Which book is interesting?
f. Which book do you think is interesting?
g. Which book do you think Ken believes is interesting?
h. Which book do you think Sarah denies that Ken believes is interesting?

LDDs must be licensed by a functional equation on the filler which defines the grammatical function of its gap (Outside-In), or a functional equation on the gap which defines the discourse function of its filler (Inside-Out). Since a WH-NP gap in English can be embedded in any number of COMPs (examples (2.43)), [Kaplan and Zaenen, 1989] proposed to use the Kleene star operator to represent the intervening grammatical functions between the filler and the gap, so that any depth of embedding can be properly represented by one functional equation (2.44). This type of functional representation is called *functional uncertainty*:

(2.44)

a. Outside-In (↑DF) = (↑ COMP* GF)
b. Inside-Out (↑GF) = ((COMP* ↑) DF)

For example, the Outside-In uncertainty functional equation for the WH phrase in (2.43h) is instantiated to (↑FOCUS) = (↑ COMP COMP COMP SUBJ).

2.6 Control

In LFG, the term "control" refers to any construction in which there is (in most languages) a non-finite verb form with no overt subject, but with particular grammatical constraints on the reference of the missing subject. Sentences in (2.45) are examples of control constructions. The bracketed constituent in each of these sentences has an unexpressed SUBJ (*the controllee*),

and the phrase "the linguist" functions as *the controller* in (2.45a-e) and "the professor" in (2.45f). (2.45g) is an example of *raising*; in transformational grammar, it is analysed that the subject is raised from the initial position within the subordinate clause, to its final position as the subject of the main clause:

(2.45)

a. [To study Navajo] would please the linguist.
b. [Studying Navajo] pleases the linguist.
c. The linguist tried [to study Navajo].
d. The linguist kept [studying Navajo].
e. The professor persuaded the linguist [to study Navajo].
f. The professor promised the linguist [to study Navajo].
g. The linguist seems [to study Navajo].

Control constructions are grouped into two categories according to whether the relation between the controller and the controllee is anaphoric or functional.

2.6.1 Anaphoric Control

[Postal, 1970] pointed out that there is a pronominal element in a clause whose head is a non-finite verb. In LFG, the pronominal element, or "zero pronoun", is represented at the level of f-structure to satisfy the Completeness condition. Figure 2.5 gives the f-structure for (2.45a). The controller and the controllee are assigned the same index 'i' indicating that there is an anaphoric link between them:

$$\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{PRED} & \text{`study}\langle \text{SUBJ,OBJ} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`Navajo'} \end{bmatrix} \end{bmatrix} \\
\text{MOOD} & \text{conditional} \\
\text{TENSE} & \text{past} \\
\text{PRED} & \text{`please}\langle \text{SUBJ, OBJ} \rangle\text{'} \\
\text{OBJ} & \begin{bmatrix} \text{DEF} & + \\ \text{PRED} & \text{`linguist'} \\ \text{NUM} & \text{sg} \\ \text{INDEX} & \text{i} \end{bmatrix}
\end{bmatrix}$$

Figure 2.5: F-structure for (2.45a)

[Bresnan, 1982a] proposes that there is a universal rule for unexpressed SUBJ pronouns:

Add the optional equation (↑GF PRED)='pro' to the lexical entry of a verb.

There are constraints on which constituent can anaphorically control an unexpressed pronoun. These constraints are defined in terms of f-structure depth of the zero pronoun (*f-command*). The definitions of f-command are shown below ([Falk, 2001]):

a. The unexpressed pronoun can only be coindexed with an f-commanding function.

b. For f-structure α, β, α f-commands β if α does not contain β and every f-structure that contains α contains β.

2.6.2 Functional Control

Functional control involves an open complement grammatical function XCOMP, whose SUBJ function is left open to be controlled by an external element, and a particular functional equation in the lexical entry for verbs determines which element serves as the subject of XCOMP.

Consider the following sentence (2.46)(=(2.45c)). This is an example of *subject control*, in which the subject of the main clause is also the subject of the XCOMP:

(2.46)

The linguist tried [to study Navajo].

The subject of the verb "study" in the open complement of (2.46) is the *subject* of the main predicate, namely "The linguist". This fact is represented by the lexical entry for the verb "try", which contains a semantic form specifying that this verb subcategorises for SUBJ and XCOMP, and a functional equation which specifies that the subject of this XCOMP is the SUBJ of "try":

(2.47)

try:
(↑PRED)='try<SUBJ, XCOMP>'
(↑XCOMP SUBJ)=(↑SUBJ)

The f-structure for (2.46) is given in Figure 2.6. The line connecting "the linguist" and the empty f-structure as the value of the XCOMP SUBJ indicates that they are in the functional control relation:

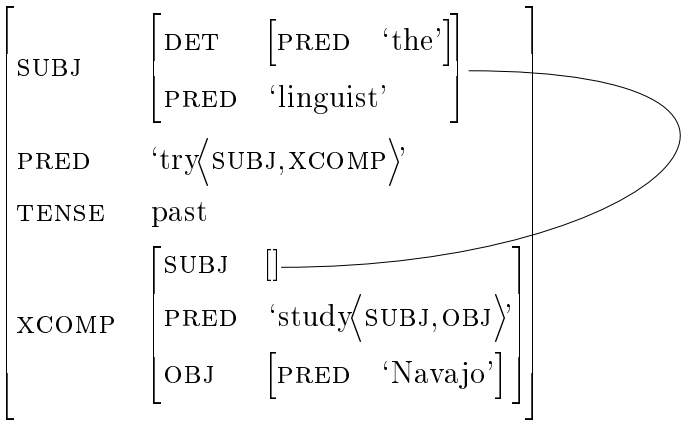

Figure 2.6: F-structure for (2.46)

Next, consider (2.48)(=(2.45f)). This is also an example of subject control:

(2.48)

The professor promised the linguist [to study Navajo].

The subject of the verb "study" in the open complement of (2.48) is the *subject* the main predicate, namely "the professor". This fact is represented by the lexical entry for the verb "promise", which contains a semantic form specifying that this verb subcategorises for SUBJ, OBJ, and XCOMP, and a functional equation which specifies that the subject of XCOMP is the SUBJ of "promise":

(2.49)

promise:
(↑PRED)='promise<SUBJ, OBJ, XCOMP>'
(↑XCOMP SUBJ)=(↑SUBJ)

The f-structure for (2.48) is given in Figure 2.7.

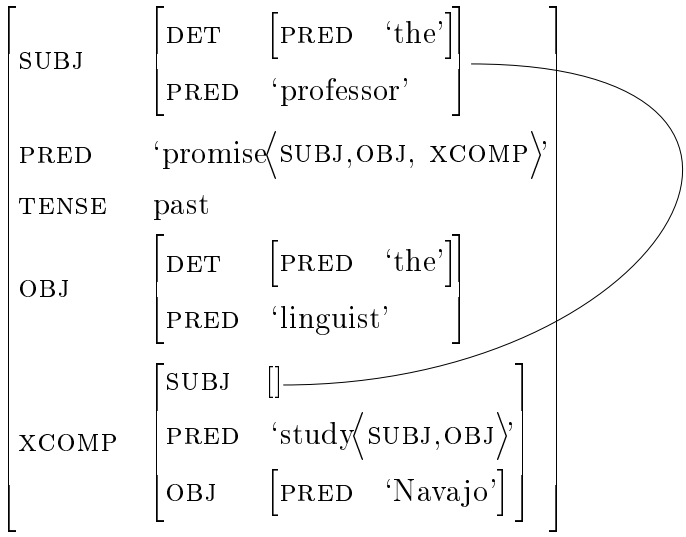

Figure 2.7: F-structure for (2.48)

Next, consider (2.50)(=(2.45e)). This is an example of *object control*, in which the object of the main predicate is the subject of the XCOMP:

(2.50)

The professor persuaded the linguist [to study Navajo].

The subject of the verb "study" in the open complement of (2.50) is the *object* the main predicate, namely "the linguist". This fact is represented by the lexical entry for the verb "persuade", which contains a semantic form specifying that this verb subcategorises for SUBJ, OBJ, and XCOMP, and a functional equation which specifies that the subject of XCOMP is the OBJ of "persuade":

(2.51)

persuade:
(↑PRED)='persuade<SUBJ, OBJ, XCOMP>'
(↑XCOMP SUBJ)=(↑OBJ)

The f-structure for (2.50) is given in Figure 2.8.

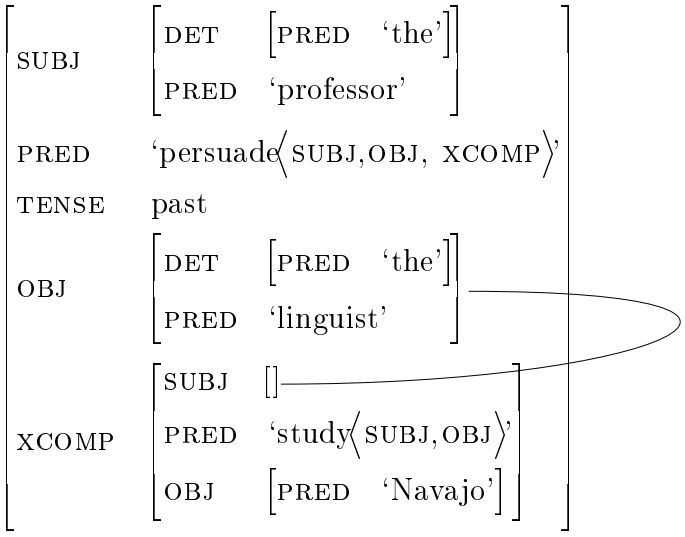

Figure 2.8: F-structure for (2.50)

2.7 Anaphora

In LFG, the term "anaphora" is used differently from the tradition of Government and Binding theory (GB) (started by [Chomsky, 1981]) and recent developments of the Minimalist Programme (MP) (started by [Chomsky, 1995]). GB/MP treat reflexive and reciprocal pronouns as one group, and other pronouns as another. GB/MP argue that the antecedent of a given reflexive or a reciprocal is determined by the structural position of the antecedent with respect to the reflexive/reciprocal in the tree structure, while this is not the case with respect to a pronoun.

LFG takes a different approach to the anaphoric relation between a pronoun and its antecedent; it is not the tree structure but the f-structure representation that defines anaphoric relations.

The conditions for binding anaphora are based on the idea of the *Minimum Complete Nucleus*, which is the smallest f-structure that contains a PRED and a SUBJ function ([Dalrymple, 2001]). The dichotomy between reflexive/reciprocal and other pronouns has been inherited from GB/MP. The binding conditions for English are as follows:

a. Reflexives and reciprocals must be bound in the Minimum Complete Nucleus which contains them.

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'the'} \end{bmatrix} \\ \text{PRED} & \text{'linguist'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{TENSE} & \text{past} \\ \text{PRED} & \text{'scare}\langle \text{SUBJ,OBJ} \rangle\text{'} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{NUM} & \text{singular} \\ \text{PERS} & \text{3rd} \\ \text{GEND} & \text{masculine} \\ \text{PRONTYPE} & \text{reflexive/*personal} \\ \text{INDEX} & \text{i} \end{bmatrix} \end{bmatrix}$$

Figure 2.9: F-structure for (2.52)

b. Pronouns must be free in the Minimal Complete Nucleus which contains them.

Consider sentence (2.52). The reflexive pronoun "himself" can have the subject of the sentence "the linguist" as its antecedent, while the pronoun "him" cannot:

(2.52)
The linguist$_i$ scared himself$_i$/*him$_i$

The f-structure for both versions of (2.52) is given in Figure 2.9:

If the reflexive pronoun has the same index as the subject "the linguist", then this reflexive pronoun is bound within the Minimum Complete Nucleus (the whole f-structure in this example), therefore the sentence is grammatical. If, on the other hand, the pronoun has the same index as the subject "the linguist", then this pronoun is bound within the Minimum Complete Nucleus and it is not free, therefore the sentence is ungrammatical.

If, on the other hand, the pronoun is not bound in the Minimum Complete Nucleus, then the sentence is grammatical:

(2.53)
The linguist$_i$ scared him$_j$.

The antecedent of a reflexive/reciprocal can be the controllee in an XCOMP. Consider (2.54).

(2.54)

The linguist$_i$ believed the informant$_j$ to have scared himself$_{*i/j}$.

The f-structure for both versions of (2.54) is given in Figure 2.10.

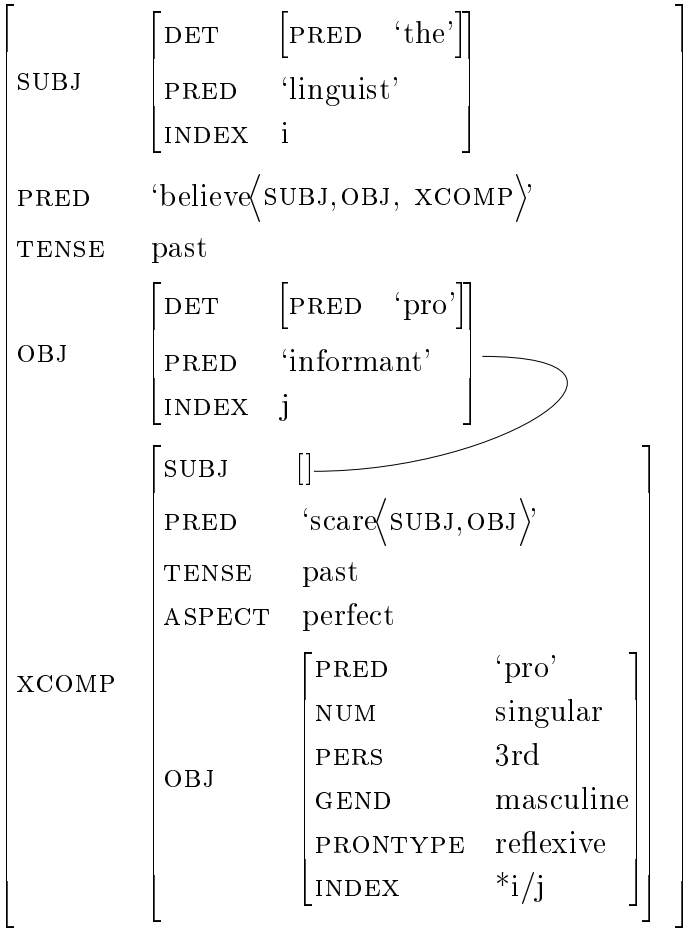

Figure 2.10: F-structure for (2.54)

If the reflexive pronoun has the same index as "the informant", which is the object of the main predicate "believe" and controls the subject within the XCOMP, then this reflexive pronoun is bound within the Minimum Complete Nucleus (the XCOMP in this example), therefore the sentence is grammatical. If, on the other hand, the reflexive pronoun has the same index as the subject "the linguist" of the main clause, then this reflexive pronoun

is bound *beyond* the Minimum Complete Nucleus, therefore the sentence is ungrammatical.

2.8 Summary

This chapter has described the basic framework of Lexical-Functional Grammar (LFG), a system of grammatical representation which was first invented by [Bresnan, 1978], [Bresnan, 1982b], [Bresnan and Kaplan, 1982], and has enjoyed continued popularity and development in applied, theoretical and computational linguistics. LFG involves a number of levels of representation for grammatical knowledge about a sentence. The three main levels of representation in the framework of LFG are *constituent structure (c-structure), functional structure (f-structure)* and *argument structure*. Two different levels of representation (c-structure and f-structure), are in correspondence with each other through functional descriptions. There are three well-formedness conditions for f-structure: Uniqueness, Coherence and Completeness Conditions. Subcategorisation of core arguments by the predicate of a clause is represented by the semantic form of the verbal predicate. Linguistic phenomena such as long-distance dependency, control, and anaphora are described in terms of f-structure representation.

Chapter 3

Core Syntactic and Morphological Aspects of Japanese

3.1 Introduction

This chapter describes the core syntactic and morphological aspects of Japanese in relation to the framework of LFG. Describing Japanese grammar in every detail needs another volume, so what is described in this chapter is core aspects which are relevant in my research. For a comprehensive linguistic profile of Japanese written in English, see e.g. [Shibatani, 1990] and [Tsujimura, 2006].

The structure of this chapter is as follows: Section 3.2 describes the non-configurationality of Japanese syntax at the sentential level and argues that the multi-structure architecture of LFG is appropriate for dealing with non-configurational languages. Section 3.3 describes the concept of "bunsetsu", syntactic units which function as the unit of syntactic investigation in Japanese grammar. I argue that these syntactic units correspond to f-structure components, and these f-structure components are combined through labelling the dependency relationship among them, to make up the f-structure for a sentence as a whole. Section 3.4 deals with topicalisation and zero pronouns, and also suggests that we need special treatments for zero pronoun identification for natural language processing of Japanese. Section 3.5 and 3.6 cover the parts of speech of Japanese, based on the account of [Masuoka and Takubo, 1992]. Section 3.5 describes the inflecting parts of speech, with some focus on the inflection forms and their functions which play important roles in the automatic annotation of grammatical functions

for syntactic bunsetsu units, the detail of which will be described in later chapters. Section 3.6 describes the non-inflecting parts of speech.

In this chapter, I try to use LFG terminology as little as possible, so that the accounts and explanations on Japanese grammar in this chapter are as much as possible free of theoretical bias, that is, they can be applicable to other syntactic frameworks and purposes depending on the readers' preferences. How these syntactic and morphological aspects will be represented in LFG will be detailed in Chapter 4.

3.2 Non-Configurationality in Japanese Syntax

Configurationality states that the order of constituents in a sentence determines the grammatical function of each constituent. In English, for example, the noun phrase which appears before a verb or an auxiliary is the (syntactic) subject of a sentence, while the noun phrase after the verb is the object of the sentence. Furthermore, the meaning of a sentence changes when the order of the noun phrases is changed. In the examples below, changing the order of noun phrases makes the sentence meaningless:

(3.1) My brother has read this book.

(3.2) ?This book has read my brother

Configurationality is reflected in the phrase structure rules annotated with functional equations in LFG accounts for English:

(3.3) $\begin{array}{lcc} \text{S} \rightarrow & \text{NP} & \text{VP} \\ & (\uparrow\text{SUBJ})=\downarrow & \uparrow=\downarrow \end{array}$

(3.4) $\begin{array}{lcc} \text{VP} \rightarrow & \text{V} & \text{NP} \\ & \uparrow=\downarrow & (\uparrow\text{OBJ})=\downarrow \end{array}$

The PS rules above state that the NP before the verb of a sentence is the subject of the sentence, and the NP after the verb is the object.

Japanese, by contrast, is a non-configurational language; in other words, the order of the constituents in a sentence is relatively free, and reordering them does not harm the intelligibility (or meaning) of the sentence:

(3.5) Watashino aniga kono honwo yonda

watashi-no	*ani-ga*	*kono*	*hon-wo*	*yom-ta*
I-part	brother-SUBJ	this	book-OBJ	read-decl.ta

"My brother read/has read this book."

(3.6) Kono honwo watashino aniga yonda

kono	*hon-wo*	*watashi-no*	*ani-ga*	*yom-ta*
this	book-OBJ	I-part	brother-SUBJ	read-decl.ta

"My brother read/has read this book."

In the examples above, changing the order of the constituents does not change the meaning of the sentence, although the listener may perceive a certain change in terms of what the speaker emphasises in the sentence.

It is important to notice that non-configurationality is not equivalent to "free word order". Japanese does not have free word order. As the example below shows, a sentence cannot be grammatical if the verb does not come at the end of a sentence (though this may be acceptable in certain context such as in poems or in lyrics):

(3.7) ? Watashino aniga yonda kono honwo

(3.8) ? Yonda watashino aniga kono honwo

Unlike for English, grammatical functions are not specified by the functional equations annotated on phrase structure rules; rather, they must be specified lexically, in particular, by the lexical properties of the particles following a noun. In the examples above, the case particle "-wo" indicates that this noun is the object of the verb, and the adverbial particle "-ga" indicates that this noun is the subject of the verb. The lexical entry for a noun with the case particle "-wo" could be as follows:

(3.9) "hon-wo": noun
PRED="HON"
↑GF = OBJ

[Bresnan, 2001] states that languages which are relatively rich in morphology may use more or less rigid phrase structure constraints, while languages which do not employ much morphology tend to have rigid, hierarchical phrase structure. Despite these language-particular differences in terms of syntactic structure, the multi-layered structural correspondence of the LFG framework can properly represent the grammatical relations among constituents within a sentence for both configurational and for non-configurational languages. For

example, the f-structures for the English sentence (3.1) and the corresponding Japanese sentences (3.5) and (3.6) are shown below. These f-structures show the common language-independent properties such as the grammatical function and the predicate of each constituent shared between the two languages. F-structures abstract away from some of the particulars of surface realisation and represent a more abstract predicate-argument structure like representation.

(3.10)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`my'} \end{bmatrix} \\ \text{PRED} & \text{`brother'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`this'} \end{bmatrix} \\ \text{PRED} & \text{`book'} \end{bmatrix} \\ \text{PRED} & \text{`read}\langle \text{SUBJ, OBJ} \rangle\text{'} \\ \text{TENSE} & \text{`past'} \end{bmatrix}$$

(3.11)
$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PADJ} & \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTCNJ} & \text{`-no'} \end{bmatrix} \\ \text{PRED} & \text{`ani; elder brother'} \\ \text{PRTADV} & \text{`-ga'} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\ \text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{INFL} & \text{`-ta'} \\ \text{TENSE} & \text{`past'} \end{bmatrix}$$

Notice that both Japanese sentences (3.5) and (3.6) with different constituent orders are both represented by the same f-structure (3.11). If there were only one level of representation, say phrase structure trees, then it would be difficult to represent the more abstract language-independent properties of given sentences across different languages, or abstract properties shared by sentences with different constituent orders. The f-structure representation enables us to represent language-independent properties of a given sentence regardless of differences in the surface syntactic level.

3.3 "Bunsetsu"(Syntactic Units) and DAG Representation of Unlabelled Dependency Relations between Bunsetsus

3.3.1 "Bunsetsu"

The previous section on non-configurational aspects of Japanese syntax may suggest that Japanese does not need phrase structure rules. We may assume that the basic phrase structure rule for Japanese is simply the following:

(3.12) $\mathrm{S} \rightarrow \left(\begin{array}{c} \mathrm{XP} \\ (\uparrow \mathrm{GF}) = \downarrow \end{array} \right)^{*} \quad \begin{array}{c} \mathrm{V} \\ \uparrow = \downarrow \end{array}$

This rule states that "XP" which refers to syntactic phrases of any category (NP, VP, AP, etc) can appear any number of times and in any order, and that a sentence ends with one verb. This rule is too permissive.

In order to account for the syntactic properties of Japanese, a certain framework other than phrase structure rules is necessary. This is the reason why the concept of "bunsetsu", or syntactic units is used in this study, along with many others in the field of Japanese computational linguistics. In this thesis, I use the term "syntactic units" with the meaning of bunsetsu.

A syntactic unit is the minimum intelligible unit of Japanese ([Hashimoto, 1948]). One syntactic unit always has one content word, and some functional word(s) can be added to it ([Nomura and Koike, 1992]). In the following example, each syntactic unit is enclosed in square brackets and labelled f_i:

(3.13) Watashino aniga kono honwo yonda.

[Watashi-no]$_{f1}$	*[ani-ga]*$_{f2}$	*[kono]*$_{f3}$	*[hon-wo]*$_{f4}$
I-part	elder.brother-SUBJ	this	book-OBJ
[yom-ta(yonda)]$_{f5}$			
read-PAST			

"My elder brother read this book."

Every sentence has one root unit which comes at the end of the sentence. Each syntactic unit except for the root unit depends on at most one syntactic unit. No unit depends on itself, and the direction of dependency is always from left to right; there is no cross dependency ([Ota, 2007], [Uchimoto et al., 2000]).

In the example above, the first unit depends on the second unit. The second depends on the last. The third unit depends on the fourth, which depends on the last. The last unit is the root unit, so it does not depend on anything.

3.3.2 Directed Acyclic Graph(DAG) Representation of Dependency

The relationship among units can be represented by a *directed acyclic graph* (DAG). Each vertex f_i in the DAG in Figure 3.1 corresponds to one of the units in the example (3.13) and each arc shows the dependency relationship:

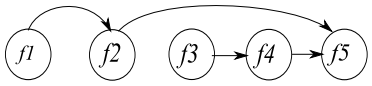

Figure 3.1: DAG for (3.13)

A DAG is a digraph without cycles ([Haggarty, 2002]). For any vertex v, there is no path that starts and ends on v. A *source* is a vertex with no incoming edges (*arcs*), while a *sink* is a vertex with no outgoing edges. A finite DAG has at least one source and at least one sink. Figure 3.2 is an example of DAG:

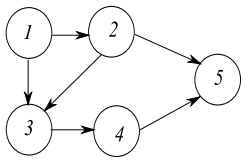

Figure 3.2: Example of a DAG

Each DAG has a *topological sort*, an ordering of the vertices such that each vertex comes before all the vertices to which it has arcs pointing. The topological sort of Figure 3.2 is Figure 3.3:

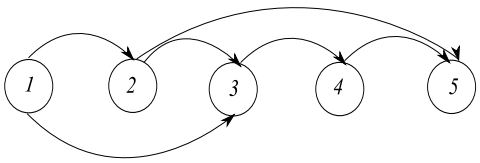

Figure 3.3: Topological Sort for Figure 3.2

The dependency relationship among syntactic units of Japanese can be represented by a labelled DAG which has the following properties, along with those of DAGs in general[1]:

[1]Dependency analysis is often used in Japanese NLP including the KNP parser

1. *Each arc is labelled with the grammatical function*: the part-of-speech of the head of a syntactic unit and the type or the absence of particle in it determine the grammatical function of this syntactic unit with respect to its target unit (target unit: the unit which the current unit is connected to by the labelled arc).

2. *The outdegree of every vertex is one, except for the outdegree of the sink*: this means that every syntactic unit, except for the root unit of the sentence, has exactly one target unit and exactly one grammatical function.

3. *The vertices are topographically ordered*: this means that every unit precedes its target unit. There is no backward dependency among syntactic units.

4. *There are no crossing arcs*: this means that no arc crosses with each other. This property is formalised as follows: For a given arc $a(v_x, v_y)$, where $x < y$, $a(v_{x+1}, v_z)$ is nested if $x + 1 < z \leq y$, and crossing if $x + 1 < y < z$.

Notice that the direction of dependency arcs is the reverse of that in Dependency Grammar, reflecting the fact that the direction of an arc specifies the direction of the flow of information from one node to another.

The DAG representation of Japanese dependency relationship resembles the model of linguistic production and comprehension by [Bresnan, 1978] based on the idea of Augmented Transfer Networks (ATN) ([Wanner and Maratsos, 1978]), in that both the DAG approach and the ATN approach process sentences sequentially, word by word in ATN and unit by unit in DAG.

The DAG representation of Japanese dependency relationships is reminiscent of Topological Dependency Grammar (TDG) ([Duchier and Debusmann, 2001]), and eXtensible Dependency Grammar (XDG) ([Debusmann, 2003]). TDG is a framework for dependency grammar with two modules of immediate dependency and linear precedence, and XDG is a generalisation of TDG, a description language for sets of labelled directed graphs. The DAG representation of Japanese dependency relationships can also be subsumed into a more general, language-independent metalanguage formalism, which is another topic of future research.

([Kurohashi and Nagao, 1998]) without a formal definition. [Suzuki et al., 2003] proposed the DAG representation of Japanese dependency relationships and applied it to question classification and sentence alignment tasks, but their representation does not label arcs with grammatical functions.

The core of Japanese syntax is to segment a sentence into units and determine the dependency among these units. The JUMAN morphological analysis system ([Kurohashi and Kawahara, 2005]) does the sentence segmentation and the KNP parser ([Kurohashi and Nagao, 1998]) does the unlabelled dependency analysis (discussed later in detail). One syntactic unit is the minimum unit to which at least one grammatical function can be assigned. The morphological information within the unit provides enough clues to determine its grammatical function. In the example (3.13) above, the second unit depends on the last unit, and the grammatical function of the second unit relative to the last is SUBJ because of the case particle "-ga" on the second unit. Similarly, the case particle "-wo" in the fourth unit specifies that this unit has the grammatical function OBJ of the root unit. The type (or absence) of the particle in a unit determines the grammatical function of the unit, and the surface order of these units is not relevant to determining the grammatical function of each unit.

The fact that (i) one syntactic unit contains one content word and that (ii) one grammatical function can be assigned to the content word leads to the idea that one syntactic unit of a sentence corresponds to part of the LFG f-structure for the sentence. The content word in a syntactic unit is the head of the unit, and thus provides the PRED value of the f-structure fragment corresponding to it, and the type (or absence) of the particle in a unit determines the grammatical function of the unit.

For example, the units in sentence (3.13) correspond to the following f-structure fragments:

(3.14)
$$_{f2}\left[\begin{array}{ll}\text{PADJ} & _{f1}\left[\begin{array}{ll}\text{PRED} & \text{'watashi; I'}\\ \text{PRTCNJ} & \text{'-no'}\end{array}\right]\end{array}\right]$$

(3.15)
$$_{f5}\left[\begin{array}{ll}\text{SUBJ} & _{f2}\left[\begin{array}{ll}\text{PRED} & \text{'ani; elder brother'}\\ \text{PRTADV} & \text{'-ga'}\end{array}\right]\end{array}\right]$$

(3.16)
$$_{f4}\left[\begin{array}{ll}\text{ADJ} & _{f3}\left[\text{PRED} \quad \text{'kono; this'}\right]\end{array}\right]$$

(3.17)
$$_{f5}\left[\begin{array}{ll}\text{OBJ} & _{f4}\left[\begin{array}{ll}\text{PRED} & \text{'hon; book'}\\ \text{PRTCS} & \text{'-wo'}\end{array}\right]\end{array}\right]$$

(3.18)
$$f5\begin{bmatrix} \text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{INFL} & \text{`-ta'} \\ \text{TENSE} & \text{past} \end{bmatrix}$$

The dependency relationships among these units combine these f-structure fragments to make up the f-structure for the sentence as a whole. The grammatical function of each unit is the label of the arc connecting the units in the dependency relationship. Figure 3.4 gives the DAG for (3.13) with arcs labelled with grammatical functions, and Figure 3.5 gives the f-structure constructed from the DAG in Figure 3.4:

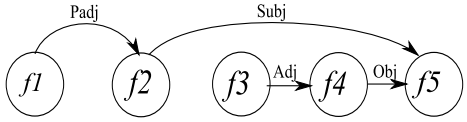

Figure 3.4: DAG for (3.13)

$$f5\begin{bmatrix} \text{SUBJ} & f2\begin{bmatrix} \text{PADJ} & \{ f1\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTCNJ} & \text{`-no'} \end{bmatrix} \} \\ \text{PRED} & \text{`ani; elder brother'} \\ \text{PRTADV} & \text{`-ga'} \end{bmatrix} \\ \text{OBJ} & f4\begin{bmatrix} \text{ADJ} & \{ f3\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\ \text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{INFL} & \text{`-ta'} \\ \text{TENSE} & \text{past} \end{bmatrix}$$

Figure 3.5: F-structure for (3.4)

It is just a matter of choice whether the syntactic tree for a sentence is a phrase-structure tree or a dependency tree; in either case, component f-structures which correspond to parts of a sentence are combined into one f-structure at the root level.

3.4 Topic and Zero Pronouns

One of the most distinctive features of Japanese is the prominence of Topic and the prolific use of zero pronouns. These two features are closely related with each other, so they are treated here in one section.

3.4.1 "Topic" in Japanese

Every Japanese grammar book deals with the distinction between the particles "wa" and "ga", sometimes in one full chapter. This is because their usage seems to be confusing to learners of Japanese and detailed explanation is required for a comprehensive understanding of them. Details aside, a syntactic unit which ends with the adverbial particle "wa" is the topic of the sentence, while a syntactic unit ending with the case particle "ga" is the subject of the sentence:

(3.19) Watashiwa kono honwo yonda

watashi-wa	*kono*	*hon-wo*	*yom-ta*
I-TOP	this	book-OBJ	read-decl.ta

"As for me, I read this book."

(3.20) Watashiga kono honwo yonda

watashi-ga	*kono*	*hon-wo*	*yom-ta*
I-SUBJ	this	book-OBJ	read-decl.ta

"It is I who read this book."

These two sentences denote the same event. What is different between them is the emphasis the speaker puts on the description of the event. In the first sentence, the speaker emphasises that he or she is talking about the referent of the topic unit, which happens to be the speaker himself. The literal translation of the first sentence will be like "As for myself, (I) read this book."[2] In the second sentence, on the other hand, the speaker emphasises that it is nobody but the referent of the subject unit who read the book. The literal translation of the second sentence will be like "It is I who read this book."

[2][Rubin, 1992] points out that Basil Hall Chamberlain, a nineteenth-century Japanologist, first used the phrase "as for" in [Chamberlain, 1907] for the analysis of Japanese topic.

In order to make the difference more conspicuous, let us put each of them into appropriate contexts. The first sentence will be an appropriate answer for a question which asks what the referent of the topic did. The dialog (3.21) in Japanese is a translation of a dialog (3.22), in which there are three persons A, B and C. A is asking B and C about what they did the day before:

(3.21)

A: Kinou naniwo shitemashitaka?
B: Watashiwa kono honwo yomimashita.
A: Anata wa?
C: Watashiwa tsurini ikimashita.

(3.22)

A: What did you do yesterday?
B: I read this book.
A: And you?
C: I went fishing.

The topic units "watashiwa" in B's utterance and C's utterance imply that each of the speakers sets the topic of the sentence by saying the topic unit first, and he or she talks about the topic and nothing else. By setting "watashi" or the first person singular pronoun as the topic of the sentence, the speaker expresses his or her intention that he or she is going to talk about the referent of the topic (it happens to be the speaker himself). In the examples above, both B's and C's utterances are simple, one-sentence utterances. The function of the particle "-wa" as a topic marker is basically the same in more complex, multi-sentence paragraphs. Therefore, the function of "-wa" is to set the topic of not only for what follows after it within the sentence, but also beyond the sentence boundary.

The second sentence "Watashi-ga kono hon-wo yonda" will be an appropriate answer for a question which asks who read the book, since the identification of the reader in this context is not the topic of the sentence; rather, it is a focus (or new information) in this context (the reason why the object unit "kono honwo" is parenthesised will be discussed in the next section):

(3.23)

A: Who read this book?
B: I did.

(3.24)

A: Darega kono honwo yomimashitaka?
B: Watashiga (kono honwo) yomimashita.

In other contexts, the following dialogue is also telling. In dialogues (3.25) and (3.26), Speaker A asks the identity of the reader of the book. Speaker B indicates the identity of the subject, with the particle "-ga" on the subject:

(3.25)
A: Is it Ken who read this book?
B: No, it is Naomi who did.

(3.26)
A: Kono honwo yondanowa Ken desuka?
B: Iie, Naomiga yomimashita.

Thus, the case particle "-ga" functions as an identifier of the subject of the verb, but not as a topic of the sentence.

Setting a certain topic by the particle "wa" sometimes implies a contrast between what is set as a topic and what is not. Therefore, in some contexts the literal translation of example (3.19) will be "I don't know what others did, but I read this book." or "I know others didn't read this book, but I did read it." In such cases, the "wa"-ending unit has the contrastive focus.

In some cases the topic "wa" and the contrastive focus "wa" appear in one sentence:

(3.27) Watashiwa kono honwa yonda

watashi-wa	*kono*	*hon-wa*	*yom-ta*
I-TOP	this	book-FOC	read-decl.ta

"I read this book, but not others."

In such cases, the sentence-initial "wa" unit is the topic, and the other "wa"-ending unit is the contrastive focus. In other words, in this case the unit order determines the grammatical function of these units. This is an instance of configurationality in a non-configurational language.

3.4.2 Topic as the Antecedent of a Zero Pronoun

There is no overt subject in sentence (3.19). The fact that there is no overt subject in the sentence does not mean that this sentence does not have a subject at all; rather, it is analysed as having a zero pronoun as the subject of the sentence. A zero pronoun refers to something which is already set in context. Recall that the function of the adverbial particle "-wa" is to set the topic of what follows after it within the sentence (or beyond). Therefore, the

topic of a sentence often can be the antecedent of the zero pronoun in the same sentence.

The more precise representation of sentence (3.19) as provided in (3.28) will be that the zero pronoun "pro" has the same index as the topic "watashi":

(3.28) $[Watashiwa]_i$ $[pro]_i$ kono honwo yonda.

The particle "-wa" indicates that in this context "watashi" is set as the topic of what follows. The grammatical function of this unit is topic, and nothing else. Apart from this topic, the presence of a zero pronoun is assumed from the fact that this sentence has no ga-marked unit. This zero pronoun must refer to what is available in this context, in this case "watashi" is the best candidate. Therefore, the topic is interpreted as the antecedent of a zero pronoun, which is the subject of the verb "yom-".

It is important to note that it is incorrect to assume that this topic is also the subject.

Zero pronouns can have grammatical functions other than subject. In the following sentence, the zero pronoun refers to the topic, and its grammatical function is object:

(3.29) $[Kono[honwa]]_i$ watashiga $[pro]_i$ yonda.

kono	*hon-wa*	*watashi-ga*	*pro*	*yom-ta*
this	book-TOPIC	I-SUBJ	pro-OBJ	read-decl.ta

"As for this book, I read it."

The particle "-wa" indicates that "kono honwa" is set in context as the topic of what follows. The grammatical function of this unit is topic, and nothing else. Apart from this topic, the presence of the zero pronoun is assumed from the fact that this sentence has no wo-marked unit, even though the verb is a transitive verb. This zero pronoun must refer to what is available in this context, in this case "kono honwa" is the best candidate. Therefore, the topic is interpreted as the antecedent of a zero pronoun, which is the object of the verb "yom-".

3.4.3 Topic Not as the Antecedent of a Zero Pronoun

Topic and zero pronouns are two independent phenomena: even though there is a preference that the topic of a sentence is the antecedent of a pro, and this topic and this pro depend on the same predicate, the existence of one of them in a sentence does not necessarily cause the existence of the other in the same sentence. One of the characteristic constructions involving a topic and a zero pronoun is the so-called "eel sentences" [Okutsu, 1978]:

(3.30) Bokuwa unagida.

boku-wa	*pro*	*unagi-da*
I(young.male)-TOP	pro	eel-copula.decl.plain.base

"I am an eel." or
"As for me, it is eel."

Literally, (3.30) means "I am an eel". But in different contexts it can mean something different. This is because the subject of the nominal predicate is a zero pronoun, and it is possible that this zero pronoun does not refer to the referent of the topic unit, but something which can be contextually appropriate as the subject of the nominal predicate.

The example above has two different interpretations. One is "I am an eel," in which the topic is the antecedent of the zero pronoun.

The other is "As for me, it is eel." The topic is not the antecedent of the zero pronoun, but functions as if it is a modifier of the sentence and the zero pronoun refers to something beyond this sentence. The context in which this interpretation is appropriate is this:

(3.31)

A: Sukina sakanawa nani?
B: Namazudesu.
A: Kimiwa?
C: Bokuwa unagida.

The English translation for (3.31) is (3.32):

(3.32)

A: What is your favorite fish?
B: It is catfish.
A: How about you?
C: As for me, it is eel.

Which of the interpretations for "Bokuwa unagida" should be chosen depends upon the semantic naturalness of the possible interpretations in a given context. In certain contexts such as fairy tales or cartoons, it is natural to say "I am an eel."

Eel sentences are not just for asking one's favorite dish or fairy tales, but are often used in more formal contexts.

3.4.4 Zero Pronouns without Reference to a Topic

A zero pronoun does not have to refer to a topic. An antecedent of a zero pronoun can be anything that is available in the context. Consider the following dialogues. The English dialogue (3.33) translates into Japanese as (3.34):

(3.33)
A: What did you do yesterday?
B: I read this book.

(3.34)
A: [pro] Kinou naniwo shiteimashitaka?
B: [pro] Kono honwo yondeimashita.

The utterances in (3.34) have neither topic nor overt subject. However, the Subject Condition[3] assumes the presence of a subject zero pronoun, and this must refer to something which is set in context. In this case, where A asks B about B's activities the day before, it is a matter of fact that A's and B's utterances are about B's activities. In other words, B's activities are set in context. Then, the zero pronouns in A's and B's utterances can refer to what has been set in context, even though this is not overtly expressed as the topic in these utterances.

In a dialogue involving two or three persons as in example (3.35), what is set in context is so obvious that it is often the case that topic units are omitted as follows:

(3.35)
A: Kinou naniwo shitemashitaka?
B: [pro] Kono honwo yomimashita.
A: Anata wa?
C: [pro] Tsurini ikimashita.

(3.36)
A: What did you do yesterday?
B: I read this book.
A: And you?
C: I went fishing.

[3][Alsina, 1996] states Subject Condition as follows:
An f-structure with propositional content must include a subject (as one of its grammatical functions) and no f-structure may include more than one subject.

In more complex situations, the presence of topic units is the key to identify the antecedent of zero pronouns. However, it is also common that zero pronouns refer to something beyond sentence boundaries, or something external to the text.

3.4.5 DAG Representation of Zero Pronouns

There are several possibilities for how zero pronouns can be represented by a DAG (cf. Section 3.3). One possibility is a coreference analysis, illustrated in the DAG in Figure 3.6 and in the f-structure in Figure 3.7. Figure 3.6 is the DAG for sentence (3.19) = (3.37):

(3.37) [Watashiwa]$_{f1}$ [pro]$_{f2}$ [kono]$_{f3}$ [honwo]$_{f4}$ [yonda]$_{f5}$.

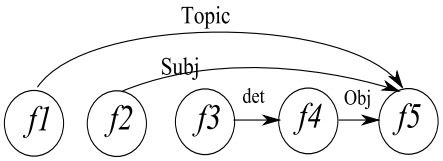

Figure 3.6: DAG for (3.27) in a coreference analysis

The f-structure generated from the DAG in Figure 3.6 is Figure 3.7:

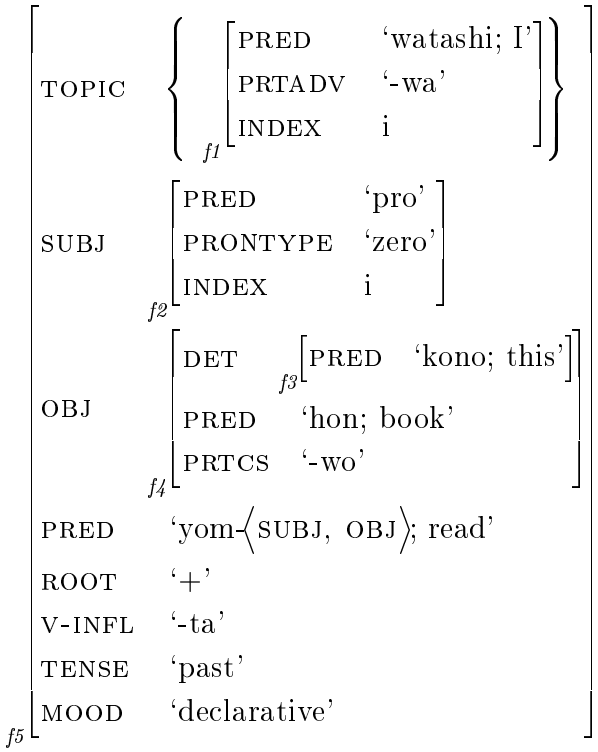

Figure 3.7: F-structure for Figure 3.6

In Figure 3.6, the presence of the zero pronoun is assumed from the fact that a verb must have a subject (Subject Condition). This zero pronoun refers to what has been set in context, in this case the topic "Watashi", meaning that this TOPIC is the antecedent of the SUBJ zero pronoun. The INDEX attributes in Figure 3.7 show that the TOPIC and the SUBJ zero pronoun refer to the same thing.

In the DAG analysis (Figure 3.6), there is no dependency relationship between a topic and a zero pronoun; both of them depend on verbal elements, but not on each other. In the f-structure analysis (Figure 3.7), the relationship between them is a type of coreference, and this is represented only at the level of f-structure; the DAG has nothing to say about the coreference relationship between the TOPIC and the SUBJ index features.

Another possibility is a filler-gap analysis, which assumes two arcs coming from TOPIC; one arc goes to the vertex which the TOPIC depends on, and the other arc (here represented as '=')goes to the SUBJ vertex:

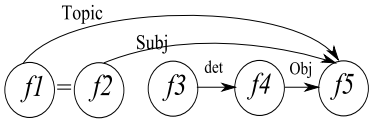

Figure 3.8: DAG for (3.27) in a filler-gap analysis

The f-structure generated from the DAG in Figure 3.8 is Figure 3.9:

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}[\,] \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f5}
$$

Figure 3.9: F-structure for Figure 3.8

In this analysis, there is a dependency relationship between a topic and a zero pronoun. The relationship between them is such that the subject is a gap and the topic is the filler for the gap.

I choose the coreference analysis throughout this thesis for the following reasons. First, the coreference analysis keeps one of the requirements of the DAG representation of Japanese dependency (cf. Section 3.3.2), namely that the outdegree of every vertex is one (in other words, every syntactic unit has only one target unit, or is connected by only one labelled arc), while the filler-gap analysis requires the outdegree of the TOPIC vertex to be more than one. Second, the filler-gap analysis contains "pred"-less core grammatical functions as gaps in f-structures, which are unnecessary in the coreference analysis. And the fact that zero pronouns can exist without reference to topic supports the idea that zero pronouns are not just gaps, but pronouns with their own PRED value. For these reasons, I choose the coreference analysis for topicalisation of Japanese.

3.4.6 Zero Pronoun Identification

Identifying the antecedent of a given zero pronoun is an important field of study in Japanese computational linguistics. Before doing this, it is of course necessary to detect the presence of zero pronouns in raw text as correctly as possible. If many zero pronouns which are assumed to be present in a text do not actually exist, then identifying their antecedents is a nonsense, however accurate the algorithm for antecedent-identification would be. On the other hand, if many of the zero pronouns in a text are ignored, then the accuracy of the antecedent-identification does not help much to improve the overall outcome. In addition, manually identifying all the zero pronouns in raw text is time-consuming and not an option for computational applications. It is desirable to identify zero pronouns automatically.

In this dissertation, zero pronoun identification for Japanese will be processed in the following steps (the basic idea of these steps are proposed in [Oya and van Genabith, 2007]): for each verb in the input sentence

Step 1. Check whether its subject is overtly expressed.
If it is, go to Step 2.
If it is not, then this verb's subject is a zero pronoun. Put a subject zero pronoun dependent on this verb in the sentence, then go to Step 2.

Step 2. Check the valence of the verb.
If it is an intransitive verb, then go to Step 6.
If it is not, then go to Step 3.

Step 3. Check whether its direct object is overtly expressed.
If it is, go to Step 4.
If it is not, then this verb's object is a zero pronoun. Put an object zero pronoun dependent on this verb in the sentence. Go to Step 4.

Step 4. Check whether this verb is a ditransitive or not.
If not, then go to Step 6.
If it is, then go to Step 5.

Step 5. Check whether its oblique is overtly expressed.
If it is, go to Stop.
If not, then this verb's oblique is a zero pronoun. Put an oblique zero pronoun dependent on this verb in the sentence, then go to Step 6.

Step 6. Stop.

It is obvious that the process above needs a lexical knowledge base providing valence information for Japanese verbs. Attempts have also been made to construct lexical knowledge bases automatically, as for example reported in [Kawahara and Kurohashi, 2004b], [Kawahara and Kurohashi, 2004a], [Kawahara and Kurohashi, 2005], etc.

In Chapter 6 of this thesis, I deal with the automatic identification of zero pronouns in detail. In particular, automatically generated f-structures for sentences taken from Kyoto Corpus are used as the linguistic resources for the identification of zero pronouns.

3.5 Inflecting Parts of Speech

3.5.1 Verbs

Verbs express actions or states. The last verb in a sentence functions as the head (or the root) of the sentence. The verb in a dependent clause functions as the head of that clause, and it depends on another verb.

Japanese verbs carry inflections expressing the tense, mood and other syntactic properties of the clause. There are two types of verbal inflections: Type-I and Type-II. The root form of Type-I verbs ends with a consonant, while the root form of Type-II ends with a vowel. Each Type is divided into subcategories according to the ending consonant or the ending vowel, which are not dealt with in greater detail here. Table 3.1 shows the inflection forms of a Type-I verb ("yom-", read) and of a Type-II verb ("mi-", see).

Table 3.1: Verbal inflections for "yom-" and "mi"

	Type-I		Type-II	
Moods	base forms	ta forms	base forms	ta forms
Declarative	yom-u	yom-ta(yonda)	mi-ru	mi-ta
Volitional	yom-o	N/A	mi-yo	N/A
Imperative	yom-e	N/A	mi-ro	N/A
Conditional	yom-eba	yom-tara(yondara)	mi-reba	mi-tara
Connective	yom-i	yom-te(yonde)	mi-	mi-te

There are two irregular verbs; "suru (do)" and "kuru (come)". They are irregular because their root forms change with inflections:

Table 3.2: Verbal inflections for "suru" and "kuru"

	suru		kuru	
Moods	base forms	ta forms	base forms	ta forms
Declarative	suru	shita	kuru	kita
Volitional	shiyo	N/A	koyo	N/A
Imperative	shiro	N/A	koi	N/A
Conditional	sureba	shitara	kureba	kitara
Connective	shi	shite	ki	kite

"Suru" is called the "sahen doushi". "Sahen" is an abbreviation for "*Sa*gyo *hen*kaku katsuyou". "Henkaku katsuyou" means "irregular inflection", and "sagyo" means that the last syllable of the root has the consonant 's'. "Doushi" meas "a verb".

The verb "suru" is often used with a noun to constitute a verbal unit. For example, the noun "kenkyuu (study)" is a noun by itself. When this noun is followed by "suru", then the compound constitutes a verbal unit:

(3.38) Watashiwa gengogakuya jinruigakuwo kenkyuushita.

watashi-wa	*gengogaku-ya*	*jinruigaku-wo*	*kenkyuu-shita*
I-TOP	linguistics-and	anthropology-OBJ	study-do.decl.ta

"I studied linguistics, anthropology, and others."

Not all Japanese nouns can be followed by "suru" to constitute verbal nouns. Nouns that can be followed by "suru" are called "sahen nouns".

Base Forms and Ta Forms

The declarative base form of a verb is its lemma, or the form registered in dictionaries. The difference between "base" and "ta" is concerned with the tenses which they express. The terms "present form" or "past form" are not used for them, because each of them can express different tense features according to the verb's lexical meaning.

Stative verbs in the base form denote a future state, a present state or a state which has continued until the present. Adverbial units disambiguate the tense feature:

(3.39) Watashiwa ieni iru.

watashi-wa *ie-ni* *ir-u*
I-TOP house-OBL exist-decl.base
"I am home."

(3.40) Watashiwa kono isshuukan zutto ieni iru.

watashi-wa *kono* *isshuukan* *zutto* *ie-ni*
I-TOP this one.week continuously house-OBL
ir-u
exist-decl.base
"I have been home for this one week."

(3.41) Watashiwa ashitamo ieni iru.

watashi-wa *ashita-mo* *ie-ni* *ir-u*
I-TOP tomorrow-FOC house-OBL exist-decl.base
"I will be home, too."

Stative verbs in the ta form denote a past state or a state which had continued until a certain point in the past:

(3.42) Watashiwa kinou ieni ita.

watashi-wa *kinou* *ie-ni* *ir-ta*
I-TOP yesterday house-OBL exist-decl.ta
"I was home yesterday."

(3.43) Watashiwa kyouno gozenmade ieni ita.

watashi-wa *kyou-no* *gozen-made* *ie-ni* *ir-ta*
I-TOP today-of morning-until house-OBL exist-decl.ta
"I had been home until this noon." or
"I was home this morning."

Action verbs in the base form denote an action or an event in the future.

(3.44) Watashiwa korekara kono honwo yomu.

watashi-wa *kore-kara* *kono* *hon-wo* *yom-u*
I-TOP from.now this book-OBJ read-decl.base
"I will read this book from now."

Action verbs in the ta-form denote an action or an event in the past.

(3.45) Watashiwa gakuseino koro kono honwo yonda.

watashi-wa	*gakusei-no*	*koro*	*kono*	*hon-wo*	*yom-ta*
I-TOP	student-of	time	this	book-OBJ	read-decl.ta

"I read this book when I was a student."

The present tense of an action verb is expressed morphologically with a suffix "-teiru" which also expresses the progressive aspect.

(3.46) Watashiwa kono honwo yondeiru.

watashi-wa	*kono*	*hon-wo*	*yom-teir-u*
I-TOP	this	book-OBJ	read-suff.-decl.base

"I am reading this book."

The perfect aspect of an action verb is expressed morphologically with a suffix "-oeru", or other constructions.

(3.47) Watashiwa kono honwo yomioeta.

watashi-wa	*kono*	*hon-wo*	*yomi-oer-ta*
I-TOP	this	book-OBJ	read-suff.-decl.ta

"I have read through this book."

Both stative verbs and action verbs in the base form can have a generic meaning which specify the subject's characteristics.

(3.48) Kinbenna gakuseiwa ookuno honwo yomu.

kinben-na	*gakusei-wa*	*ooku-no*	*hon-wo*	*yom-u*
diligent-attr.base	student-TOP	many-of	book-OBJ	read-decl.base

"Diligent students read a lot of books."

The ta forms of Type-I verbs show phonological changes; for example, "yom+ta" is actually pronounced as "yonda". In this thesis, the phonological changes of verbs are indicated by the parenthesised form next to the base form. For the phonological details of Japanese, see [Tsujimura, 2006].

Declarative Forms

Declaratives are the most basic, plain form of verbs.

(3.49) Watashiwa mainichi issatsu honwo yomu.

Watashi-wa	*mainichi*	*issatsu*	*hon-wo*	*yom-u*
I-TOP	every.day	one.book	book-OBJ	read-decl.base

"I read one book every day."

(3.50) Watashiwa gakuseino koro mainichi issatsu honwo yonda.

Watashi-wa	*gakusei-no*	*koro*	*mainichi*	*issatsu*
I-TOP	student-particle	when	every.day	one.book
hon-wo	*yom-ta.*			
book-OBJ	read-decl.PST			

"When I was a student, I used to read one book a day."

Declaratives do not have the progressive meaning; the progressive aspect is indicated by the connective form followed by the verbal suffix "-iru" (cf. Section 3.3.3);

(3.51) Watashiwa ima honwo yondeiru.

Watashi-wa	*ima*	*hon-wo*	*yom-te-ir-u*
I-TOP	now	book-OBJ	read-conn.ta-suff.-decl.base

"Now I am reading a book."

(3.52) Watashiwa gakuseino koro chomusukiiwo yondeita.

Watashi-wa	*gakusei-no*	*koro*	*chomusukii-wo*
I-TOP	student-particle	when	Chomsky-OBJ
yom-te-ir-ta			
read-conn.ta-suff.-decl.ta			

"When I was a student, I was reading Chomsky."

Declaratives are also used for the main predicate of a relative clause. Since Japanese does not have relative pronouns, the inflection of the verb is the clue to identify the clause boundary of a relative clause (cf. Section 3.6.5). If a verb in declarative form appears before a noun, then the verb is the head of the relative clause modifying the noun.

(3.53) Watashiga yonda hon.

Watashi-ga	*yom-ta*	*hon*
I-SUBJ	read-decl.ta	book

"(the) book I read"

Verbs in the declarative form can be followed by an auxiliary:

(3.54) Watashiwa kono honwo yomutsumorida.

Watashi-wa	*kono*	*hon-wo*	*yom-u-tsumorida*
I-TOP	this	book-OBJ	read-decl.base-AUX.decl.base

"I am going to read this book."

(3.55) Watashiwa kono honwo yondabakarida.

Watashi-wa	*kono*	*hon-wo*	*yom-ta-bakarida*
I-TOP	this	book-OBJ	read-decl.ta-AUX.decl.base

"I have just finished reading this book."

Volitional Forms

Verbs in the volitional base form express the will of the speaker, or invitation for an action. The choice between them depends on the subject: if the subject is the first person, then it is the expression of the will of the speaker, while it is an invitation if the subject is not the first person.

(3.56) Watashiwa kono honwo yomo.

Watashi-wa	*kono*	*hon-wo*	*yom-o*
I-TOP	this	book-OBJ	read-vol.base

"I will read this book."

(3.57) Ashita eigani iko.

Ashita	*eiga-ni*	*ik-o*
tomorrow	movie-OBL	go-vol.base

"Let's go to a movie tomorrow."

In the second sentence in the examples above, the subject is a zero pronoun. There are many instances of zero pronouns in Japanese utterances and texts, and the interpretation depends on the context.

Imperative Forms

Verbs in the imperative form have the imperative mood. Imperatives do not have the ta form:

(3.58) Kono honwo yome.

Kono *hon-wo* *yom-e.*
This book-OBJ read-imp.base
"Read this book."

Verbs in the imperative form by themselves sound rather rude and they are avoided in actual use. The more friendly style of imperative is the connective ta form, and the politeness increases with the verbal suffix "-kudasaru" in its connective base form. The politeness further increases with the verbal suffix "-masu" after "-kudasaru", and the "-masu" is in the imperative form:

(3.59) Kono honwo yonde.

Kono *hon-wo* *yom-te*
this book-OBJ read-conn.ta
"Why don't you read this book?"

(3.60) Kono honwo yondekudasai.

Kono *hon-wo* *yom-te-kudasai*
this book-OBJ read-conn.ta-suff.conn.base
"Please read this book."

(3.61) Kono honwo yondekudasaimase.

Kono *hon-wo* *yom-te-kudasai-mase*
this book-OBJ read-conn.ta-suff.conn.base-suff.imp.base
"Would you please read this book?"

It is possible (and required in certain contexts) to further increase the politeness by adding other elements.

Conditional Forms

Verbs in the conditional have conditional mood. In the examples below, since the subject of the verb "read" is zero-pronominalised and its identity depending on the context, it also depends on the context whose life changes:

(3.62) Kono honwo yomeba jinseiga kawaru.

Kono	*hon-wo*	*yom-eba*	*jinsei-ga*	*kawar-u.*
This	book-OBJ	read-cond.base	life-SUBJ	change-decl.base

"If one reads this book, (his or her) life will change."

(3.63) Kono honwo yondara, jinseiga kawatta.

Kono	*hon-wo*	*yom-tara,*	*jinsei-ga*	*kawar-ta*
This	book-OBJ	read-cond.ta	life-SUBJ	change-decl.ta

"After reading this book, (my) life changed."

The difference of meaning between the conditional base form and the conditional ta form is not temporal, but propositional; the conditional base form implies that the action of the verb actually has not taken place, while the conditional ta form does not have any such implication;

(3.64) Kono honwo yomeba jinseiga kawatta.

**Kono*	*hon-wo*	*yom-eba*	*jinsei-ga*	*kawar-ta*
This	book-OBJ	read-cond.base	life-SUBJ	change-decl.ta

(3.65) Kono honwo yondara jinseiga kawaru.

Kono	*hon-wo*	*yom-tara*	*jinsei-ga*	*kawar-u.*
This	book-OBJ	read-cond.ta	life-SUBJ	change-decl.base

"Life will change after reading this book."

This property is also the reason why the terms "present form" or "past form" are not used for these forms.

Connective Forms

The function of connectives is to show that a verb in this form is the main predicate of a subordinate clause.

(3.66) Kono honwa yomi, ano honwa suteru.

Kono *hon-wa* *yom-i,* *ano* *hon-wa*
This book-TOP read-conn.base that book-TOP
sute-ru
throw.away-decl.base
"I will read this book, while I will throw away that book."

(3.67) Kono honwa yonde, ano honwa suteta.

Kono *hon-wa* *yom-te,* *ano* *hon-wa*
This book-TOP read-conn.ta, that book-OBJ
sute-ta
throw.away-decl.ta
"I read this book and threw away that book."

A verb in this form can be followed by a verbal suffix:

(3.68) Watashiwa kono honwo yomitsuzuketa.

Watashi-wa *kono* *hon-wo* *yom-i-tsuzuke-ta*
I-TOP this book-OBJ read-conn.base-SUF-decl.ta
"I kept on reading this book"

3.5.2 Transitive-Intransitive Pairs

Some verbs constitute Transitive-Intransitive pairs. These are verbs that share the basic meaning, and the difference of their ending forms indicates the transitivity or intransitivity of the verb. Consider the examples below; both of the verbs "ak-u" and "ake-ru" are concerned with the action of opening a door, and "ak-u" is an intransitive verb with the door being its subject, while "ake-ru" is a transitive verb with the door being its object:

(3.69) Doaga aita.

doa-ga *ak-ta*
door-SUBJ open-decl.ta
"The door opened."

(3.70) Anega doawo aketa.

ane-ga *doa-wo* *ake-ta*
sister-SUBJ door-OBJ open-decl.ta
"(My) sister opened the door."

The verbs which constitute a transitive-intransitive pair are two different verbs, which basically express the same event, but differ in terms of whether the event should be expressed transitively or intransitively. Notice that the two verbs do not share the root form; the root form of the intransitive is "ak-", while the root form of the transitive is "ake-". This fact suggests that the transitive-intransitive pairs are not motivated by morphological operations (such as conditional forms, connective forms mentioned above) to derive different inflectional realisations of one root verb. This makes a good contrast with English, in which a transitive-intransitive pair can be made by two completely identical verbs, such as "open" (intransitive) and "open" (transitive):

[Yoshikawa, 1989] lists the morphological types of transitive-intransitive pairs:

Table 3.3: Transitive-Intransitive Pairs

	Intransitive ending	Transitive ending	Examples	
1	-ARU	-U	husag*ar-u*	husag-*u*
2	-ARU	-ERU	ag*ar-u*	ag*e-ru*
3	-U	-ERU	ak-*u*	ak*e-ru*
4	-ERU	-U	tor*e-ru*	tor-*u*
5	-ERU	-ASU	nur*e-ru*	nur*as-u*
6	-RERU	-SU	tao*re-ru*	tao*s-u*
7	-U	-ASU	kawak-*u*	kawak*as-u*
8	-IRU	-ASU	nob*i-ru*	nob*as-u*
9	-IRU	-OSU	och*i-ru*	ot*os-u*
10	-RU	-SU	noko*r-u*	noko*s-u*
11	-RU	-SERU	no*r-u*	no*se-ru*
12	-IERU	-ESU	k*ie-ru*	k*es-u*

The first column lists the intransitive endings, and the second column the transitive endings. Intransitive and transitive endings are capitalised. The third column shows an example verb for each of the intransitive endings, and the fourth shows an example verbs for each transitive ending. The intransitive or transitive endings in these example verbs are italicised.

Notice that the place of the hyphens in the intransitive and transitive endings and the place of hyphens in the examples are different. This stresses the fact that the boundary between intransitive or transitive ending and the rest of the verb does not necessarily correspond to the boundary between the root and its inflection.

Table 3.3 shows that: for the verbs in the transitive-intransitive pairs,

a. a verb is a transitive if it ends with "SU" or "SERU" (types 5 to 12 in Table 3.3);

b. a verb is a transitive if it ends with "-U" or "-ERU" and if it constitutes a transitive-intransitive pair with a verb which ends with "-ARU" (types 1 to 2 in Table 3.3);

c. a verb is a intransitive if it ends with '-U" or "-ERU" and if it constitutes a transitive-intransitive pair with a verb which ends with "-ASU" (type 5 and type 7 in Table 3.3);

d. if a verb has the ending "-U" and it constitutes a transitive-intransitive pair with a verb which ends with "-ERU", it is ambiguous (types 3 and 4 in Table 3.3).

When we have to determine the valence of verbs in a given context, the first step is to focus on the verbs which end with "-U" and "-ERU". If one of the verbs belongs to type 3 or type 4, the lemma forms of these verbs do not help to determine their valence. We need to use some measures other than morphology. I will return to this in the context of zero-pronoun identification in Section 6.6.3.

Not all Japanese verbs belong to transitive-intransitive pairs; some verbs are transitive without intransitive counterparts (no-counterpart transitives), while some others are intransitive without transitive counterparts (no-counterpart intransitives). Morphology does not help much to identify which verb is a no-counterpart transitive or a no-counterpart intransitive.

Still, we can say that verbs ending with "su" tend to have transitivity; actually, the majority of "su" ending verbs are transitive verbs without their intransitive counterparts: checking 628 "-SU" ending verbs in the JUMAN dictionary, I found that 609 of them are transitive.

3.5.3 Adjectives

Attributive adjectives express the characteristics of, or a subjective judgement on, a noun. Japanese adjectives have inflections and they need no

copula in their predicative use. Japanese adjectives are classified into two categories according to their inflection forms: i-adjectives and na-adjectives. I-adjectives (pronounced "ee-adjectives") end with "-i" both in the attributive and predicative present tense, and na-adjectives end with "-na" in the attributive usage and "-da" in the predicative present tense.

Table 3.4 gives the inflection forms of an i-adjective "akai":

Table 3.4: Inflection forms of the i-adjective "akai"

moods	base forms	ta forms
declarative	aka-i	aka-katta
conditional	aka-kereba	aka-kattara
adverbial	aka-ku	aka-kute, aka-kattari

The examples below show the attributive and predicative uses of an adjective "akai(red)". The attributive (in (3.71)) and the predicative (in (3.72)) have the same form:

(3.71) akai hana

aka-i *hana*
red-decl.base flower
"a red flower"

(3.72) kono hanawa akai

kono *hana-wa* *aka-i*
this flower-TOP flower-decl.base
"This flower is red."

The conditional form expresses the conditional mood. This is used in the predicative only, and it is not used as the head of the main clause:

(3.73) Kono hanaga akakereba yokattanoni

kono *hana-ga* *aka-kereba* *yo-katta-noni*
this flower-SUBJ red-cond.base good-decl.ta-particle
"If this flower had been red, it would have been better."

The adverbial form of an adjective is used when the adjective depends on another inflecting element, or it is followed by an adjectival or a verbal suffix. Adjectives in the adverbial form function as adverbs:

(3.74) akaku saita hana

aka-ku *saki-ta* *hana*
red-adv.base bloom-decl.ta flower
"a flower which bloomed red."

(3.75) akakute ookii hana

aka-kute *ooki-i* *hana*
red-adv.ta big-decl.base flower
"a red big flower."

(3.76) kono hanawa akakunai

kono *hana-wa* *aka-ku-nai*
this flower-TOP red-adv.base-NEG.decl.base
"This flower is not red."

There are three systems for na-adjectives: plain, formal and polite. Table 3.5 shows the inflection forms of na-adjectives in the plain style:

Table 3.5: Inflection forms of the na-adjective "kireida (beautiful)" in the plain style:

moods	base forms	ta forms
declarative	kirei-da	kirei-datta
conditional	N/A	kirei-dattara
adverbial	kirei-ni	kirei-de, kirei-dattari
attributive	kirei-na	N/A

The examples below involve the na-adjective "kireida (beautiful)":

(3.77) kireina hana

kirei-na *hana*
beautiful-att.base.PLAIN flower
"a beautiful flower"

(3.78) kono hanawa kireida

kono *hana-wa* *kirei-da*
this flower-TOP beautiful-decl.base.PLAIN
"This flower is beautiful."

The three-way system of inflections of na-adjectives resembles those of the copula "-da". This fact leads some researchers to decompose a na-adjective into an "adjectival noun" and a copula. According to this analysis, the na-adjective "kireida" consists of an na-adjective "kirei" and a copula "-da". The "adjectival noun" analysis focuses on the fact that na-adjectives in predicative use are morphologically similar to "noun+copula" construction. This construction denotes the identity of the referent of the subject:

(3.79) Kenga kono chiimuno riidaada

Ken-ga *kono* *chiimu-no* *riidaa-da*
Ken-SUBJ this team-particle leader-copula.decl.base.PLAIN
"Ken is the leader of this team."

In the example above, the referent of the noun followed by the copula (the leader) is identical to the referent of the subject (Ken). On the other hand, the referent of the "adjectival noun" followed by the copula ("kirei" in the example) is not identical to the referent of the subject, but to an attribute of the subject. Morphological similarity between na-adjectives and noun+copula construction does not lead to semantic similarity. Besides, in many cases the "adjectival noun" cannot be used without the copula. If "kireida", for example, is used without "-da" and is followed by a case particle, it sounds odd, even though not ungrammatical. Therefore, it is better to analyse a na-adjectives as one word, rather than dividing them into a noun and a copula.

3.5.4 Adjectival and Verbal Suffixes

The function of adjectival and verbal suffixes in Japanese is to show various grammatical features such as voice, aspect, tense and mood. For example, the verbal suffix '-iru' is used to indicate progressive aspect:

(3.80) Watashiwa kono honwo yomu

watashi-wa	*kono*	*hon-wo*	*yom-u*
I-TOP	this	book-OBJ	read-decl.base

"I read this book."

(3.81) Watashiwa kono honwo yondeiru

watashi-wa	*kono*	*hon-wo*	*yom-te-iru*
I-TOP	this	book-OBJ	read-conn.ta-suff.decl.base

"I am reading this book."

The verbal suffix "-aru" shows that the verb's action has been completed:

(3.82) Watashiwa kono honwo yondearu

watashi-wa	*kono*	*hon-wo*	*yom-te-aru*
I-TOP	this	book-OBJ	read-conn.ta-suff.decl.base

"I have already read this book."

There are a variety of adjectival and verbal suffixes with different meanings (the dictionary of JUMAN registers 52 of them). Since I cannot describe all of them in detail, here I present some of them which are relevant to the discussions later in this thesis.

Negation

Negation of a verb or an adjective is expressed by the negative suffix "-(a)nai". Type-I verbs (verbs whose root ends with a consonant) are followed by "-anai", while Type-II verbs (verbs whose root ends with a vowel) and adjectives are followed by "-nai":

(3.83) Watashiwa kono honwo yomanai.

watashi-wa	*kono*	*hon-wo*	*yom-ana-i*
I-TOP	this	book-OBJ	read-NEG-decl.base

"I don't read this book."

(3.84) Watashiwa sono eigawo minai.

watashi-wa *sono* *eiga-wo* *mi-na-i*
I-TOP the movie-OBJ watch-NEG-decl.base
"I don't watch the movie."

The inflections of "-(a)nai" are the same as those of i-adjectives. JUMAN calls this and other suffixes with the same inflections "adjectival predicate suffixes".

Causatives

Causative voice is expressed by the causative suffixes "-aser-" and "-saser-". Type-I verbs are followed by "-aseru", Type-II verbs are followed by "-saseru".

(3.85) Watashiwa ototoni kono honwo yomaseta.

watasi-wa *ototo-ni* *kono* *hon-wo* *yom-ase-ta*
I-TOP yonger.brother-OBL this book-OBJ read-caus.-decl.ta
"I made my younger brother read this book."

(3.86) Watashiwa ototoni kono eigawo misaseta

watashi-wa *ototo-ni* *kono* *eiga-wo*
I-TOP younger.brother-OBL this movie-OBJ
mi-sase-ta
watshi-caus.-decl.ta
"I made my younger brother watch this movie."

Causative suffixes inflect as Type-II verbs; the roots of both of the causative suffixes end with a vowel. They can be followed by further suffixes.

Causative suffixes change the number of arguments the verb can take; in the example above, the transitive verb "yom-" comes to have three arguments. The change can be illustrated by the subcategorisation frames:

(3.87) yom-<SUBJ, OBJ>
agent theme

(3.88) yom-ase-<SUBJ, OBL, OBJ>
agent causee theme

The SUBJ in the transitive "yom-" is the agent of the action, and the OBJ is the theme of the action. In the causative "yom-ase-", the SUBJ is the agent of the causing action, and the OBL is the causee, or the agent of the caused action, corresponding to the SUBJ of the transitive "yom-".

Some researchers, for example [Matsumoto, 1996], argue that Japanese causatives can be either biclausal or monoclausal. "Biclausal" means that the causative suffix is the head of the main clause, and the verbal root is the head of the subordinate clause. "Monoclausal" means that the suffixed verbal root is the head of the main clause. [Matsumoto, 1996] also argues that the choice of biclausal and monoclausal reflects the difference in the strength of causation, pointing out that permissive causative sentences show biclausality, while coercive causative sentences do not. Considering that the argument in favour of biclausality is based on intuitive judgement, I do not adopt this analysis but instead assume that a verb with a causative suffix is always one word, derived through a lexical rule.

Passives

Passive voice is expressed by the passive suffixes "-areru" or "-rareru". Both follow the root of a verb, and they inflect as Type-II verbs.

Passive voice of Japanese is categorized into direct passive and indirect passive (or adversative passive); direct passive is the passive in which the object of a verb in the active voice turned into the subject, and the subject of the verb in the active voice is expressed by the oblique-case noun phrase. Only transitive verbs can be in the direct passive:

(3.89) Ototoga watashino tsubowo kowashita.

Ototo-ga	*watashi-no*	*tsubo-wo*	*kowas-ta*
younger.brother-SUBJ	I-of	vase-OBJ	break-decl.ta

"My younger brother broke my vase."

(3.90) Watashino tsuboga ototoni kowasareta.

watashi-no	*tsubo-ga*	*ototo-ni*	*kowas-are-ta*
I-of	vase-SUBJ	younger.brother-OBJ	break-pass.-decl.ta

"My vase was broken by my younger brother."

The mapping between thematic roles and grammatical functions changes as follows:

(3.91) kowas-<SUBJ, OBJ>
agent theme

(3.92) kowas-arer-<SUBJ, OBL>
theme agent

Indirect passive, on the other hand, expresses that the referent of the subject is influenced by, or suffered from the action or the result of the action caused by the referent of the oblique. The change can be illustrated by the subcategorisation frames. Notice that the "agent" in both frames refers to the same individual:

(3.93) kowas-<SUBJ, OBJ>
agent theme

(3.94) kowas-arer-<SUBJ, OBL, OBJ>
patient agent theme

(3.95) Ototoga watashino tsubowo kowashita.

Ototo-ga	*watashi-no*	*tsubo-wo*	*kowas-ta*
younger.brother-SUBJ	I-of	vase-OBJ	break-decl.ta

"My younger brother broke my vase."

(3.96) Watashiwa tsubowo ototoni kowasareta.

watashi-wa	*tsubo-wo*	*ototo-ni*	*kowas-are-ta*
I-TOP	vase-OBJ	younger.brother-OBL	break-pass.-decl.ta

"I suffer from my younger brother's breaking my vase."

Unlike direct passive, intransitive verbs can be in the indirect passive.

(3.97) Imotoga naita.

imoto-ga	*nak-ta*
younger.sister-SUBJ	weep-decl.ta

"My sister wept."

(3.98) Watashiwa imotoni nakareta.

watashi-wa	*imoto-ni*	*nak-are-ta*
I-TOP	younger.sister-OBL	weep-pass.-decl.ta

"I suffer from my sister's weeping."

The subcategorisation frames below illustrate the indirect passivisation of an intransitive verb:

(3.99) nak-<SUBJ>
agent

(3.100) nak-arer-<SUBJ, OBL>
patient agent

As in the case of causative suffixes, passive suffixes change the valence of the verb, and a verb with a passive suffix is one word. In the case of the indirect passive, the SUBJ is mapped onto the thematic role patient, not onto the agent.

Benefactives

Benefactive voice is one of the characteristics of Japanese. It expresses the speaker's subjective judgement on the benefit of the action. This voice is closely related to the honorific system of Japanese, which basically focuses on where the benefit of a given action comes from; if it comes from someone with high status, then it is expressed as if it "comes down". For more detail on the Japanese honorific system, see [Shibatani, 1990] and [Tsujimura, 2006].

Benefactive suffixes are in charge of expressing benefactive voice. For example, the suffix "-kureru" emphasises the speaker's gratitude that the action by the subject is beneficial to the speaker. This suffix shares the form with the full verb "kureru", which means 'give':

(3.101) Anega watashino shukudaiwo tetsudatta.

ane-ga	*watashi-no*	*shukudai-wo*	*tetsudaw-ta*
elder.sister-SUBJ	I-of	assignment-OBJ	help-decl.ta

"My elder sister helped my assignment."

(3.102) Ane-ga watashino shukudaiwo tetsudattekureta.

ane-ga	*watashi-no*	*shukudai-wo*
elder.sister-SUBJ	I-of	assignment-OBJ

tetsudaw-te-kure-ta
help-conn.ta-give-decl.ta

"My elder sister gave me the benefit of her helping my assignment."

The suffix "-morau", on the other hand, emphasises the speaker's gratitude that the action by the referent of the oblique is beneficial to the speaker, who is expressed by the topic. This suffix shares the form with the full verb "morau", which means 'receive':

(3.103) Watashiwa shukudaiwo aneni tetsudattemoratta.

watashi-wa *shukudai-wo* *ane-ni*
I-TOP assignment-OBJ elder.sister-OBL
tetsudaw-te-moraw-ta
help-conn.ta-receive-decl.ta
"I received (from my elder sister) the benefit of her helping my assignment."

These three sentences express the same event, but they show differences in the speaker's attitudes toward the event. The first sentence is just a description of the event. The second expresses that the benefit is directed from the subject's action to the speaker. The third expresses that the benefit comes from the oblique's action to the topic, which happens to be the speaker in this example.

As in the cases of causative suffixes and passive suffixes, the benefactive suffix "-morau" changes the valence of the verb, and a verb with this benefactive suffix is analysed as one word.

"Possibility" or "Spontaneity"

"Possibility" or "spontaneous" suffixes change a transitive verb into an intransitive verb, and the object of the transitive verb remains the subject, while the agent of the verb's action is deleted or topicalised:

(3.104) Kokokara Fujisanga mieru.

koko-kara *Fujisan-ga* *mi-eru*
here-from Mt.Fuji-SUBJ see-suff.decl.base
"One can see Mt.Fuji from here."

In example (3.104) above, the verb does not refer to a particular event of someone seeing Mt.Fuji, but the state that there is a possibility of anyone seeing Mt.Fuji from there. The literal meaning of this will be "Mt.Fuji is in the state of being possible to be seen from here." No agent of seeing Mt.Fuji is presupposed by this sentence. The voice specified by this verbal suffix is

called “jihatsu tai (spontaneous voice)” or “kanou tai(possible voice)”. These two voices are distinguished by semantics rather than morphology, as they are both expressed by the same suffix “-eru”.

If the agent of the event must somehow be expressed with a “verbal root + possible/spontaneous suffix” verbal unit, a topicalised noun unit serves this purpose. In this case, the voice will be “possible”, rather than spontaneous:

(3.105) Watashiwa kono honga yomeru.

watashi-wa	*kono*	*hon-ga*	*yom-eru*
I-TOP	this	book-SUBJ	read-suff.decl.base

“As for me, this book is readable.” or
“I can read this book.”

In example 3.105, the verb does not refer to a particular event of reading the specific book, but to the state that the referent of the topic is in the state of being able, or having the potential, to read the book. The literal meaning of this sentence will be “As for me, this book is in the state of being able/possible to be read.”

3.5.5 Copulas

Copulas come after a noun to make a nominal predicate. The functions of nominal predicates are almost equivalent to the “be” copula of English accompanied with an NP.

(3.106) Watashiwa gakuseida.

watashi-TOP	*gakusei-da*
I-TOP	student-Copula.plain.decl.base

“I am a student.”

(3.107) Korega kotaedearu.

kore-ga	*kotae-dearu*
This-SUBJ	answer-Copula.formal.decl.base

“This is the answer.”

(3.108) Shachowa kaigichuudesu.

shacho-wa *kaigi-chu-desu*
manager-TOP meeting-during-Copula.polite.decl.base
"The manager is at the meeting now."

The inflections for copulas are almost the same as those of na-adjectives. The only difference is the plain attributive base form.

3.5.6 Auxiliaries

Auxiliaries in Japanese express various moods. Auxiliaries follow a verb, a verbal suffix, or an i-adjective in the declarative, and a na-adjective or a copula in the attributive:

(3.109) Kenwa kono honwo yomubekida.

Ken-wa *kono* *hon-wo* *yom-u-bekida*
Ken-TOP this book-OBJ read-decl.base-AUX
"Ken must read this book."

(3.110) Kenwa Naomini kono honwo yomaserubekida.

Ken-wa *Naomi-ni* *kono* *hon-wo*
Ken-TOP Naomi-OBL this book-OBJ
yom-ase-ru-bekida
read-cause-decl.base-AUX
"Ken must have Naomi read this book."

(3.111) Kono tenga hushizennanoda.

kono *ten-ga* *hushizen-na-noda*
this point-SUBJ unnatural-attr.base-AUX
"It is this point that is unnatural."

(3.112) Korega kotaenanoda.

kore-ga *kotae-na-noda*
this-SUBJ answer-copula.attr.base-AUX
"This IS the answer."

The inflection forms of auxiliaries are mainly those of na-adjectives, but some auxiliaries have their particular inflection forms.

3.6 Non-Inflecting Parts of Speech

3.6.1 Nouns

Nouns refer to entities, either physical or abstract, and they function as an argument or an adjunct of a verb. In Japanese, a noun is the head of a noun unit, and its grammatical function is determined by the type (or absence) of the particle right after the head noun (cf. Section 3.6.2). Nouns can be classified into categories according to their meanings, or according to their morpho-syntactic properties. This section focuses on part-of-speech alternations of nouns derived from verbs and adjectives, formal nouns which are used as complementisers or postpositions, and types of classifiers for nouns.

Nouns Derived from Verbs and Adjectives

This section introduces some instances of both lexicalised and productive nominalisations of verbs and adjectives. Some nouns are derived from other parts of speech such as verbs and adjectives, and sometimes they are listed as nouns in dictionaries. On the other hand, there are some productive ways of nominalisations of verbs and adjectives, which are not listed as nouns in dictionaries. When a morphological analyser outputs these nominalised verbs just as verbs in an NLP application, then this leads to incorrect analyses. Therefore, nominalisation of inflecting parts of speech is relevant to natural language processing of Japanese, including my study.

Some verbs in their connective forms function as nouns, and are listed as nouns in dictionaries. In the example below, the noun "yomi" is derived from the connective form of a verb "yomu", and this noun means something like "prediction of the future according to the speculation on the current situation":

(3.113) Ano seijikaha yomiga amakatta.

ano	*seijika-wa*	*yomi-ga*	*amak-atta*
that	politician-TOP	reading-SUBJ	sweet-decl.ta

lit. "As for the politician, (his) reading was sweet." or "That politician predicted the situation wrong."

A more productive way of nominalisation of verbs consists of a verb followed by a nominal suffix such as "kata" (meaning "way" or "process"):

(3.114) Ano gakuseino yomikatawa okashikatta

ano *gakusei-no* *yomi-kata-wa* *okashi-katta*
that student-of read-way-TOP funny-decl.ta
"The student's way of reading was funny."

The "connective + kata" construction is so productive that virtually every verb can be used in this construction. The meaning of the simple nominalised connective form of a verb and that of the "connective + kata" nominalisation construction are different: in general, the former refers to the result of the action of the verb, while the latter refers to the process of the action of the verb.

Adjectives are nominalised by some nominal suffixes. One of the examples is "-sa" which follows the root of both an i-adjective and a na-adjective:

(3.115) Kono heyawa akarui

kono *heya-wa* *akaru-i*
this room-TOP bright-decl.base
"This room is bright."

(3.116) Kono heyano akarusawo hakatta

kono *heya-no* *akaru-sa-wo* *hakar-ta*
this room-of bright-SUFF-OBJ measure-decl.base
"(Pro) measured the brightness of this room."

(3.117) Watashino anewa kireida

watashi-no *ane-wa* *kirei-da*
I-of sister-TOP beautiful-decl.base
"My sister is beautiful."

(3.118) Watashino aneno kireisani Kenwa odoroita

watashi-no *ane-no* *kirei-sa-ni* *Ken-wa*
I-of sister-of beautiful-SUFF-OBL Ken-TOP
odorok-ta
surprise-decl.ta
"Ken was surprised by my sister's beauty."

Formal Nouns

Formal nouns have only abstract, idiomatic meaning and are always modified by a clause or a postpositional phrase. They function like an English complementiser "that". The presence of a formal noun in syntax is more important than its meaning. In the example below, the formal noun "koto" takes a clause and is followed by a particle:

(3.119) Kenwa Naomiga rikonshitakotowo shiranakatta.

Ken-wa *[Naomi-ga* *rikonshi-ta]-koto-wo*
Ken-TOP [Naomi-SUBJ divorce-decl.ta]-that-OBJ
shir-ana-katta
know-NEG-decl.ta
"Ken didn't know that Naomi had divorced."

The formal noun "no" can appear in the same construction:

(3.120) Kenwa Naomiga rikonshitanowo shiranakatta.

Ken-wa *[Naomi-ga* *rikonshi-ta]-no-wo*
Ken-TOP [Naomi-SUBJ divorce-decl.ta]-that-OBJ
shir-ana-katta
know-NEG-decl.ta
"Ken didn't know that Naomi had divorced."

Some verbs exclusively choose "no", and others prefer "koto", but in many cases they are interchangeable. [Kuno, 1983] argues that "koto" introduces an abstract idea or knowledge, and "no" introduces a concrete, perceivable event. The choice of formal nouns will be relevant to automatic generation and surface realisation of natural language, if we are to make the output as natural as possible.

Other formal nouns constitute adverbial units which can be classified into four categories: temporal, causal, manner and others. These formal nouns function like conjunctions or postpositions. The following four examples illustrate each category:

(3.121) Temporal; "saichuu", during
Kaigino saichu(ni) Kenwa okashinakotowo iidashita.

kaigi-no *saichu(-ni)* *Ken-wa* *okashi-na-koto-wo*
meeting-of during(-OBL) Ken-TOP funny-attr.base-thing-OBJ
ii-dashi-ta
say-SUFF-decl.ta
"In the meeting, Ken began to talk about funny things."

(3.122) Causal; "sei", because
Kenno okashina hanasino seide kaigiga sanjuppun nobita.

Ken-no *okashi-na* *hanasi-no* *sei-de* *kaigi-ga*
Ken-of funny-attr.base talk-of because-part meeting-SUBJ
sanjuppun *nobi-ta*
thirty.minutes get.longer-decl.ta
"Because of Ken's funny talk, the meeting got thirty minutes longer than planned."

(3.123) Manner; "you", as if
Karewa [toritsukareta] youni uchuno bouchounitsuite hanashita.

Kare-wa *[toritsuk-are-ta]* *you-ni* *uchu-no*
He-TOP possess-PASS-decl.ta like-OBL universe-of
bouchou-ni-tsuite *hanas-ta.*
expansion-OBL-about talk-decl.ta
"He talked about the expansion of the Universe as if he had been possessed."

(3.124) Others; "ippou", on the other hand
Ippoude gichouwa inemurishiteita.

ippou-de *gichou-wa* *inemuris-tei-ta*
the.other.side-part chairperson-TOP take.a.nap-PROG-decl.ta
"On the other hand, the chairperson was taking a nap."

These formal nouns can be used like a complementiser, introducing a clause (cf. Section 4.2.10). For example, the formal noun "youni" in the sentence (3.123) follows a clause, and this noun does not function as any of the arguments of the head verb of the clause.

Numerical Classifiers

Numerical classifiers are the type of nominal suffixes which specify the number of the noun. One numerical classifier attaches to one numerical noun, and this "numerical noun + numerical classifier" construction constitutes one unit. This unit (numerical unit) functions as an adverbial unit, or as a nominal unit when followed by case particles "-ga", "-wo" or "-ni", or as a postpositional unit when followed by other particles. The choice of numerical classifiers depends on the noun they modify. For example, a numerical classifier "satsu" is used exclusively for books:

(3.125) Kenwa honwo daitai sensatsu motteiru.

Ken-wa	*hon-wo*	*daitai*	*sen-satsu*
Ken-TOP	book-OBJ	about	one.thousand-NC:book
mot-tei-ru			
have-prog-decl.base			

"Ken has about a thousand books."

In the example above, the numerical unit "sensatsu" functions as an adverbial unit for the root verb. This construction is known as *quantifier floating* ([Dowty and Brodie, 1984], [Miyagawa, 1989]). The numerical unit depends on the root verb, and not the nominal unit "honwo". Notice that there is no backward dependency in Japanese, and that the adverbial unit "sensatsu" cannot depend on (hence modify) the nominal unit "honwo". It is expected that treating numerical units as adverbial units will keep the analysis simple.

Unit Classifiers

The term "unit" in unit classifiers does not mean "syntactic units" (as I have been using the term in this thesis). Unit classifiers are the type of nominal suffixes which specify particular units, such as time, date, currencies, order, etc. One unit classifier follows after a numerical, and this constitutes a noun which indicates the amount of the unit:

(3.126) Time; "-pun", minutes
Sampun

san-pun
three-UC:minute
"three minutes"

(3.127) Currency; “-doru”, dollar
Hyakunijuugodoru

hyaku-nijuu-go-doru
one.hundred-twenty-five-UC:dollar
“one hundred twenty five dollar”

(3.128) Date; “-gatsu”, month, “-nich”, day
Kugatsusanjuunichi

ku-gatsu-sanjuu-nichi
nine-UC:month-thirty-UC:day
“the 30th of September”

(3.129) Frequency; “-kai”, time
Hyakkai mawaru

hyaku-kai *mawar-u*
one.hundred-UC:times turn-decl.base
“to turn one hundred times”

(3.130) Order; “-bamme”, Xth
Migikara kazoete yombamme

migi-kara *kazoe-te* *yon-bamme*
right-from count-conn.ta four-UC:order
lit.“Counting from the right, the fourth” or
“fourth from the right”

3.6.2 Particles

The particle after the head noun determines the grammatical function of the noun unit. The absence of particle in a nominal unit also determines its grammatical function, since some nouns can function as adverbs when they are not followed by a particle. For example, the noun “ashita” means “tomorrow”, and this noun can function differently according to the particle right after it:

(3.131) Ashitawa tenkiga iidarou

ashita-wa *tenki-ga* *ii-darou*
tomorrow-TOP weather-SUBJ good-AUX.decl.base
"As for tomorrow, the weather will be nice."

(3.132) Ashitaga watashino tanjobida

ashita-ga *watashi-of* *tanjobi-da*
tomorrow-SUBJ I-part birthday-COPL.decl.base
"Tomorrow is my birthday."

(3.133) Akarui ashitawo tsukurimashou.

akaru-i *ashita-wo* *tukur-i-mashou*
bright-conn.base tomorrow-OBJ create-conn.base-SUFF.VOL
"Let's make tomorrow bright."

(3.134) Tsurini ikunowa ashitani shita.

tsuri-ni *ik-u-no-wa* *ashita-ni* *shi-ta*
fishing-OBL go-decl.base-no-TOP tomorrow-OBL do-decl.ta
"I decided to go fishing tomorrow."

(3.135) Ashitakara shingakkiga hajimaru.

ashita-kara *shingakki-ga* *hajimar-u*
tomorrow-from new academic year-SUBJ start-decl.base
"The new academic year starts tomorrow."

(3.136) Atarashii terebiga ashita haitatsusareru

atarashi-i *terebi-ga* *ashita* *haitatu-s-areru*
new-decl.base TV-SUBJ tomorrow deliver-do-SUFF.decl.base
"A new TV will be delivered (to somewhere) tomorrow."

Particles are classified into a number of categories according to their functions. This section explains the functions of particles in each category. Since it is impossible to cover all the details of how particles are used and how they function in each usage, I chose only a few of them for each category and the examples are the most basic and typical ones.

Case Particles

Case particles are used to specify the grammatical function of the noun unit they attach to. The grammatical functions they can specify are SUBJECT, OBJECT, OBLIQUE, and Postpositional adjunct (PADJ). The JUMAN dictionary registers the following 12 case particles: -ga, -wo, -ni, -kara, -to, -de, -made, -yori, -no, -nite, -tto. In general, one case particle has several meanings. For example, a noun unit with the case particle "-kara" can mean the following:

(3.137) The starting point of a movement:
Iekara gakkouni aruiteitta.

Ie-kara *gakkou-ni* *aruk-ite-ik-ta*
house-from school-to walk-conn.base-go-decl.ta
"I walked to school from home."

(3.138) The source of giving:
Anekara kono honwo moratta.

ane-kara *kono* *hon-wo* *moraw-ta*
elder.sister-from this book-OBJ receive-decl.ta
"My elder sister gave me this book."

(3.139) The starting time of an action:
Kujikara jugyouga hajimaru.

ku-ji-kara *jugyou-ga* *hajimar-u*
nine-time-from class-SUBJ start-decl.base
"The class starts at nine o'clock."

(3.140) The reason of an action:
Ichigyouno puroguramumisukara daikibona sisutemudaunga okita.

ichi-gyou-no *puroguram-misu-kara* *daikibo-na*
one-line-of program-mistake-from large.scale-attr.base
sisutem-daun-ga *okir-ta*
system-down-SUBJ take.place-decl.ta
"A large-scale system down took place because of one line of program mistake."

(3.141) The reason of giving a particular judgement
Shouraino dekigotowo genzaino jyoukyoukara yosokusuru.

shourai-no *dekigoto-wo* *genzai-no* *jyoukyou-kara*
future-of event-OBJ present-of situation-from
yosoku-sur-u
prediction-do-decl.base
"Predicting future events from the present situation"

(3.142) The material of something:
Sakewa komekara dekiteiru.

sake-wa *kome-kara* *deki-te-ir-u*
Japanese.sake-TOP rice-from make-conn.ta-suff.-decl.base
"Japanese sake is made from rice."

It seems that this case particle has the prototypical meaning of specifying the "source" of an action or thing, and the verb it depends on and the head noun of the noun unit determine the meaning that this case particle "-kara" has.

Case particles "-ga" and "-wo" specify the grammatical function SUBJ and OBJ, respectively, and these are rare cases where there seems to be a one-to-one correspondence between a case particle and a grammatical function.

Adverbial Particles

Adverbial Particles produce topic units or focus units from noun units. The JUMAN dictionary registers 24 adverbial particles. The adverbial particle "-wa" is one of the most frequent ones, specifying the topic of a sentence. The adverbial particle "-mo" is the second most frequent one, meaning "too". As for other adverbial particles, they have various meanings which would be expressed adverbially in English, for example "too", "even" or "only":

(3.143) Watashiwa kono honwo yonda.

watashi-wa *kono* *hon-wo* *yom-ta*
I-TOP this book-OBJ read-decl.ta
"I read this book."

(3.144) Watashi*mo* kono honwo yonda.

watashi-mo *kono* *hon-wo* *yom-ta*
I-too this book-OBJ read-decl.ta
"I, too, read this book."

(3.145) Watashiwa kono hon*mo* yonda.

watashi-wa *kono* *hon-mo* *yom-ta*
I-TOP this book-too read-decl.ta
"I read this book, too."

(3.146) Watashi*sae* kono honwo yonda.

watashi-sae *kono* *hon-wo* *yom-ta*
I-even this book-OBJ read-decl.ta
"Even I read this book(, therefore others should do so)."

(3.147) Watashiwa kono hon*sae* yonda.

watashi-wa *kono* *hon-sae* *yom-ta*
I-TOP this book-even read-decl.ta
"I read even this book(, therefore it is natural for me to have read other books)."

(3.148) Watashi*dake* kono honwo yonda.

watashi-dake *kono* *hon-wo* *yom-ta*
I-only this book-OBJ read-decl.ta
"Only I read this book(, while others didn't)."

(3.149) Watashiwa kon hon*dake* yonda.

watashi-wa *kono* *hon-dake* *yom-ta*
I-TOP this book-only read-decl.ta
"I read this book(, but not other books)."

Conjunctive Particles

Conjunctive particles are used for conjoining nouns or clauses.

(3.150) Kento Naomi

Ken-to *Naomi*
Ken-and Naomi
"Ken and Naomi"

(3.151) Sugu ikukara mattete.

sugu *ik-u-kara* *mat-te-ite.*
soon come-decl.base-part wait-conn.base-be-imp
"I'm coming soon, so wait for a moment."

In sentence (3.151), the conjunctive particle "kara" means literally "since", thus the literal translation of (3.151) would be "Since I'm coming soon, wait for a moment."

Conjunctive particles connect a noun to a noun or a clause to a clause. The connection can be coordination or subordination. A coordination noun unit can be followed by a case particle or an adverbial particle, and its grammatical function is specified for all the coordinates, as in sentence (3.152). Both Ken and Naomi are the subject of the verb. A subordination noun unit can be followed by a case particle or an adverbial particle, but it only specifies the noun which it attaches to. The subject of the verb in sentence (3.153) is "imouto", a sister, and the conjunctive particle has the function of giving the grammatical function "possessive" to the noun, in this case "Ken":

(3.152) Kento Naomiga goukakushita

[Ken-to *Naomi]-ga* *goukaku-shita*
[Ken-coord.part Naomi]-SUBJ pass.an.exam-do.decl.ta
"Ken and Naomi passed the exam."

(3.153) Kenno imoutoga goukakushita

Ken-no *imouto-ga* *goukaku-shita*
Ken-coord.part sister-SUBJ pass.an.exam-do.decl.ta
"Ken's sister passed the exam."

It seems that subordination of nouns by conjunctive particles should be treated as a postpositional unit depending on the noun unit, in which the conjunctive particle is a kind of postposition which takes a noun unit as its argument, and this unit depends on the following noun unit. In example (3.153), the unit "Ken-no" is a postpositional unit with the grammatical function possessive (or just "postpositional", since this particle can have meanings other than possession, depending on the meaning of the noun it attaches to), and this depends on the following noun unit "imouto-ga".

When a conjunctive particle connects clauses, the connection can also be coordination or subordination. The distinction between a coordinated clause and a subordinated clause depends on the scope of the topic phrase, and both of them are sentential adjuncts (SADJs) to another clause.

Sentence-Ending Particles

Sentence-ending particles appear at the end of a sentence to express various types of modal meanings, such as question, assertion, prohibition:

(3.154) Kono honwo yomimashitaka

kono	*hon-wo*	*yom-i-mas-ta-ka*
this	book-OBJ	read-conn.base-polite-decl.ta-Q

"Have you read this book?"

(3.155) Kono honwo yomimashitane

kono	*hon-wo*	*yom-i-mas-ta-ne*
this	book-OBJ	read-conn.base-polite-decl.ta-part

"You have read this book, haven't you?"

(3.156) Kono honwa yomuna

kono	*hon-wa*	*yom-u-na*
this	book-OBJ	read-decl.base-part

"As for this book, you shouldn't read this."

The particles in the examples above have different functions. In sentence (3.154), the sentence-ending particle "-ka" indicates that this sentence is a question. Without it, the sentence is just a declarative sentence; "Kono honwo yomimashita (I read this book)." In sentence (3.155), the sentence-ending particle "-ne" indicates that the speaker intends to make sure that his

knowledge (in this case, the listener's having read the book) is right. Again, the sentence without the sentence-ending particle is a declarative sentence. The "-na" in sentence (3.156) indicates that the speaker orders the listener not to read the book. One of the interesting functions of sentence-ending particles is that some of them, in particular "-ne" and "-sa", can be used to emphasise the syntactic unit (bunsetsu) boundaries in a sentence:

(3.157) Watashiwane konone honwone yomimashita

watashi-wa-ne *kono-ne* *hon-wo-ne*
I-TOP-part this-part book-OBJ-part
yom-i-mas-ta
read-conn.base-polite-decl.ta
"I have read this book."

Example (3.157) shows that sentence-ending particles are not purely "sentence-ending", but they can indicate the ending point of syntactic units, so that the listener would pay more attention to the unit. This is a simple and reliable measure to approach the problem of syntactic unit boundary identification.

More than One Particle in a Single Unit

In some fixed expressions, more than one particle can appear in a single unit. For example, the expression "-ka-dou-ka", which means "whether", has two conjunctive particles "-ka" which introduces a question, and this is followed by the case particle "-wo".

(3.158) Kono honwo yomubekikadoukawo senseini tazuneta.

kono *hon-wo* *yom-u-beki-ka-dou-ka-wo*
this book-OBJ read-decl.base-aux.root-Q-dem-Q-OBJ
sensei-ni *tazune-ta*
teacher-OBL ask-decl.ta
"I asked (my) teacher whether I should read this book."

3.6.3 Non-Inflecting Adjectives, or 'Rentaishi'

Rentaishi (literally rentai "attributive" + shi "part of speech") are non-inflecting adjectives. They modify a noun, and they do not admit predicative usage, i.e., they cannot be the root of a sentence or a clause. They are divided into several categories, and those that are derived from inflecting parts of

speech need special treatments by morphological analysers or parsers, so that they can be correctly distinguished from strings that are identical in terms of morphology.

Non-inflecting Adjectives from Verbs in the Base Form

Rentaishi in this category have the same form as the verbs from which they are derived, but the meaning is not simply the adjectivisation of these verbs. For example, the rentaishi "aru" means "certain", and has the same form as the verb "aru", which means "exist":

(3.159) aru jinbutsu

aru	*jinbutsu*
certain	person

"a certain person"

If a parser analyses sentence (3.159) as (3.160), then this is syntactically correct, but often semantically odd:

(3.160) aru jinbutsu

ar-u	*jinbutsu*
exist-decl.base	person

"a person who exists"

The distinction between "aru" as a rentaishi and "aru" as a verb depends on the syntactic environment in which they appear. If "aru" appears before a noun and nothing depends on this, then this "aru" is a rentaishi. If "aru" appears before a noun and something depends on this "aru", then this "aru" is the head of the relative clause which modifies the noun. For example, the "aru" in sentence (3.161) is a verb, since an oblique unit depends on "aru":

(3.161) Takai chiini aru jinbutsu

taka-i	*chii-ni*	*ar-u*	*jinbutsu*
high-decl.base	position-OBL	exist-decl.base	person

"a person who is at a high position"

Non-Inflective Adjectives Derived from Verbs in the Ta Form

Some rentaishi are derived from verbs in the ta form, and the derived meaning is different from the original meaning. The rentaishi "komatta" meaning "troublesome", for example, is the same form as the ta form of a verb "komaru" which means "be troubled".

(3.162) komatta hito

komatta	*hito*
troublesome	person

"a troublesome person"

If something depends on "komatta", and this is one of the arguments or adjuncts of the verb "komaru", then it is obvious that this "komatta" is a verb:

(3.163) Shakkinde komatta hito

shakkin-de	*komar-ta*	*hito*
loan-particle	be.troubled-decl.ta	person

"A person who suffered from loans"

If a parser analyses the sentence "komatta hito" as in (3.164), then this is simply wrong in terms of semantics. The analysis is syntactically possible, but the interpretation of this phrase by native speakers of Japanese is "a troublesome person".

(3.164) komatta hito

komar-ta	*hito*
be.troubled-decl.ta	person

"a person who suffered"

These examples show the importance of morphological analysers' proper treatment of non-inflecting adjectives derived from inflecting parts of speech. One approach to this issue is to list all the possible rentaishi in the dictionary of a morphological analyser and analyse input accordingly. The JUMAN dictionary has 81 entries for rentaishi, and these are all lexicalised, fixed rentaishi. However, this approach cannot cover new rentaishi which have not been listed in the dictionary of the analyser. Even though it seems that rentaishi is not as productive as other parts of speech, the problem remains to be solved in future research.

3.6.4 Pronouns

Pronouns in Japanese are categorised into two classes: demonstrative pronouns and personal pronouns. Other pronominal categories we know in English are expressed by other means in Japanese.

Reflexive pronouns and possessive pronouns do not exist as pronominal categories; possession is expressed by a noun, either pronominal or not, with the postposition "-no" (cf. Section 3.6.2.3). Reflexivity can be expressed by a noun "jibun" for all the three persons and numbers, or by a personal pronoun plus "jishin", such as "watashijishin (myself)"

Relative pronouns do not exist at all in Japanese. The verbal inflection of the head verb of a clause and the syntactic environment in which this clause appears are the clues to determine whether it is a relative clause or not. The following two sections introduce demonstrative pronouns and personal pronouns.

Demonstrative Pronouns

Demonstrative pronouns refer to something which the speaker refers to in reality, or which is in the context. Japanese demonstrative pronouns are divided into four categories in terms of deixis, and each category has different demonstratives for different meanings. Table 3.6 summarises the demonstrative pronominal system of Japanese. The rows give the categories, and the columns the meanings:

Table 3.6: The system of demonstrative pronouns

	thing	place	direction	adjectival	state	adverbial
near speaker	kore	koko	kochira	kono + N	konna	kou
near listener	sore	soko	sochira	sono + N	sonna	sou
far from both	are	asoko	achira	ano + N	anna	aa
indefinite	dore	doko	dochira	dono + N	donna	dou

Adjectival demonstrative pronouns (in the fourth column from the left) always appear before a noun, thus functioning as determiners for the noun.

The demonstrative pronouns "kono", "ano", "sono" and "dono" function as determiners, like "this" or "that" when used as demonstrative determiners, such as "this book" or "that book". However, they are analysed as one syntactic unit depending on a noun unit. The demonstrative pronoun "kono" in example (3.165) is one unit depending on the next unit "honwo". These

two units constitute what would be expressed by an NP in English, namely "this book":

(3.165) Watashiwa kono honwo yonda

watashi-wa	*kono*	*hon-wo*	*yom-ta*
I-TOP	this	book-OBJ	read-decl.ta

"I read this book."

As Table 3.6 shows, the first syllable of each demonstrative indicates which category it belongs to. For example, when a speaker refers to something near herself, then she refers to it by "kore":

(3.166) Korewa nani?

kore-wa	*nani*
dem.thing.near.speaker-TOP	what

"What is this?"

When a speaker refers to something near the listener, then she refers to it by "sore":

(3.167) Sorewa nani?

sore-wa	*nani*
dem.thing.near.listener-TOP	what

"What is it?"

When a speaker refers to something distant from both the speaker and the listener, then she refers to it by "are":

(3.168) Arewa nani?

are-wa	*nani*
dem.thing.far-TOP	what

"What is that?"

The indefinite demonstratives cannot appear with the topic particle "-wa". This is because the topic of a sentence must be something definite:

(3.169) Dorega kimino hon?

dore-ga	*kimi-no*	*hon*
dem.thing.indef	you-of	book

"Which is your book?"

Personal Pronouns

Personal pronouns refer to the person or the people in context. One of the characteristics of Japanese is the variety of personal pronouns used by different genders and different ages, with different social status, regions, occupations, degrees of politeness, etc. For example, the Japanese Wikipedia has an entry "1st person pronouns of Japanese language" and this entry explains 50 different 1st person pronouns (actually, this entry only talks about 1st person singular pronouns and does not talk about the 1st person plural pronouns corresponding to each of the singular ones). The function of 1st person singular pronouns is not just referring to the speaker, but showing some of the personal attributes of the speaker. For example, the differences between the 1st person singular pronouns show the difference of the referent of the subject in the following sentences. All of them mean "I read this book":

(3.170) Bokuwa kono honwo yonda. (The subject is a boy, or an adult talking in a nonformal context.)

Watashiwa kono honwo yomimashita. (The subject is an adult, or a child in a formal context.)

Watakusiwa kono honwo yomimashita. (The subject is an adult talking in a formal context.)

Atashiwa kono honwo yonda. (The subject is a woman in a nonformal context.)

Orewa kono honwo yonda. (The subject is a man trying to sound masculine.)

Washiwa kono honwo yonda. (The subject is an old man, or a man born and raised in Western Japan.)

In English, "I" is always "I" anywhere anytime, while in Japanese the first person singular changes (sometimes should be changed) according to the time, place and occasion. This may evoke some interesting sociolinguistic ideas, but these are not relevant here. The variety of personal pronouns can be explained differently as follows. These so-called personal pronouns are not actually "pronouns", but ordinary nouns which happen to have the characteristic that they exclusively refer to the 1st person singular. These nouns have their own meaning, such as "boy", "male", "woman", "old man",

etc. and what they have in common is the grammatical feature which can be expressed LFG style as "PERSON=SINGULAR".

3.6.5 Adverbs

Adverbs are used to modify a predicate. They are divided into several semantic categories such as manner, degree, amount, tense and aspect. Some adverbs function as sentential adverbs ([Masuoka and Takubo, 1992]).

Manner Adverbs

Manner adverbs express the manner of an action. In Sentence (3.171), "yukkuri" is a manner adverb modifying the predicate "yom-":

(3.171) Watashiwa kono honwo yukkuri yondeiru.

watashi-wa	*kono*	*hon-wo*	*yukkuri*	*yom-te-ir-u*
I-TOP	this	book-OBJ	slowly	read-conn.ta-suff-decl.base

"I read this book slowly."

Japanese manner adverbs include a quite extensive number of mimetic words, which are divided into onomatopoeia and ideophones ([Tsujimura, 2006]).

Degree Adverbs

Degree adverbs express the degree of a state. In Sentence (3.172), "taihen" is a degree adverb modifying the predicate "omoshiro-" (interesting):

(3.172) Kono honwa taihen omoshiroi.

kono	*hon-wa*	*taihen*	*omoshiro-i*
this	book-TOP	very	interesting-decl.base

"This book is very interesting."

Amount Adverbs

Amount adverbs express the amount of an argument of a predicate. Numerical classifiers belong to this category. In Sentence (3.173), "takusan" is an amount adverb:

(3.173) Watashiwa honwo takusan yonda.

watashi-wa *hon-wo* *takusan* *yom-ta*
I-TOP book-OBJ many read-base.ta
"I read many books."

When the object is zero-pronominalised, amount adverbs (and numerical classifiers before the verb) indicate the amount of the object:

(3.174) Watashiwa takusan yonda.

watashi-wa *takusan* *yom-ta*
I-TOP many read-base.ta
"I read many."

Amount adverbs can also indicate the amount of a zero-pronominalised subject. Both in (3.175) and (3.176), "takusan" indicates the amount of the subject.

(3.175) Gakuseiga takusan kita.

gakusei-ga *takusan* *ki-ta*
student-SUBJ many come-base.ta
"Many students have come."

(3.176) Takusan kita.

takusan *ki-ta*
many come-base.ta
"Many have come."

Amount adverbs can be followed by the particle "-no" and depend on the argument which it modifies:

(3.177) Watashiwa takusanno honwo yonda.

watashi-wa *takusan-no* *hon-wo* *yom-ta*
I-TOP many book-OBJ read-base.ta
"I read many books."

Tense Adverbs

Tense adverbs specify the tense of an event. In Sentence (3.178), the main verb "yomu" is in declarative base form, hence the tense feature is nonpast (cf. Section 3.5.1). The adverb "mousugu" (near future) further specifies that the action will take place in near future:

(3.178) Watashiwa kono honwo mousugu yomu.

watashi-wa	*kono*	*hon-wo*	*mousugu*	*yom-u*
I-TOP	this	book-OBJ	near.future	read-decl.base

"I will soon read this book."

Aspect Adverbs

Aspect adverbs specify aspectual features of sentences. In Sentence (3.179), "yok-u" is an adjective in adverbial base form (cf. Section 3.5.3), which means "often", and it functions as an aspect adverb indicating a repetition of action:

(3.179) Watashiwa kono honwo yoku yondeiru.

watashi-wa	*kono*	*hon-wo*	*yok-u*
I-TOP	this	book-OBJ	good/often-adv.base

yom-te-ir-u
read-conn.ta-suff-decl.base
"I often read this book."

3.7 Summary

This chapter described the core syntactic and morphological properties of Japanese. Section 3.2 described the non-configurationality of Japanese and points out that the multi-level architecture of LFG is appropriate for dealing with non-configurational languages just as it is for configurational languages. Section 3.3 introduced the idea of "bunsetsu", which function as the unit of syntactic investigation for Japanese. Each of these "bunsetsu" in a sentence corresponds to a piece of f-structure, and these f-structure pieces are combined through the labelled dependency relationships represented by the DAG containing the "bunsetsu" units as vertices, and define the f-structure for the sentence as a whole. A rough definition of DAG-based dependency representations is also provided for further cross-linguistic investigations. Section 3.4

dealt with topicalisation by a particular particle and zero pronouns, and also suggests that we need special treatments for zero pronoun identification for natural language processing of Japanese. Section 3.5 described the inflecting parts of speech, with some focus on the inflection forms and their functions which play important roles in the automatic annotation of grammatical functions of syntactic units, the detail of which will be described in later chapters. Section 3.6 describes the non-inflecting parts of speech.

Chapter 4

An LFG-Based Description of Core Japanese Grammar

4.1 Introduction

This chapter describes core aspects of Japanese grammar based on the framework of LFG. In particular, I look at how grammatical functions and grammatical features are encoded in Japanese. In doing this, the core aspects of Japanese grammar presented in Chapter 3 will be relevant. As mentioned in Chapter 3, Japanese sentences are divided into "bunsetsu", or *syntactic units*, and each of the dependencies among these units has a unique grammatical function which is specified by the particle at its end, or by the part of speech of the head of the syntactic unit. Sometimes, however, ambiguity arises which grammatical function should be assigned. Information which is lexically or morphologically encoded in each syntactic unit is integrated through the dependency relations into the f-structure for the sentence as a whole. I use the DAG representation introduced in Chapter 3 to visualise the basic dependency relationships between the syntactic units for each example sentence.

The structure of this chapter is as follows: Section 4.2 describes how grammatical functions are encoded in Japanese, and what meaning they have in actual sentences. The grammatical functions described are: TOPIC, FOCUS, SUBJ, OBJ, OBL, PADJ (postpositional adjunct), ADJ, DET, SADJ (sentential adjunct), REL (relative clause), COMP, and COORD. In this study, TOPIC and FOCUS are treated as grammatical functions, rather than "discourse functions". This is because the function assignment of TOPIC and FOCUS in Japanese is in principle the same as the function assignment for other grammatical functions. TOPIC and FOCUS are specified by particu-

lar particles on syntactic units, along with other grammatical functions. The DAG representation of the dependency relations among units for each of the example sentences, and the f-structure representations corresponding to the paths of the DAG are given for the example sentences. Section 4.3 explains each of the grammatical features in Japanese f-structures. The grammatical features dealt with in this section are basic ones which are relevant cross-linguistically: TENSE, ASPECT, MOOD, and VOICE. Finally, Section 4.4 summarises the chapter. In general, LFG f-structure representations are richer than the corresponding DAG representations in that (1) f-structures represent grammatical features in addition to grammatical relations and (2) f-structures represent control relations (anaphoric / functional).

4.2 Grammatical Functions

4.2.1 TOPIC

The grammatical function TOPIC is specified by the adverbial particle "-wa" after a noun phrase. TOPIC often provides a contextual referent for a zero pronoun (in the same sentence) whose grammatical function is either SUBJ or OBJ, but not always (cf. Section 3.4.3). So far in this thesis, we have seen many cases of TOPIC providing a referent for a SUBJ zero pronoun. Example (4.1) is a TOPIC setting up a referent for an OBJ zero pronoun. Figure 4.1 is the DAG representation of the dependency relationships among the syntactic units in (4.1), and Figure 4.2 is the f-structure for (4.1). The format of "Sentence-DAG-F-structure" is used throughout this section if not stated otherwise. The existence of the OBJ pro is based on the observation that the verb "yom- (read)" is a transitive verb, in other words, a reading event entails the existence of something to be read:

(4.1) Kono honwa aniga mainichi yondeiru.

[kono]$_{f1}$	*[hon-wa]*$_{f2}$	*[ani-ga]*$_{f3}$	*[mainichi]*$_{f4}$	*[pro]*$_{f5}$
this	book-TOP	elder.brother-SUBJ	every.day	pro-OBJ

[yom-te-ir-u]$_{f6}$
read-conn.ta-suff.-decl.base
"As for this book, my elder brother reads this every day."

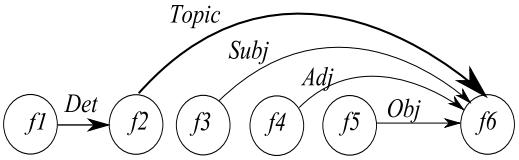

Figure 4.1: DAG for (4.1)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f2}\begin{bmatrix} \text{DET} & {}_{f1}\begin{bmatrix}\text{PRED} & \text{`kono; this'}\end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix}\text{PRED} & \text{`ani; elder brother'} \\ \text{PRTCS} & \text{`-ga'}\end{bmatrix} \\
\text{OBJ} & {}_{f5}\begin{bmatrix}\text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i}\end{bmatrix} \\
\text{ADJ} & \left\{ {}_{f4}\begin{bmatrix}\text{PRED} & \text{`mainichi; every day'}\end{bmatrix} \right\} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-te-'} \\
\text{SUF} & \text{`-ir-'} \\
\text{SUF-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`present'} \\
\text{MOOD} & \text{`declarative'} \\
{}_{f6}\text{ASPECT} & \text{`progressive'}
\end{bmatrix}
$$

Figure 4.2: F-structure for (4.1)

The big round brackets around the TOPIC value represent that the value can be a set of topics. Note that in contrast to the DAG in Figure 4.1, the f-structure represents the anaphoric control between the TOPIC and the OBJ (in terms of sharing the referential index i).

The adverbial particle "-wa" can follow another case particle except for "-ga" and "-wo". The grammatical function of such noun units is not TOPIC, but OBL (when the case particle is "-ni") as in example (4.2) or PADJ (when the case particle is not "-ni") as in example (4.3):

(4.2) imotoniwa kono honwo yomaseta.

$[pro]_{f1}$	$[imoto\text{-}ni\text{-}wa]_{f2}$	$[kono]_{f3}$	$[hon\text{-}wo]_{f4}$
pro-SUBJ	younger.sister-OBL-prtadv	this	book-OBJ

$[yom\text{-}ase\text{-}ta]_{f5}$
read-caus.-decl.ta

"As for my younger sister, I had her read this book".

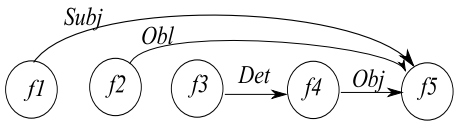

Figure 4.3: DAG for (4.1)

$$
\begin{bmatrix}
\text{SUBJ} & {}_{f1}\begin{bmatrix}\text{PRED} & \text{`pro'}\end{bmatrix} \\
\text{OBL} & {}_{f2}\begin{bmatrix}\text{PRED} & \text{`imoto; younger sister'} \\ \text{PRTCS} & \text{`-ni'} \\ \text{PRTADV} & \text{`-wa'}\end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}\text{DET} & {}_{f3}\begin{bmatrix}\text{PRED} & \text{`kono; this'}\end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'}\end{bmatrix}_{f4} \\
\text{PRED} & \text{`yom-ase-}\langle \text{SUBJ, OBL, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`root'} \\
\text{SUF} & \text{`-ase-'} \\
\text{SUF-INFL} & \text{`decl.ta'} \\
\text{VOICE} & \text{`causative'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`decl'}
\end{bmatrix}_{f5}
$$

Figure 4.4: F-structure for (4.2)

(4.3) Nagasakikarawa mainichi funega deteiru

[Nagasaki-kara-wa]$_{f1}$ *[mainichi]*$_{f2}$ *[fune-ga]*$_{f3}$
Nagasaki-from-prtadv every.day ship-SUBJ
[de-te-i-ru]$_{f4}$
depart-conn.ta-suf-decl.base
"From Nagasaki, ships departs every day."

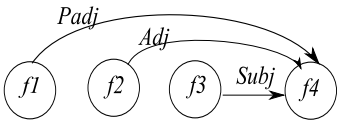

Figure 4.5: DAG for (4.3)

$$
\begin{bmatrix}
\text{PADJ} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{'Nagasaki'} \\ \text{PRTCS} & \text{'-kara'} \\ \text{PRTADV} & \text{'-wa'} \end{bmatrix} \right\} \\
\text{ADJ} & \left\{ {}_{f2}\begin{bmatrix} \text{PRED} & \text{'mainichi; every day'} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{'fune; ship'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\
\text{PRED} & \text{'de-}\langle \text{SUBJ} \rangle\text{; depart'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-te-'} \\
\text{SUF} & \text{'-i-'} \\
\text{SUF-INFL} & \text{'-ru'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'decl'} \\
\text{ASPECT} & \text{'progressive'}
\end{bmatrix}_{f4}
$$

Figure 4.6: F-structure for (4.3)

Once a syntactic unit is given its unique grammatical function, then this unit cannot be further specified by any other grammatical function, unless there is a zero pronoun reference involved. In addition, the order of a case particle and an adverbial particle is fixed: an adverbial particle cannot precede a case particle. For example, *"Watashiwa imotowani kono honwo yomaseta." is ungrammatical. In example (4.2), the case particle specifies the grammatical function of the noun phrase, and the adverbial particle "-wa" after the case particle just marks that the unit has topical status.

In some constructions the TOPIC in a sentence does not set up a referent for any zero pronoun, but instead functions as a modifier for its target unit. The "eel sentence" is an example of such constructions:

(4.4) Bokuwa unagida

[boku-TOP]$_{f1}$	*[pro]*$_{f2}$	*[unagi-da]*$_{f3}$
I(young.male)-TOP	pro-SUBJ	eel-Copula.plain.decl.base

"I am an eel." or "As for me, it is eel."

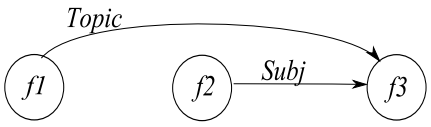

Figure 4.7: DAG for (4.4)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`boku; I}\big(\text{YOUNG MALE}\big)\text{'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PRED} & \text{`unagi; eel'} \\
\text{ROOT} & \text{`+'} \\
\text{COPULA} & \text{`-da'} \\
\text{COPULA-INFL} & \text{`}\phi\text{'} \\
\text{TENSE} & \text{`present'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f3}
$$

Figure 4.8: F-structure for (4.4)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`boku; I}\big(\text{YOUNG MALE}\big)\text{'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{j} \end{bmatrix} \\
\text{PRED} & \text{`unagi; eel'} \\
\text{ROOT} & \text{`+'} \\
\text{COPULA} & \text{`-da'} \\
\text{COPULA-INFL} & \text{`}\phi\text{'} \\
\text{TENSE} & \text{`present'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f3}
$$

Figure 4.9: Another F-structure for (4.4)

Example (4.4) has two different interpretations. One (Figure 4.8) is "I am an eel," in which the TOPIC shares its referential index with the SUBJ zero pronoun, and the referent of the TOPIC is also the referent of the SUBJ of the sentence.

The other interpretation (Figure 4.9) is "As for me, it is eel." Here TOPIC does not share its referential index with the SUBJ zero pronoun but functions as a modifier of the sentence and the zero pronoun refers to something beyond this sentence.

Which of these interpretations should be chosen depends upon the semantic naturalness of the possible interpretations in a given context. In certain contexts such as fairy tales or cartoons, it is natural to say "I am an eel." Eels aside, sentences of this type involving a Topic sharing their referential index with a zero pronoun with a core grammatical function are often found in text and in conversation in a formal setting.

4.2.2 FOCUS

The grammatical function FOCUS is specified by one of the adverbial particles (cf. Section 3.6.2.2) except for "-wa" after a noun phrase. A noun unit with FOCUS is the antecedent of a zero pronoun in the same sentence.

(4.5) Watashimo kono honwo yonda.

[watashi-mo]$_{f1}$	*[pro]*$_{f2}$	*[kono]*$_{f3}$	*[hon-wo]*$_{f4}$	*[yom-ta]*$_{f5}$
I-FOC	pro-SUBJ	this	book-OBJ	read-decl.ta

"I read this book(, and so did others)"

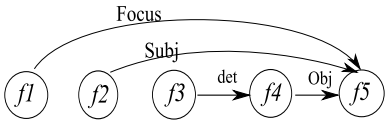

Figure 4.10: DAG for (4.5)

$$
\begin{bmatrix}
\text{FOCUS} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-mo'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
{}_{f5}\begin{bmatrix} \text{MOOD} & \text{`declarative'} \end{bmatrix}
\end{bmatrix}
$$

Figure 4.11: F-structure for (4.5)

The OBJ of a sentence can also be focused by an adverbial particle. In (4.6), the object is focused to show that there are some other books that the speaker has read:

(4.6) Watashiwa kono honmo yonda.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[kono]*$_{f3}$	*[hon-mo]*$_{f4}$	*[pro]*$_{f5}$
I-TOP	pro-SUBJ	this	book-FOC	pro-OBJ

[yom-ta]$_{f6}$
read-decl.ta

"I read this book(, along with other books)."

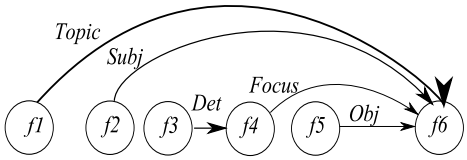

Figure 4.12: DAG for (4.6)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{FOCUS} & \left\{ \begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTADV} & \text{`-mo'} \\ \text{INDEX} & \text{j} \end{bmatrix}_{f4} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f5}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{j} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
{}_{f6}\text{MOOD} & \text{`declarative'}
\end{bmatrix}
$$

Figure 4.13: F-structure for (4.6)

When there is one TOPIC unit and one FOCUS unit and they depend on a transitive verb, then the problem arises which of them is the antecedent of which of the two zero pronouns (required by the transitive verb). This problem can be solved only by the semantic affinity of the nouns and the verb involved. In example (4.7), the TOPIC unit is the antecedent of the SUBJ zero pronoun, while the FOCUS unit is the antecedent of the OBJ

zero pronoun.

(4.7) Watashiwa konna honwa yomanai.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[konna]*$_{f3}$	*[hon-wa]*$_{f4}$	*[pro]*$_{f5}$
I-TOP	pro-SUBJ	such.a	book-FOC	pro-OBJ

[yom-ana-i]$_{f6}$
read-NEG-decl.base
"I won't read such a book like this."

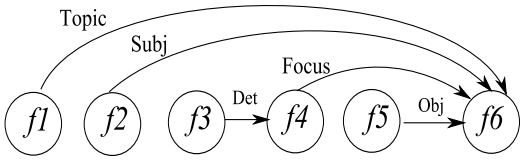

Figure 4.14: DAG for (4.7)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{FOCUS} & \left\{ {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`konnna; such a'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wa'} \\ \text{INDEX} & \text{j} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f5}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{j} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{SUF} & \text{`-ana-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`present'} \\
{}_{f6}\text{MOOD} & \text{`negative'}
\end{bmatrix}
$$

Figure 4.15: F-structure for (4.7)

4.2.3 SUBJ

Noun units followed by the case particle "-ga" have the grammatical function SUBJ (cf. Section 3.6.2). Figure 4.16 is the DAG representation of the dependency among syntactic units in Example (4.8), and Figure 4.17 is the

f-structure generated from Figure 4.16:

(4.8) Watashiga kono honwo yonda.

[watashi-ga]$_{f1}$	*[kono]*$_{f2}$	*[hon-wo]*$_{f3}$	*[yom-ta]*$_{f4}$.
I-SUBJ	this	book-OBJ	read-decl.ta

"I read this book."

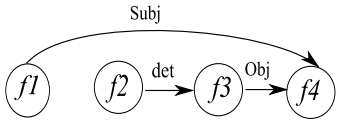

Figure 4.16: DAG for (4.8)

$$
\begin{bmatrix}
\text{SUBJ} & f1\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\
\text{OBJ} & f2\begin{bmatrix} \text{DET} & f3\begin{bmatrix}\text{PRED} & \text{'kono; this'}\end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{MOOD} & \text{'declarative'} \\
\text{TENSE} & \text{'past'}
\end{bmatrix}_{f4}
$$

Figure 4.17: F-structure for (4.8)

In written Japanese, it is rare to find a nominal syntactic unit without any particle which functions as one of the core grammatical functions. However, in spoken Japanese this is sometimes the case, and it is intelligible:

(4.9) Watashi kono hon yonda

How to determine the grammatical function of nominal units without case particles will be an important issue when dealing with spoken Japanese.

For both written and spoken Japanese, if a clause does not have any unit with "-ga", then the subject is a zero pronoun, as the Subject Condition

states that every verbal predicate must have a SUBJ ([Bresnan and Kanerva, 1989], [Dalrymple, 2001]). For Japanese, this condition is satisfied at the level of f-structure, not at the phonological level of representation:

(4.10) Kono honwo yonda.

[pro]$_{f1}$	*[kono]*$_{f2}$	*[hon-wo]*$_{f3}$	*[yom-ta]*$_{4}$
pro-SUBJ	this	book-OBJ	read-decl.ta

"Someone read this book."

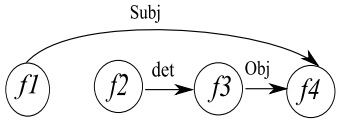

Figure 4.18: DAG for (4.10)

$$
\begin{bmatrix}
\text{SUBJ} & \underset{f1}{\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix}} \\
\text{OBJ} & \underset{f3}{\begin{bmatrix} \text{DET} & \underset{f2}{\begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix}} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix}} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{MOOD} & \text{'declarative'} \\
\text{TENSE} & \text{'past'}
\end{bmatrix}_{f4}
$$

Figure 4.19: F-structure for (4.10)

It is often the case that the subject of a clause is topicalised, i.e. the subject unit has the adverbial particle "-wa" at its end. Here the TOPIC shares its referential index with the SUBJ pro:

(4.11) Watashiwa kono honwo yonda.

[watashi-wa]$_{f1}$	*[pro]$_{f2}$*	*[kono]$_{f3}$*	*[hon-wo]$_{f4}$*	*[yom-ta]$_{f5}$*
I-TOPIC	pro-SUBJ	this	book-OBJ	read-decl.ta

"I read this book."

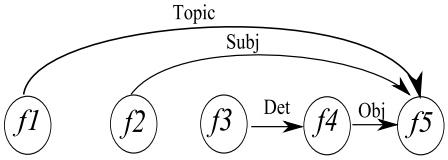

Figure 4.20: DAG for (4.11)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{MOOD} & \text{'declarative'} \\
{}_{f5}\text{TENSE} & \text{'past'}
\end{bmatrix}
$$

Figure 4.21: F-structure for (4.11)

The Meaning of SUBJ

A syntactic unit with SUBJ is often the agent of the event denoted by the main verbal predicate in the active voice, or the theme of the state denoted by the main adjectival predicate:

There are some instances where the SUBJ of the clause (or the syntactic unit which contains the grammatical function SUBJ) is not the agent of the

event. One such instance is a clause in the passive voice. The other is a clause which has a head with a spontaneous or possible verbal suffix (cf. Section 3.5.3.5).

In these cases, it is common that the syntactic unit has the case particle "-ga". These instances support the claim that the case particle "-ga" does not indicate the agent of an action or the theme of a state, but it indicates the grammatical function SUBJ, whatever thematic role it corresponds to.

4.2.4 OBJ

Noun units followed by the case particle "-wo" have the grammatical function OBJ, as we have already seen in examples in the previous sections and in Chapter 3.

As in the case of the grammatical function SUBJ, there are many instances of zero pronouns which would be assigned with the grammatical function OBJ, if they were phonologically or morphologically absent. The referent of such zero pronouns is something that is taken for granted by the speaker and the hearer in the context, or something the speaker believes that is taken for granted:

(4.12) Watashiga yonda.

[watashi-ga]$_{f1}$	*[pro]*$_{f2}$	*[yom-ta]*$_{f3}$
I-SUBJ	pro-OBJ	read-decl.ta

"I read (something)."

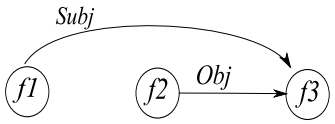

Figure 4.22: DAG for (4.12)

$$
f3\begin{bmatrix}
\text{SUBJ} & f1\begin{bmatrix}\text{PRED} & \text{`watashi; I'}\end{bmatrix} \\
\text{OBJ} & f2\begin{bmatrix}\text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i}\end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{MOOD} & \text{`declarative'} \\
\text{TENSE} & \text{`past'}
\end{bmatrix}
$$

Figure 4.23: F-structure for (4.12)

The problem is that it is difficult to distinguish verbs which do not subcategorise for a wo-marked noun unit (hence OBJ) from transitive verbs whose object is zero-pronominalised, based on the syntactic environment presented in a particular sentence alone.[1] Unlike identifying SUBJ zero pronouns, simply supposing the presence of OBJ zero pronoun in all clauses which do not have a wo-marked noun unit leads to incorrect analyses postulating OBJ zero pronouns for intransitive verbs. In (4.13), the verb "agar" is an intransitive verb, and the absence of a wo-marked verb noun does not cause violation of the Completeness Condition (cf. Chapter 2):

(4.13) Sekiyuno nedanga mata agatta.

[sekiyu-no]$_{f1}$	*[nedan-ga]*$_{f2}$	*[mata]*$_{f3}$	*[agar-ta]*$_{f4}$
oil-PADJ	price-SUBJ	again	rise-decl.ta

"The price of oil has risen again."

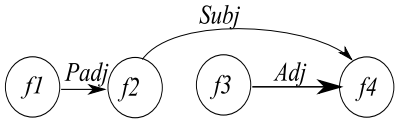

Figure 4.24: DAG for (4.13)

[1] Of course this task would be easier if we had access to comprehensive subcategorisation information for all Japanese verbs.

$$
\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix} \text{PADJ} & \{ \begin{bmatrix} \text{PRED} & \text{`sekiyu; oil'} \\ \text{PRTCNJ} & \text{`-no'} \end{bmatrix}_{f1} \} \\ \text{PRED} & \text{`nedan; price'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix}_{f2} \\
\text{ADJ} & \{ \begin{bmatrix} \text{PRED} & \text{`mata; again'} \end{bmatrix}_{f3} \} \\
\text{PRED} & \text{`agar-}\langle \text{SUBJ} \rangle\text{; rise'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f4}
$$

Figure 4.25: F-structure for (4.13)

If, on the other hand, we do nothing about OBJ zero pronouns, then we produce incorrect analyses overlooking OBJ zero pronouns for transitive verbs which do not have an overt OBJ noun unit. In the absence of exhaustive subcategorisation information, the distinction must be made not by local syntactic environment, but other linguistic features such as the morphology of the verb (cf. Section 3.5.2), the semantics of the verb, and the context in which a given sentence appears.

The Meaning of OBJ

OBJ is less ambiguous than other grammatical functions. A syntactic unit with the grammatical function OBJ is usually the theme of the event expressed by the main verbal predicate of the clause in which it appears.

4.2.5 OBL

The grammatical function OBL is specified by the case particle "-ni" attached to a noun unit.

The Meaning of OBL

A noun unit with OBL has one of the following meanings:

(4.14) The receiver in an event of giving:

Watashiga imoutoni kono honwo kashita.

[watashi-ga]$_{f1}$	*[imouto-ni]*$_{f2}$	*[kono]*$_{f3}$	*[hon-wo]*$_{f4}$
I-SUBJ	younger.sister-OBL	this	book-OBJ

[kas-ta]$_{f5}$
lend-decl.ta
"I lent this book to my younger sister."

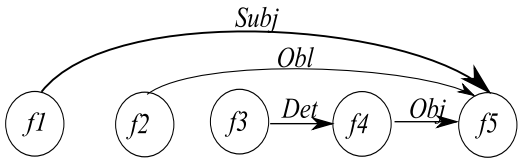

Figure 4.26: DAG for (4.14)

$$
\begin{bmatrix}
\text{SUBJ} & {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\
\text{OBL} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'imoto; younger sister'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'kas-}\langle \text{SUBJ, OBJ, OBL} \rangle\text{; lend'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{TENSE} & \text{'past'} \\
{}_{f5}\text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.27: F-structure for (4.14)

(4.15) The goal of an event of moving:

Kinou Uenoni itta.

[pro]$_{f1}$	*[kinou]*$_{f2}$	*[Ueno-ni]*$_{f3}$	*[ik-ta]*$_{f4}$
pro-SUBJ	yesterday	Ueno-OBL	go-decl.ta

"(Someone) went to Ueno yesterday."

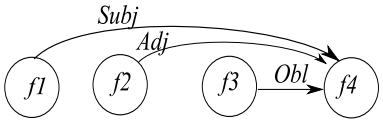

Figure 4.28: DAG for (4.15)

$$
\begin{bmatrix}
\text{SUBJ} & f1\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{ADJ} & \left\{ f2\begin{bmatrix} \text{PRED} & \text{'kinou; yesterday'} \end{bmatrix} \right\} \\
\text{OBL} & f3\begin{bmatrix} \text{PRED} & \text{'Ueno'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{PRED} & \text{'ik-}\langle \text{SUBJ, OBL} \rangle\text{; go'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{TENSE} & \text{'past'} \\
f4\text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.29: F-structure for (4.15)

(4.16) The agent of a passive sentence:

Watashino kabinga aneni kowasareta.

[watashi-no]$_{f1}$	*[kabin-ga]*$_{f2}$	*[ane-ni]*$_{f3}$	*[kowas-are-ta]*$_{f4}$
I-PADJ	vase-SUBJ	elder.sister-OBL	break-passive-decl.ta

"This vase was broken by my elder sister."

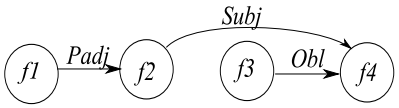

Figure 4.30: DAG for (4.16)

$$
\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix}
\text{PADJ} & {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTCJ} & \text{'-no'} \end{bmatrix} \\
\text{PRED} & \text{'kabin;vase'} \\
\text{PRTCS} & \text{'-ga'}
\end{bmatrix}_{f2} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{'ane; elder sister'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix}_{f3} \\
\text{PRED} & \text{'kowas-}\langle \text{SUBJ, OBL} \rangle\text{; break'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'zero'} \\
\text{SUF} & \text{'-are-'} \\
\text{SUF-INFL} & \text{'-ta'} \\
\text{VOICE} & \text{'passive'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}_{f4}
$$

Figure 4.31: F-structure for (4.16)

(4.17) The causee of a causative sentence:

Watashiwa ototoni sono honwo yomaseta.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[ototo-ni]*$_{f3}$	*[sono]*$_{f4}$
I-TOP	pro-SUBJ	younger.brother-OBL	the
[hon-wo]$_{f5}$	*[yom-ase-ta]*$_{f6}$		
book-OBL	read-causative-decl.ta		

"I had my younger brother read the book."

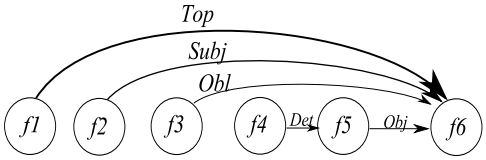

Figure 4.32: DAG for (4.17)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`ototo; younger brother'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{OBJ} & {}_{f5}\begin{bmatrix} \text{DET} & {}_{f4}\begin{bmatrix} \text{PRED} & \text{`sono; the'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-ase-}\langle \text{SUBJ, OBJ, OBL} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`zero'} \\
\text{SUF} & \text{`-ase-'} \\
\text{SUF-INFL} & \text{`-ta'} \\
\text{VOICE} & \text{`causative'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f6}
$$

Figure 4.33: F-structure for (4.17)

In order to make the function name more precise, it would be better to distinguish OBL which is mapped onto the agent role and OBL which is mapped onto a thematic role other than agent. However, this study does not choose this, for the sake of simplicity.

That fact that a noun unit with "-ni" can mean the agent of a passive or a causative is the reason why this particle is considered to specify one of the core functions. Other case particles, except for "-ga", are not used for

specifying the agent of a passive, a causative or any construction, hence they are peripheral.

4.2.6 PADJ

Noun units followed by one of the case particles other than “-ga”, “-wo” or “-ni” have the grammatical function PADJ (Postpositional ADJunct). Each of the PADJ case particles has various meanings. For example, ”-kara” has the following meanings ([Masuoka and Takubo, 1992]):

- Starting point
- The source of information
- The data for a judgement
- The date of an event
- The material of a product

(4.18) Starting point: Ekikara ieni mukatta.

[pro]$_{f1}$	*[eki-kara]*$_{f2}$	*[ie-ni]*$_{f3}$	*[mukaw-ta]*$_{f4}$.
pro-SUBJ	station-PADJ	house-OBL	go-decl.ta

"I went home from the station."

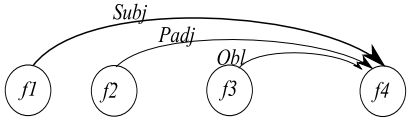

Figure 4.34: DAG for (4.18)

$$
\begin{bmatrix}
\text{SUBJ} & \underset{f1}{\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix}} \\
\text{PADJ} & \left\{ \underset{f2}{\begin{bmatrix} \text{PRED} & \text{'eki; station'} \\ \text{PRTCS} & \text{'-kara'} \end{bmatrix}} \right\} \\
\text{OBL} & \underset{f3}{\begin{bmatrix} \text{PRED} & \text{'ie; home'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix}} \\
\text{PRED} & \text{'mukaw-}\langle \text{SUBJ, OBL} \rangle\text{; go to'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{TENSE} & \text{'past'} \\
\underset{f4}{} \text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.35: F-structure for (4.18)

(4.19) The source of information:

Watashikara anatani hanashiga aru.

[watashi-kara]$_{f1}$	*[anata-ni]*$_{f2}$	*[hanashi-ga]*$_{f3}$	*[ar-u]*$_{f4}$
I-PADJ	you.(polite)-OBL	talk-SUBJ	exist-decl.base

(Literally)"There is something to talk from me to you."
"I have something to tell you."

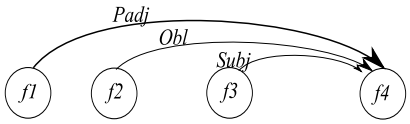

Figure 4.36: DAG for (4.19)

$$
\begin{bmatrix}
\text{PADJ} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTCS} & \text{'-kara'} \end{bmatrix} \right\} \\
\text{OBL} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'anata; you}\big(\text{POLITE}\big)\text{'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{'hanashi; talk'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\
\text{PRED} & \text{'ar-}\big\langle \text{SUBJ, OBL} \big\rangle\text{; exist'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'nonpast'} \\
{}_{f4}\text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.37: F-structure for (4.19)

(4.20) The data for a judgement:

Samazamana jijitsukara ketsuronwo michibikidashita.

[pro]$_{f1}$	*[samazama-na]*$_{f2}$	*[jijitsu-kara]*$_{f3}$	*[ketsuron-wo]*$_{f4}$
pro	various-att.base.plain	fact-PADJ	conclusion-OBJ

[michibikidas-ta]$_{f5}$
induce-decl.ta

"(Someone) induced a conclusion from various facts"

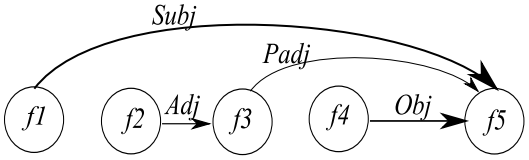

Figure 4.38: DAG for (4.20)

$$
\begin{bmatrix}
\text{SUBJ}_{f1} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PADJ} & \left\{ \begin{bmatrix} \text{ADJ}_{f2} & \begin{bmatrix} \text{PRED} & \text{'samazama; various'} \\ \text{ADJ-INFL} & \text{'-na'} \end{bmatrix} \\ \text{PRED} & \text{'jijitsu; fact'} \\ \text{PRTCS}_{f3} & \text{'-kara'} \end{bmatrix} \right\} \\
\text{OBJ}_{f4} & \begin{bmatrix} \text{PRED} & \text{'ketsuron; conclusion'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'michibikidas-}\langle \text{SUBJ, OBJ} \rangle\text{; induce'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{TENSE}_{f5} & \text{'past'}
\end{bmatrix}
$$

Figure 4.39: F-structure for (4.20)

(4.21) The date of an event:

Ashitakara kakikyukada.

[pro]$_{f1}$ *[ashita-kara]*$_{f2}$ *[kakikyuka-da]*$_{f3}$
pro-SUBJ tomorrow-PADJ summer.vacation-copula.plain.decl.base
(Literally) "(It) is the summer vacation from tomorrow."
"The summer vacation will start tomorrow."

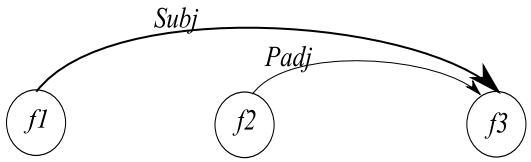

Figure 4.40: DAG for (4.21)

$$
\begin{bmatrix}
\text{SUBJ} & {}_{f1}\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PADJ} & \left\{ {}_{f2}\begin{bmatrix} \text{PRED} & \text{'ashita; tomorrow'} \\ \text{PRTCS} & \text{'-kara'} \end{bmatrix} \right\} \\
\text{PRED} & \text{'kakikyuka; summar vacation'} \\
\text{ROOT} & \text{'+'} \\
\text{COPULA} & \text{'des-; be'} \\
\text{COPULA-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'nonpast'} \\
{}_{f3}\text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.41: F-structure for (4.21)

(4.22) The material of a product:

Nihonshuwa komekara tsukurareru.

[nihonshu-wa]$_{f1}$ *[pro]*$_{f2}$ *[kome-kara]*$_{f3}$
Japanese.sake-TOP pro-SUBJ rice-PADJ
[tsukur-are-ru]$_{f4}$
make-passive-decl.base
"Japanese sake is made of rice."

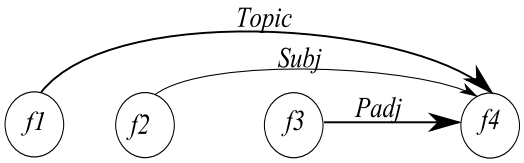

Figure 4.42: DAG for (4.22)

$$
\begin{bmatrix}
\textsc{topic} & \left\{ {}_{f1}\begin{bmatrix} \textsc{pred} & \text{'Nihonshu; Japanese sake'} \\ \textsc{prtcs} & \text{'wa'} \\ \textsc{index} & \text{i} \end{bmatrix} \right\} \\
\textsc{subj} & {}_{f2}\begin{bmatrix} \textsc{pred} & \text{'pro'} \\ \textsc{prontype} & \text{'zero'} \\ \textsc{index} & \text{i} \end{bmatrix} \\
\textsc{padj} & \left\{ {}_{f3}\begin{bmatrix} \textsc{pred} & \text{'kome; rice'} \\ \textsc{prtcs} & \text{'-kara'} \end{bmatrix} \right\} \\
\textsc{pred} & \text{'tsukur-}\langle \textsc{subj}, \textsc{obj} \rangle\text{; make'} \\
\textsc{suf} & \text{'-are-'} \\
\textsc{suf-infl} & \text{'-ru'} \\
\textsc{tense} & \text{'nonpast'} \\
\textsc{mood} & \text{'declarative'} \\
{}_{f4}\textsc{voice} & \text{'passive'}
\end{bmatrix}
$$

Figure 4.43: F-structure for (4.22)

Adjunct or Not?

In this study, I assume that noun units with case particles other than "-ga", "-wo" and "-ni" are not subcategorised for by the verb, hence the name PADJ, or PostpositionalADJunct is given to them.

In principle, it would be possible to analyse them as "governable grammatical functions", which are subcategorised for by the predicate ([Dalrymple, 2001]). However, such an analysis would have to consider the semantics of each noun unit and the predicate. When it comes to the automatic annotation of grammatical function, this semantic consideration would make the annotation process complex and liable to output incorrect annotations for noun units with PADJ case particles.

If probabilistic approaches to automatic case-frame extraction find that certain verbs have a high probability to appear with a noun unit attached with one of the PADJ case particles, then these verbs seem to subcategorise for a noun unit with this particle, and this constitutes important information provided by a probabilistic approach to linguistic data. In such cases, the name of the grammatical function should be something other than PADJ, since it is no longer an adjunct; however, a problem remains how high the coocurrence probability should be for such units to have the status of arguments, not adjuncts. The notion of "governable grammatical function" must be substantiated through further study, especially for "pro-drop" languages including Japanese.

In order to avoid these problems, in this study I have chosen to treat noun units attached with case particles other than "-ga", "-wo", and "-ni" as having the grammatical function PADJ, while not completely excluding the possibility that this "adjunct" can be found being subcategorised for by the verb it depends on.

4.2.7 ADJ

The grammatical function ADJ is an abbreviation for ADJunct; it is instantiated by adverbs, adjectives in the adverbial form (cf. Section 3.5.2), non-inflecting adjectives modifying a noun, or noun units which have no particle at their right periphery. A unit with the function ADJ modifies the unit it depends on.

(4.23) ADJ projected by an adverb:
Watashiwa kono honwo shocchuu yondeiru.

$[watashi\text{-}wa]_{f1}$	$[pro]_{f2}$	$[kono]_{f3}$	$[hon\text{-}wo]_{f4}$	$[shocchuu]_{f5}$
I-TOP	pro-SUBJ	this	book-OBJ	frequently

[yom-te-ir-u]$_{f6}$
read-conn.ta-suff-decl.base
"I read this book frequently."

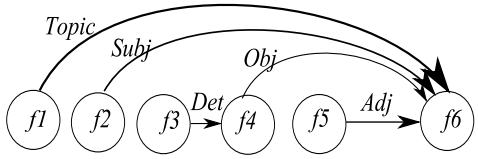

Figure 4.44: DAG for (4.23)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\left[\text{PRED} \quad \text{'kono; this'}\right] \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{ADJ} & \left\{ {}_{f5}\left[\text{PRED} \quad \text{'shocchuu; frequently'}\right] \right\} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-te-'} \\
\text{SUF} & \text{'-ir-'} \\
\text{SUF-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'present'} \\
\text{MOOD} & \text{'declarative'} \\
{}_{f6}\text{ASPECT} & \text{'progressive'}
\end{bmatrix}
$$

Figure 4.45: F-structure for (4.23)

(4.24) ADJ projected by an adjective in the adverbial form:

Watashiwa kono honwo yoku yondeiru.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[kono]*$_{f3}$	*[hon-wo]*$_{f4}$
I-TOP	pro-SUBJ	this	book-OBJ
[yok-u]$_{f5}$		*[yom-te-ir-u]*$_{f6}$	
good/often-adv.base		read-conn.ta-suff-decl.base	

"I often read this book."

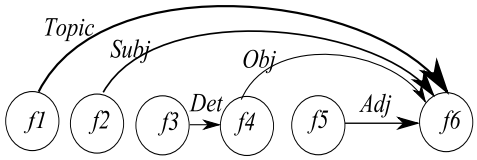

Figure 4.46: DAG for (4.24)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{ADJ} & \left\{ {}_{f5}\begin{bmatrix} \text{PRED} & \text{`yok-; good/often'} \\ \text{ADJ-INFL} & \text{`-u'} \end{bmatrix} \right\} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-te-'} \\
\text{SUF} & \text{`-ir-'} \\
\text{SUF-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`present'} \\
\text{MOOD} & \text{`declarative'} \\
{}_{f6}\text{ASPECT} & \text{`progressive'}
\end{bmatrix}
$$

Figure 4.47: F-structure for (4.24)

(4.25) ADJ projected by a noun unit with no particle:
Watashiwa kono honwo mainichi yondeiru.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[kono]*$_{f3}$	*[hon-wo]*$_{f4}$	*[mainichi]*$_{f5}$
I-TOP	pro-SUBJ	this	book-OBJ	every.day

[yom-te-ir-u]$_{f6}$
read-conn.ta-suff-decl.base
"I read this book every day."

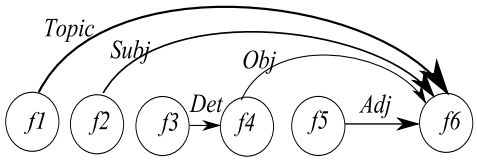

Figure 4.48: DAG for (4.25)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix} \text{DET} & {}_{f3}\left[\text{PRED} \quad \text{'kono; this'}\right] \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{ADJ} & \left\{ {}_{f5}\left[\text{PRED} \quad \text{'mainichi; every day'}\right] \right\} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-te-'} \\
\text{SUF} & \text{'-ir-'} \\
\text{SUF-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'present'} \\
\text{MOOD} & \text{'declarative'} \\
\text{ASPECT} & \text{'progressive'}
\end{bmatrix}_{f6}
$$

Figure 4.49: F-structure for (4.25)

4.2.8 DET

In Japanese, the grammatical function DET is projected by adjectival demonstrative pronouns (cf. Section 3.6.3). Since they appear without particle, it is possible to analyse them as having the grammatical function ADJ. However, they are concerned with the definiteness of what they depend on, not just modifying their local heads; therefore, they are given the grammatical function DETerminer, not ADJ.

The demonstrative pronoun "kono" has already been frequently used in the examples so far. Some adjectival demonstrative pronouns express the speaker's subjective judgement on a noun, or interrogative adjectives:

(4.26) Speaker's judgement on a noun; e.g., "konna":

Konna honwa yomanai.

$[pro]_{f1}$	$[konna]_{f2}$	$[hon\text{-}wa]_{f3}$	$[pro]_{f3}$	$[yom\text{-}ana\text{-}i]_{f7}$
pro-SUBJ	such.a	book-FOC	pro-OBJ	read-NEG-decl.base

"I won't read such a book like this."

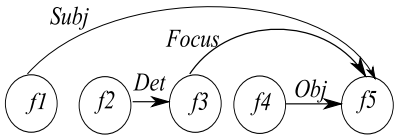

Figure 4.50: DAG for (4.26)

$$
\begin{bmatrix}
\text{FOCUS} & \left\{ \begin{bmatrix} \text{DET} & {}_{f1}\begin{bmatrix}\text{PRED} & \text{`konna; such a '}\end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f2} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix}\text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{j}\end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix}\text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i}\end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{SUF} & \text{`-ana-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`present'} \\
{}_{f5}\text{MOOD} & \text{`negative'}
\end{bmatrix}
$$

Figure 4.51: F-structure for (4.26)

(4.27) Interrogative adjective; e.g., "donna":

Kimiwa donna honga sukidesuka?

[kimi-wa]$_{f1}$ *[donna]*$_{f2}$ *[hon-ga]*$_{f3}$ *[suki-desu-ka]*$_{f4}$
you-TOP what.kind.of book-SUBJ like-aux-Q
(Literally) "As for you, what kind of book is (your) favorite?" or "What kind of book do you like?"

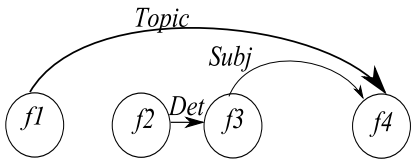

Figure 4.52: DAG for (4.27)

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ {}_{f1}\begin{bmatrix} \text{PRED} & \text{`kimi; you'} \\ \text{PRTADV} & \text{`-wa'} \end{bmatrix} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix} \text{DET} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{`donna; what kind of'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\
\text{PRED} & \text{`suki}\langle \text{SUBJ} \rangle\text{; favorite'} \\
\text{AUX} & \text{`-des-'} \\
\text{AUX-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`present'} \\
{}_{f4}\text{MOOD} & \text{`interrogative'}
\end{bmatrix}
$$

Figure 4.53: F-structure for (4.27)

4.2.9 REL

Relative clauses project the grammatical function REL*ative clause*. The head of a relative clause can be a verb or an adjective. A relative clause depends on a noun unit, and the head verb or adjective has a zero pronoun as one of its arguments which refers to this noun unit.

Relative clauses are such that the last morpheme of its head verb unit ends with the base form or the ta form, and it is not the root of a sentence. The last morpheme can be a verb, a verbal suffix, an auxiliary or an adjective. One of the core arguments of the head of a relative clause is a zero pronoun which refers to the noun on which the relative clause depends:

(4.28) Watashiga kinou yonda honga miataranai.

[watashi-ga]$_{f1}$ *[kinou]*$_{f2}$ *[pro]*$_{f3}$ *[yom-ta]*$_{f4}$ *[hon-ga]*$_{f5}$
[I-SUBJ yesterday pro-OBJ read-decl.ta]$_{rel}$ book-SUBJ
[miatar-ana-i]$_{f6}$
can.be.found-NEG-decl.base

(Literally) "The book which I read yesterday cannot be found" or "I cannot find the book which I read yesterday."

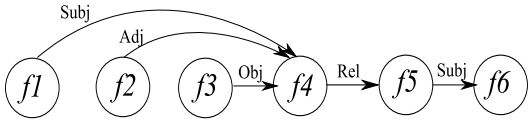

Figure 4.54: DAG for (4.28)

$$
\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix}
\text{REL} & \left\{ \begin{bmatrix}
\text{SUBJ} & {}_{f1}\begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\
\text{ADJ} & \left\{ {}_{f2}\begin{bmatrix} \text{PRED} & \text{`kinou; yesterday'} \end{bmatrix} \right\} \\
\text{OBJ} & {}_{f3}\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f4} \right\} \\
\text{PRED} & \text{`hon; book'} \\
\text{PRTADV} & \text{`-ga'} \\
\text{INDEX} & \text{i}
\end{bmatrix}_{f5} \\
\text{PRED} & \text{`miatar-}\langle \text{SUBJ} \rangle\text{; can be found'} \\
\text{ROOT} & \text{`+'} \\
\text{SUF} & \text{`-ana-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`present'} \\
\text{MOOD} & \text{`negative'}
\end{bmatrix}_{f6}
$$

Figure 4.55: F-structure for (4.28)

When both of the core arguments are zero-pronominalised, then ambiguity arises with regards to which zero pronoun refers to which antecedent. In (4.29), the subject zero pronoun of the relative clause refers to the topic of the main clause, while the object zero pronoun refers to the noun on which the relative clause depends.

(4.29) Watashiwa kinou yonda honwo imotoni ageta.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[pro]*$_{f3}$	*[kinou]*$_{f4}$	*[pro]*$_{f5}$
I-TOP	pro-SUBJ	[pro-SUBJ	yesterday	pro-OBJ
[yom-ta]$_{f6}$	*[hon-wo]*$_{f7}$	*[imoto-ni]*$_{f8}$		*[ager-ta]*$_{f9}$
read-decl.ta]$_{rel}$	book-OBJ	younger.sister-OBL		give-decl.ta

"I gave my sister the book which I had read yesterday."

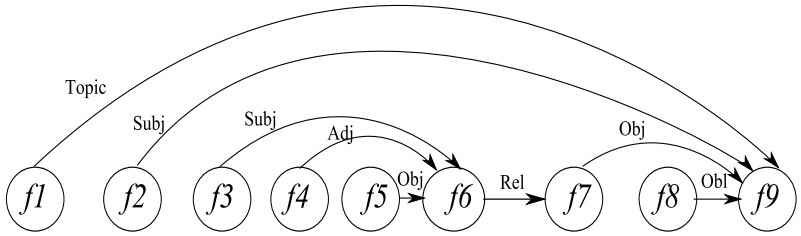

Figure 4.56: DAG for (4.29)

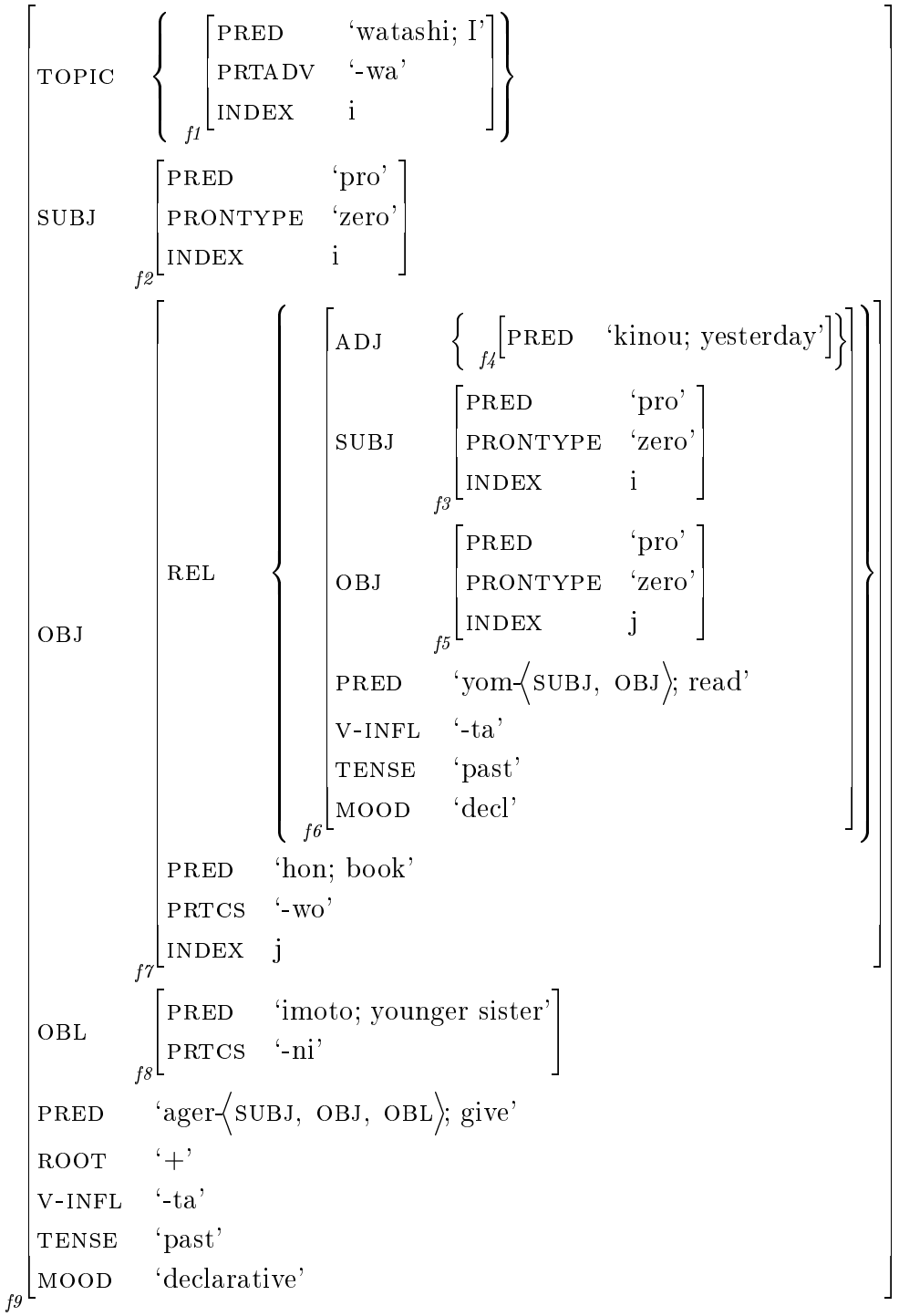

Figure 4.57: F-structure for (4.29)

In (4.30), the object zero pronoun of the relative clause refers to the topic of the main clause, while the subject zero pronoun refers to the noun on which the relative clause depends.

(4.30) kono honwa kinou yonda hitoni yogosareta.

[kono]$_{f1}$	*[hon-wa]*$_{f2}$	*[pro]*$_{f3}$	*[pro]*$_{f4}$	*[kihou]*$_{f5}$	*[pro]*$_{f6}$
this	book-TOP	pro-SUBJ	pro-SUBJ	yesterday	pro-OBJ
[yom-ta]$_{f7}$	*[hito-ni]*$_{f8}$	*[yogos-are-ta]*$_{f9}$			
read-decl.ta	person-OBL	dirty-pass.-decl.ta			

"This book is dirtied by the person who read it yesterday."

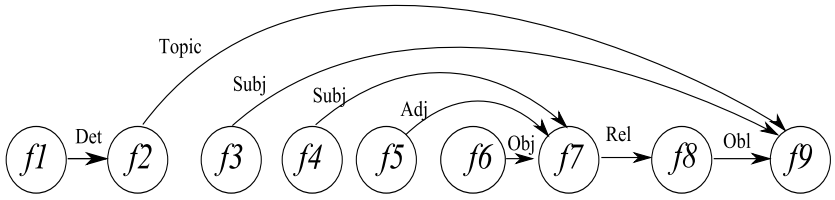

Figure 4.58: DAG for (4.30)

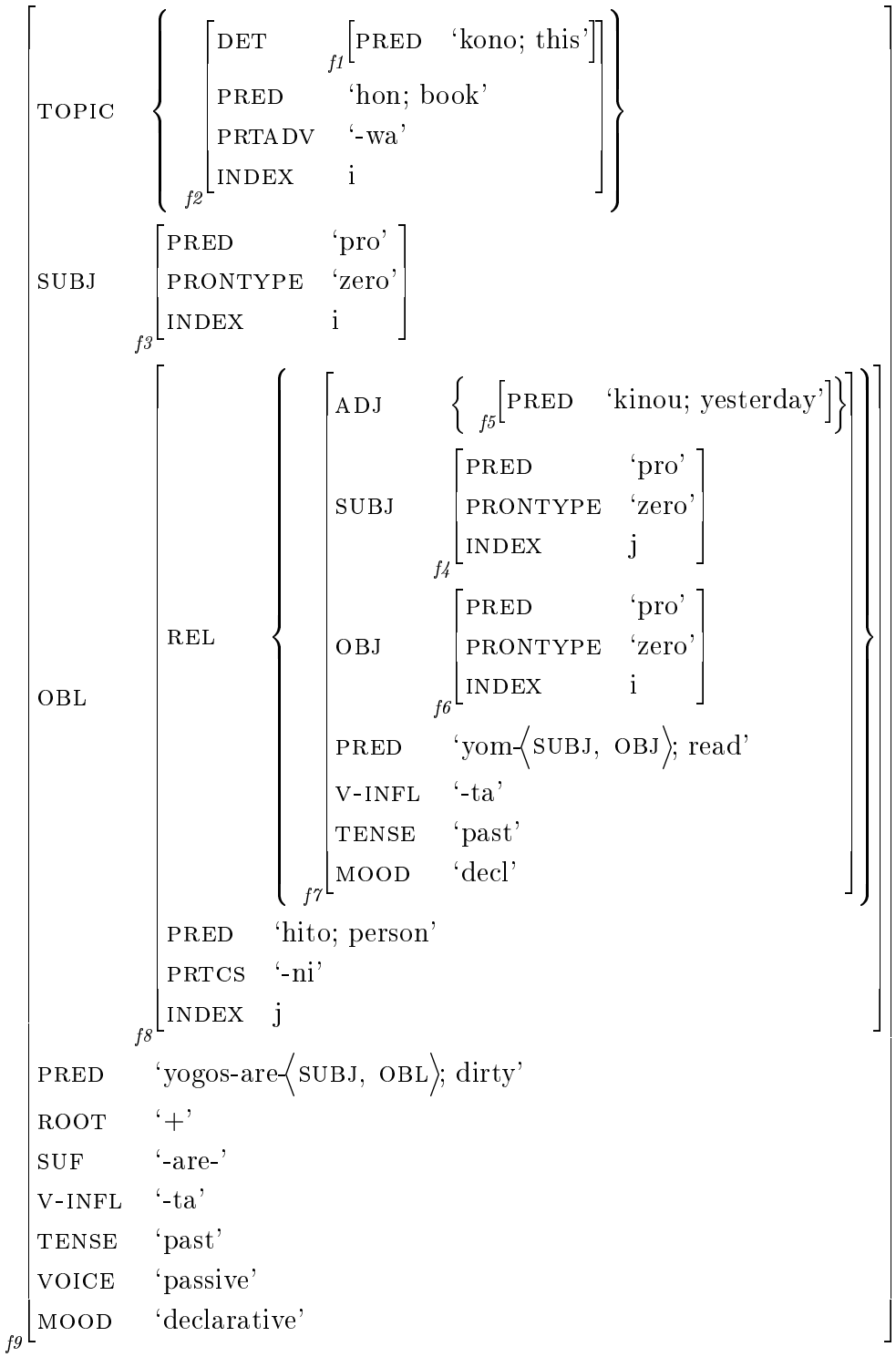

Figure 4.59: F-structure for (4.30)

When the subject of the main clause is topicalised by the adverbial particle "-wa", it can be referred to by the zero pronoun within the relative clause. Notice that the noun unit "watashiga" in (4.28) is within the relative clause, while the noun unit "watashiwa" in (4.29) and "kono honwa" in (4.30) are outside the relative clause. This is because topic units have wider scope as regards dependency relations than other core-grammatical-function units such as subject or object units.

Attributive Adjectives as Relative Clauses

I propose in this study that Japanese attributive adjectives modifying a noun should be treated as relative clauses. This is because Japanese adjectives can be the main predicate of a sentence without copula, just like verbs; Japanese adjectives subcategorise for SUBJ, and they have verblike inflections (cf. Section 3.5.3). In Sentence (4.31), the adjective "omoshiro-i" (interesting) depends on "hon-wo" (book), which is coindexed with the subject PRO of the adjective (Figure 4.61):

(4.31) Watashiwa omoshiroi honwo ototoni kashita.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[pro]*$_{f3}$	*[omoshiro-i]*$_{f4}$
I-TOP	pro-SUBJ	pro-SUBJ	interesting-decl.base
[hon-wo]$_{f5}$	*[ototo-ni]*$_{f6}$		*[kas-ta]*$_{f7}$
book-OBJ	younger.brother-OBL		lend-decl.ta

"I lent an interesting book to my younger brother." or "I lent my younger brother a book which was interesting."

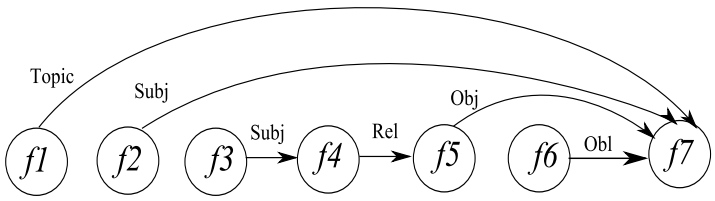

Figure 4.60: DAG for (4.31)

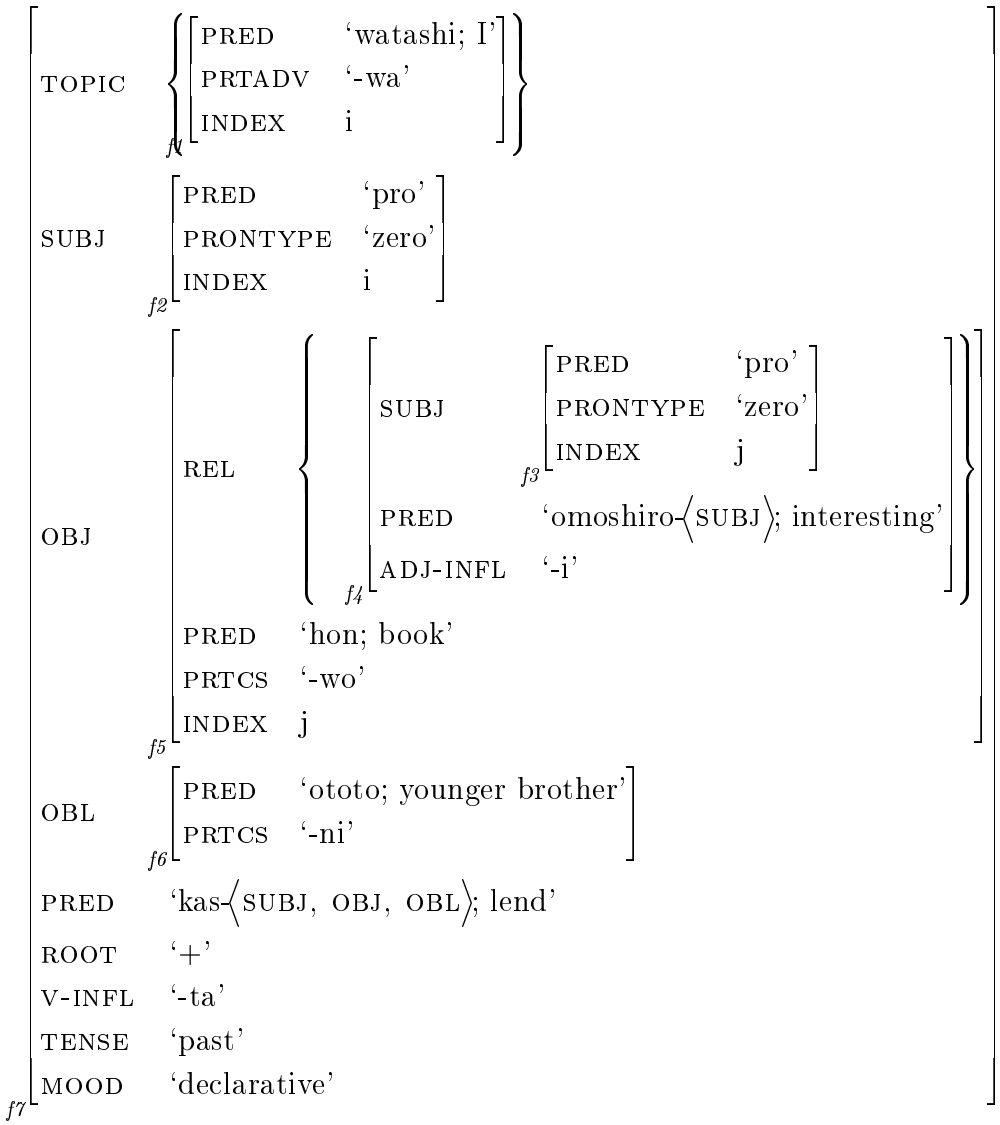

Figure 4.61: F-structure for (4.31)

4.2.10 COMP

The grammatical function COMP labels a subordinate clause followed by a case particle "-to" which is equivalent to an English complementiser "that":

(4.32) Ookina yumewo moteto senseiwa watashitachini itta.

$[pro]_{f1}$	$[ookina]_{f2}$	$[yume\text{-}wo]_{f3}$	$[mot\text{-}e\text{-}to]_{f4}$
pro-SUBJ	big	dream-OBJ	have-imperative-COMP
$[sensei\text{-}wa]_{f5}$	$[pro]_{f6}$	$[watashi\text{-}tachi\text{-}ni]_{f7}$	$[iw\text{-}ta]_{f8}$
sensei-TOP	pro-SUBJ	1st-pl-OBL	say-decl.ta

"Have a big dream", said (our) teacher to us.

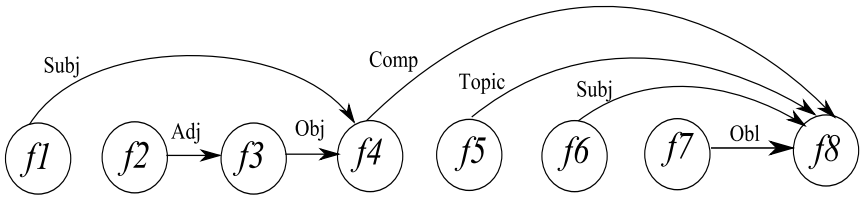

Figure 4.62: DAG for (4.32)

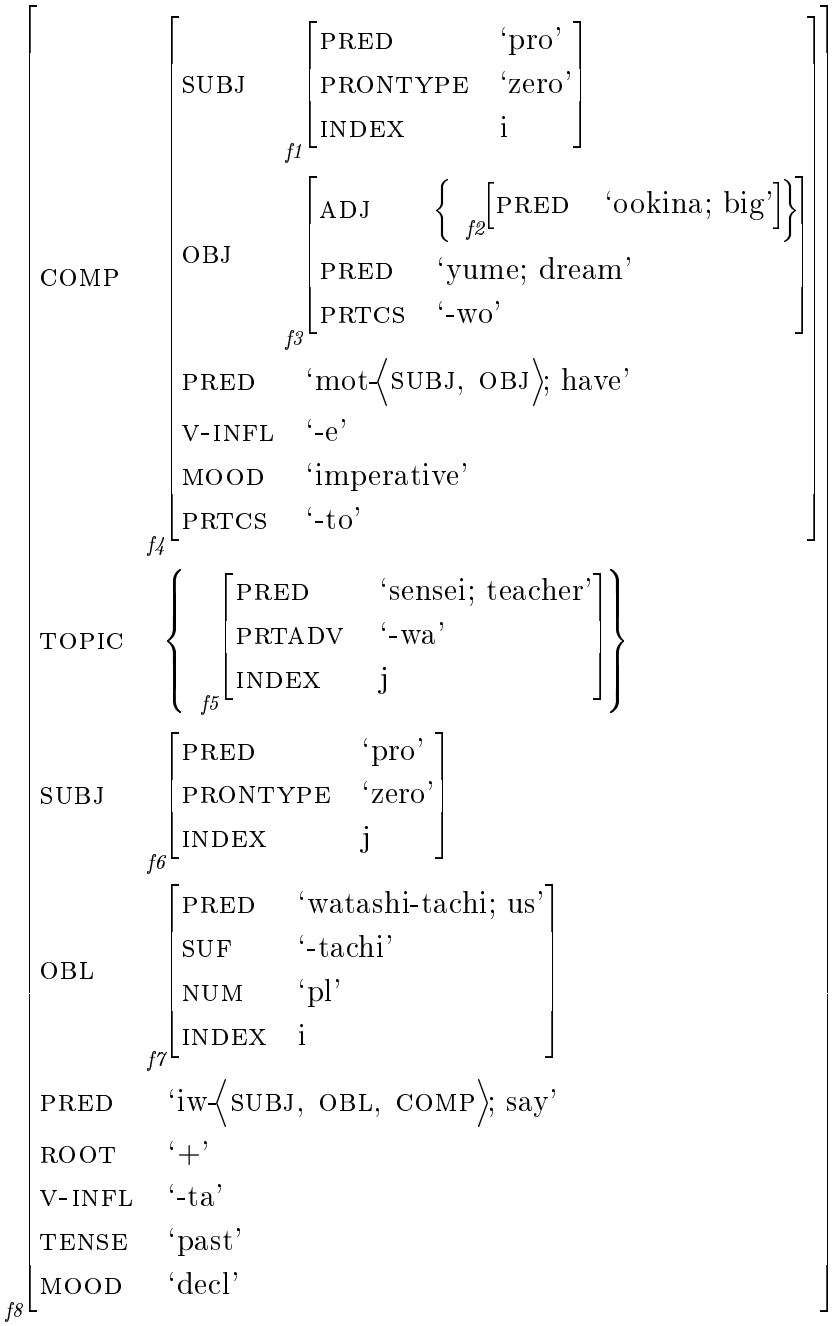

Figure 4.63: F-structure for (4.32)

Appositional COMP

COMP is also the grammatical function of appositional clauses. Appositional clauses are such that the last morpheme of their head verb unit ends with the base form or the ta form, and it is not the root of a sentence. The last morpheme can be a verb, a verbal suffix, an auxiliary or an adjective. None of the core arguments of the head of an apposition clause is related to the noun on which the appositional clause depends. [Tsujimura, 2006] calls this construction "relative clauses without gaps", pointing out there is no "gap" in the clause, unlike normal relative clauses in which there is a filler-gap relationship between the main noun as the filler and one of the missing arguments of the verb as the gap. [Masuoka and Takubo, 1992] call this construction "Naiyou-setsu (Content clauses)", pointing out the fact that the clause in this construction describes the content of the noun on which this clause depends:

(4.33) Hahawa watashiga ototoni kono honwo kashita jijitsuwo shiranakatta.

[Haha-wa]$_{f1}$	*[pro]*$_{f2}$	*[watashi-ga]*$_{f3}$	*[ototo-ni]*$_{f4}$	*[kono]*$f5$
mother-TOP	pro-SUBJ	I-SUBJ	younger.brother	this
[hon-wo]$_{f6}$	*[kas-ta]*$_{f7}$	*[jijitsu-wo]*$_{f8}$	*[shir-anakatta]*$_{f9}$	
book-OBJ	lend-decl.ta	fact-OBJ	know-NEG.past	

"My mother didn't know the fact that I had lent the book to my younger brother."

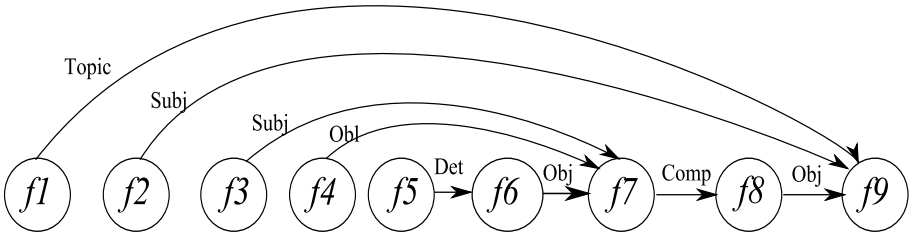

Figure 4.64: DAG for (4.33)

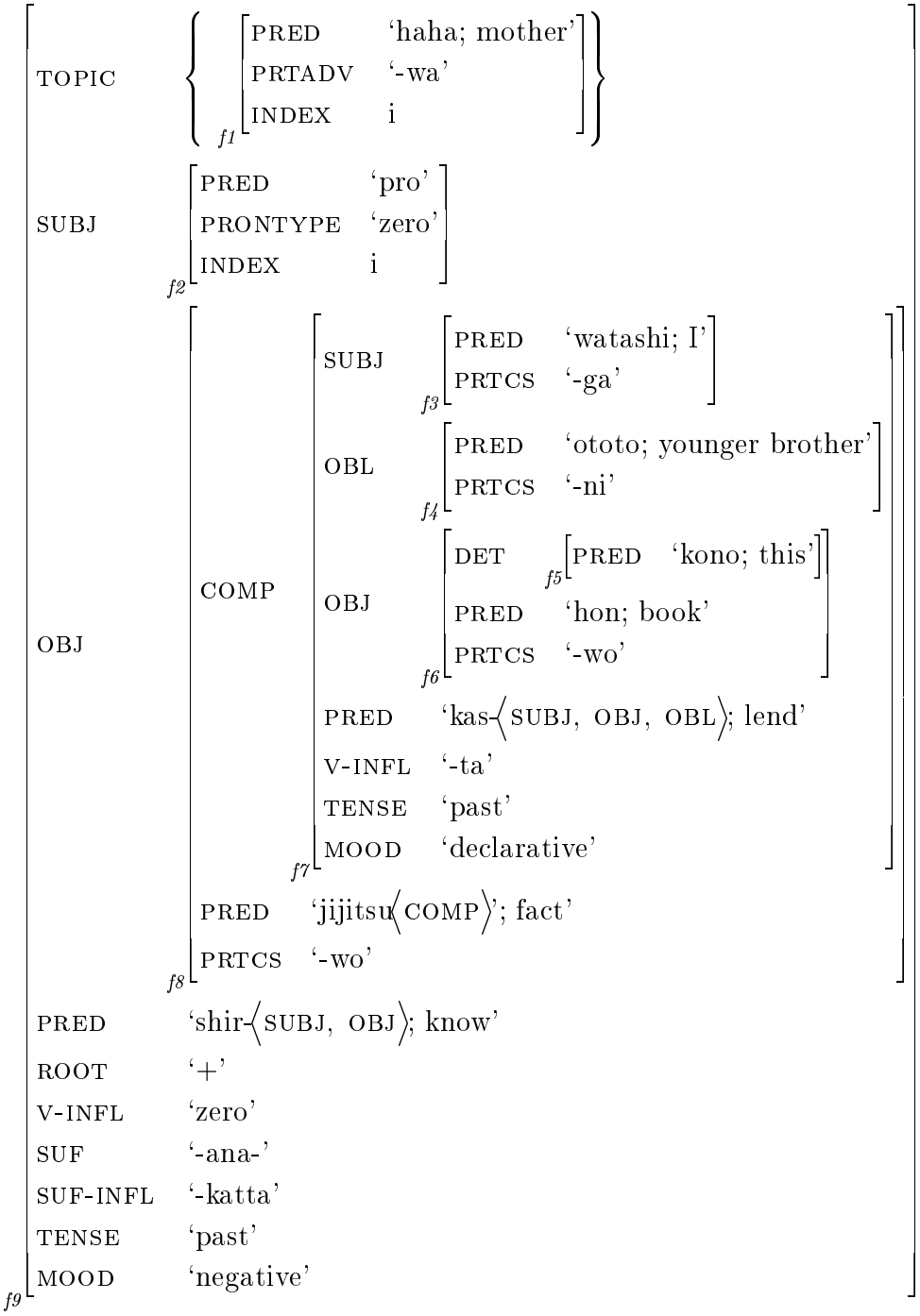

Figure 4.65: F-structure for (4.33)

As for the grammatical function for such appositional constructions, ambiguity arises between REL and COMP in terms of zero pronoun resolution. For example, consider (4.34):

(4.34) Gakkaiwa karega mitsuketa jijitsuwo hiteishita.

[gakkai-wa]$_{f1}$	*[pro]*$_{f2}$	*[kare-ga]*$_{f3}$	*[pro]*$_{f4}$	*[mitsuke-ta]*$_{f5}$
association-TOP	pro-SUBJ	3rd.sg.-SUBJ	pro-OBJ	find-decl.ta
[jijitsu-wo]$_{f6}$	*[hitei-shi-ta]*$_{f7}$			
fact-OBJ	denial-do-decl.ta			

"The association denied the fact that he found" or
"The association denied the fact that he found it."

If a verb unit depends on a noun unit, and a zero pronoun which refers to this noun unit is a dependent of this verb unit, then the grammatical function of this verb unit is REL.

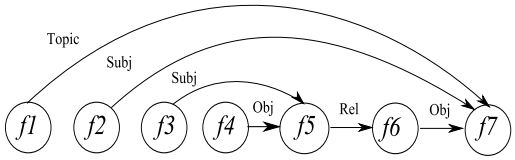

Figure 4.66: DAG for (4.34) with "REL"

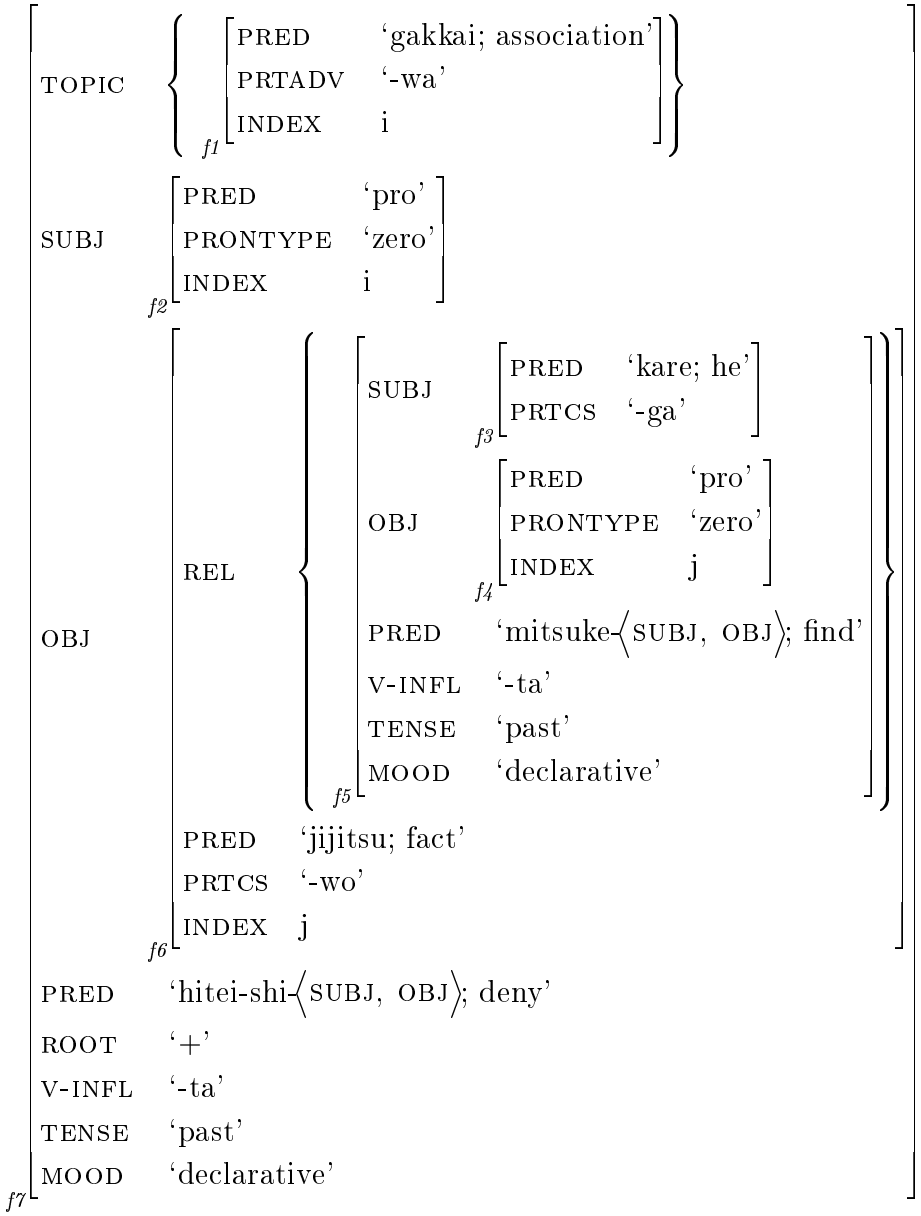

Figure 4.67: F-structure for (4.34) with "REL"

If, on the other hand, there is no zero pronoun depending on the head verb of an appositional clause (namely, all the arguments of the verb are physically realised), or no zero pronoun in the appositional clause refers to the noun unit on which the construction depends, then this verb unit has the grammatical function COMP, and is an apposition to this noun.

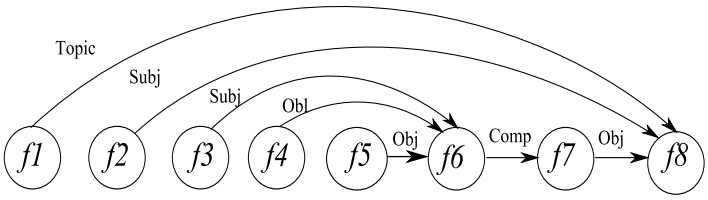

Figure 4.68: DAG for (4.34) with "COMP"

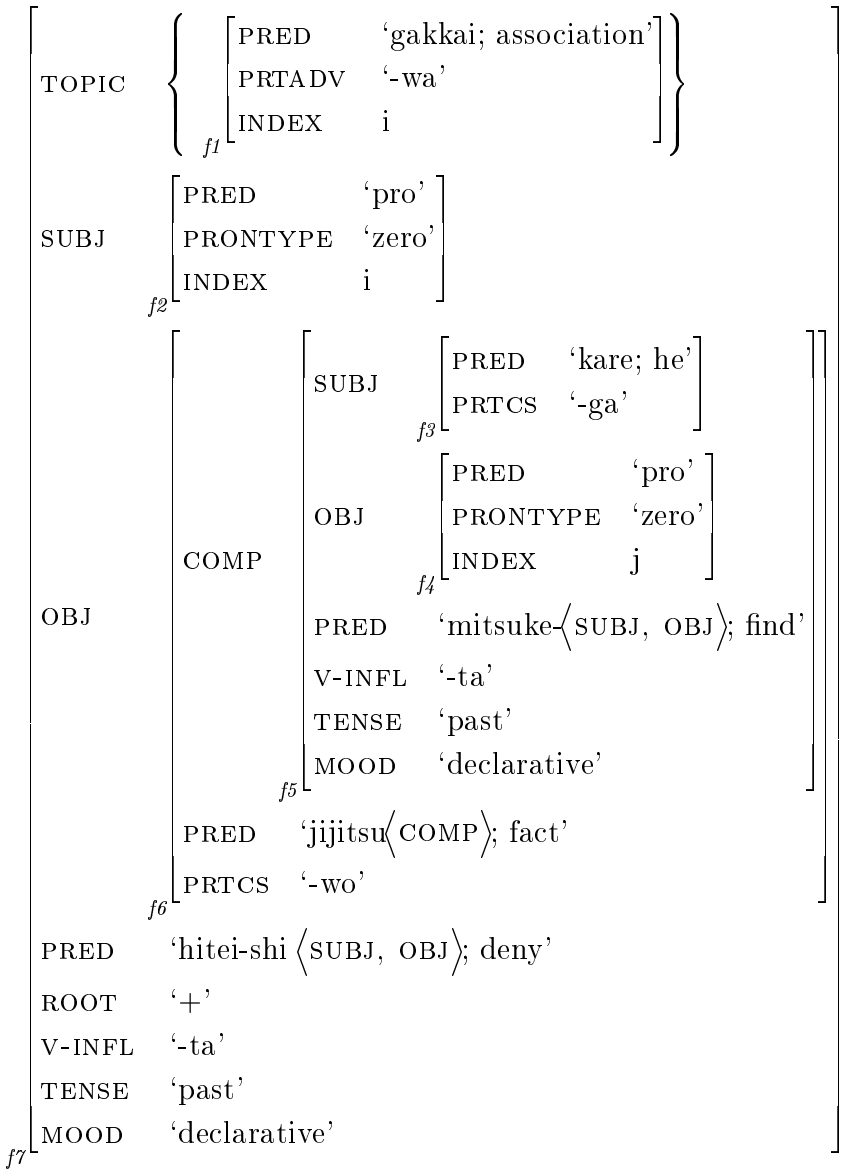

Figure 4.69: F-structure for (4.34) with "COMP"

The distinction between REL and appositional COMP depends on zero-pronoun resolution of unrealised arguments of the verbal head of a clause dependent on a noun. One of the approaches to approximate the distinction between COMP and REL is to find nouns which often take appositional clauses under certain syntactic or semantic circumstances in a given corpus, and if one of these nouns appears after a clause in the same circumstance in new, unseen text, then analyse this clause as appositional to this noun. For example, the observation that formal nouns (cf. 3.6.1) take an

appositional clause with a high probability in a training corpus leads to the analysis that they also take an appositional clause in the test corpus. This approach has been taken by KNP, which includes some rules for the distinction which explicitly specify the nouns which often take an appositional clause. In addition, this approach has also been studied probabilistically ([Fujimoto et al., 2002], [Kawahara and Kurohashi, 2002]).

4.2.11 SADJ

Sentential adjuncts or SADJ are the grammatical functions which are assigned to verbal units whose head has inflections other than the base form or the -ta form (cf. Section 3.5.1), and which depend on another verbal unit. The head of SADJ has the connective form of an inflecting part of speech (4.35), the conditional form of an inflecting part of speech (4.36) or a conjunctive particle attached to it (4.37):

(4.35) Kono honwo yonde jinseiga kawatta.

[pro]$_{f1}$	*[kono]*$_{f2}$	*[hon-wo]*$_{f3}$	*[yom-te]*$_{f4}$	*[jinsei-ga]*$_{f5}$
pro-SUBJ	this	book-OBJ	read-conn.ta	life-SUBJ
[kawar-ta]$_{f6}$				
change-decl.ta				

"After reading this book, (my) life changed."

(4.36) Kono honwo yomeba jinseiga kawaru.

[pro]$_{f1}$	*[kono]*$_{f2}$	*[hon-wo]*$_{f3}$	*[yom-eba]*$_{f4}$	*[jinsei-ga]*$_{f5}$
pro-SUBJ	this	book-OBJ	read-cond.base	life-SUBJ
[kawar-u]$_{f6}$				
change-decl.base				

"If one reads this book, (his or her) life will change"

(4.37) Kono honwo yominagara jinseinitsuite kangaeta.

[pro]$_{f1}$	*[kono]*$_{f2}$	*[hon-wo]*$_{f3}$	*[yom-i-nagara]*$_{f4}$	*[pro]*$_{f5}$
pro-SUBJ	this	book-OBJ	read-conn.base-while	life-about
[jinsei-nitsuite]$_{f6}$	*[kangae-ta]*$_{f7}$			
think-decl.ta				

"While reading this book, I thought about life."

The DAG representation and f-structure for (4.35) are Figure 4.70 and Figure 4.71, respectively:

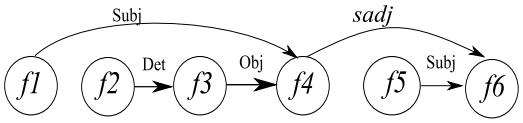

Figure 4.70: DAG for (4.35)

$$
\begin{bmatrix}
\text{SADJ} & \left\{ \begin{bmatrix}
 & \text{SUBJ} & \underset{f1}{\begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix}} \\
 & \text{OBJ} & \underset{f3}{\begin{bmatrix} \text{DET} & \underset{f2}{\begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix}} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix}} \\
 & \text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
 & \text{V-INFL} & \text{`-te'}
\end{bmatrix} \right\} \\
\text{SUBJ} & \underset{f5}{\begin{bmatrix} \text{PRED} & \text{`jinsei; life'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix}} \\
\text{PRED} & \text{`kawar-}\langle \text{SUBJ} \rangle\text{; change'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
\underset{f6}{\text{MOOD}} & \text{`declarative'}
\end{bmatrix}
$$

Figure 4.71: F-structure for (4.35)

4.2.12 COORD

Coordination is expressed by a case particle "-to" or "-ya", or a punctuation mark. Coordination is an instance where the DAG dependency among syntactic units does not yield the correct f-structure representation. Consider (4.38):

(4.38) Watashiwa kono honto ano honwo yonda.

[watashi-wa]$_{f1}$	*[pro]*$_{f2}$	*[kono]*$_{f3}$	*[hon-to]*$_{f4}$	*[ano]*$_{f5}$	*[hon-wo]*$_{f6}$
I-TOP	pro-SUBJ	[this	book-and	that	book]-OBJ
[yom-ta]$_{f7}$					
read-decl.ta					

"I read this book and that book."

The dependency DAG of this sentence is below:

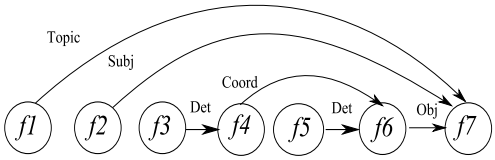

Figure 4.72: DAG for (4.38)

Notice that (4.72) does not properly represent the coordination of $f4$ and $f6$, in that in the DAG $f4$ depends on $f6$, hence in the f-structure (Figure 4.73) directly projected from the DAG in Figure 4.72, the f-structure for the first conjunct ("kono hon-to" in this example) is embedded into the f-structure for the second conjunct ("ano hon-wo" in this example):

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f1} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f2} \\
\text{OBJ} & \begin{bmatrix} \text{COORD} & \left\{ \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`kono'} \end{bmatrix}_{f3} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-to'} \end{bmatrix}_{f4} \right\} \\ \text{DET} & \begin{bmatrix} \text{PRED} & \text{`ano'} \end{bmatrix}_{f5} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix}_{f6} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}_{f7}
$$

Figure 4.73: F-structure from Figure 4.72

Coordinates must not be embedded, but they should be treated as elements of a set which has one grammatical function. Therefore, the dependency relationship must be modified as follows; the coordinates depend on one "dummy" syntactic unit which has the grammatical function of the last coordinate in the DAG in Figure 4.72. The DAG representation for the example sentence (4.38) is then (4.74):

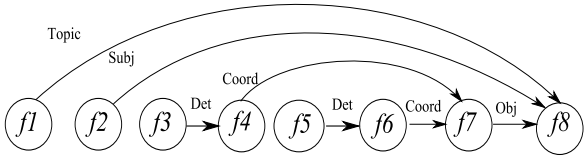

Figure 4.74: DAG for (4.38)

The syntactic unit $f7$ is the dummy unit on which the coordinates depend. This unit contains the case particle "-wo", hence this has the gram-

matical function OBJ. The f-structure projected from (4.74) is (4.75). The coordinates are analysed as elements of the set in the f-structure $f7$:

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f1} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f2} \\
\text{OBJ} & \begin{bmatrix} \text{COORD} & \left\{ \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'kono'} \end{bmatrix}_{f3} \\ \text{PRED} & \text{'hon; book'} \end{bmatrix}_{f4}, \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'ano'} \end{bmatrix}_{f5} \\ \text{PRED} & \text{'hon; book'} \end{bmatrix}_{f6} \right\} \\ \text{PRTCS} & \text{'-TO'} \end{bmatrix} \\
 & \quad \begin{bmatrix} \text{PRTCS} & \text{'-wo'} \end{bmatrix}_{f7} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{V-INFL} & \text{'-ta'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}_{f8}
$$

Figure 4.75: F-structure for (4.38)

Later in this thesis, Section 6.4 deals with how the KTC4/KNP representation for coordination is properly treated in the LFG annotation process using functional equations.

Coordination by "-to" indicates that the speaker has enumerated everything that are the case, while coordination by "-ya" indicates that the speaker has picked up the examples that are the case:

(4.39) Watashiwa gengogakuya jinruigakuwo kenkyuushita.

watashi-wa	*gengogaku-ya*	*jinruigaku-wo*	*kenkyuu-shi-ta*
I-TOP	linguistics-and	anthropology-OBJ	study-do-decl.ta

"I studied linguistics, anthropology, and others."

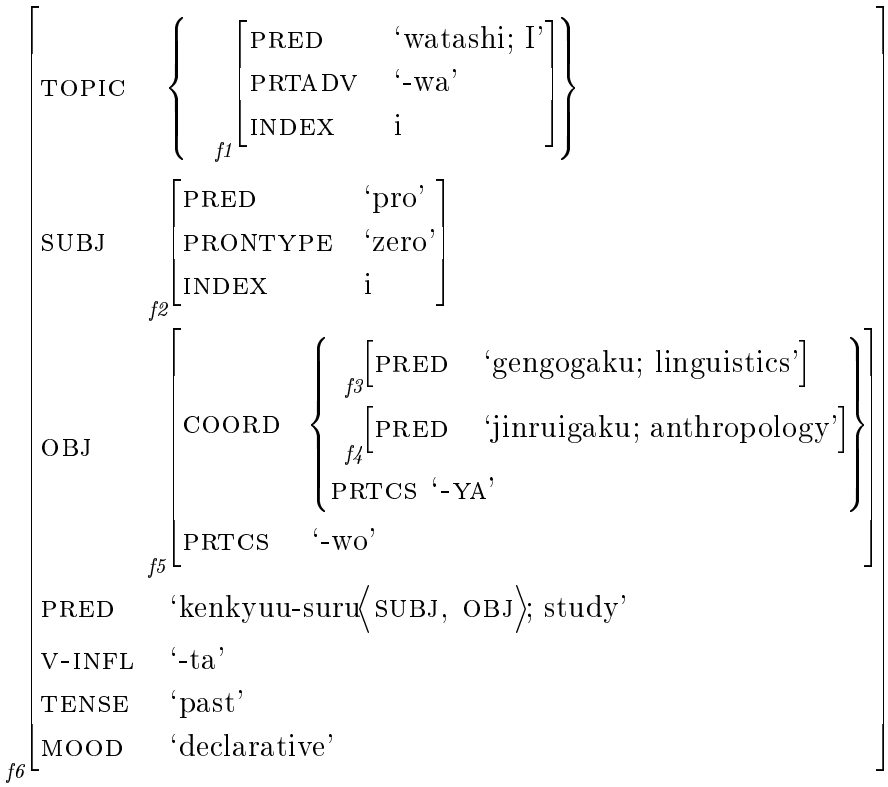

Figure 4.76: F-structure for (4.39)

4.3 Grammatical Features

This section introduces the grammatical features used in the Japanese f-structure representation in this thesis. These features are mainly derived from information provided by each of the morphemes in a syntactic unit, and these pieces of information are combined to form the f-structure for the syntactic unit. However, we do not always find a simple one-to-one correspondence between a morpheme and an atomic grammatical feature; there are many instances where, even though the value of a particular grammatical feature is atomic, this value is represented by more than one morpheme, i.e. resulting in many-to-one correspondences between morphemes and a grammatical feature. In addition, there are idiomatic constructions in which the interpretation of a unit is not the sum of the interpretations of its parts. Furthermore, syntactic information, such as adverbial units modifying a verbal unit, must be taken into account, especially identifying the modal feature of a given sentence. This section presents the grammatical features in the Japanese f-structures developed in this thesis, and approaches the issues at stake in terms of the correspondence between the form and function. The grammatical features dealt with in this section are: TENSE, ASPECT, MOOD, and VOICE.

Before starting the presentation of the grammatical features, it is necessary to make some comments on how inflectional features are presented in the f-structures of this study. As far as the correspondence between the form and function is concerned, the forms of inflecting morphemes in syntactic units, especially those in verbal units, play important roles in establishing the correspondence. In this study, all the inflectional forms are considered to be the value of an attribute of INFL, and INFL is subdivided into part-of-speech categories, such as follows:

1. V-infl (verbal inflection)
2. A-infl (adjectival inflection)
3. AUX-infl (auxiliary inflection)
4. SUF-infl (suffix inflection)

The projected features at the level of f-structures take the surface form as their values. If the surface form of a morpheme is the same as its lemma form, then the INFL value is zero, represented as "ϕ". If there are more than one morpheme of the same category in one syntactic unit, then they

are represented in the linear order of their appearance. It is postulated that there is no hierarchical relationship among them.

Along with these inflection features, the root of each inflecting form is the value of the feature V, AUX, SUF. In the f-structures of this study, the surface forms of all inflections are represented as the values of relevant grammatical features. This is because it will be possible to take the patterns of inflection forms in f-structure representations, and utilise these patters for constructing language models or other purposes, for example, natural language generation which takes f-structures as input, which is an issue of further study in future.

4.3.1 Tense

The grammatical category Tense is a feature which locates the state or the action expressed by a sentence onto a timescale. In general, there are three values for Tense: past, present and future. How these values are expressed varies from language to language. In English, the present and the past tense are expressed by verbal morphology, and the future tense is expressed by the auxiliary verbs "will" and "shall". In Japanese, verbal morphology specifies the tense of a sentence, but the story does not end here.

For tense, the inflection forms of Japanese verbs are divided into two classes (base forms and ta forms). Ta forms express the past tense, while base forms never express the past tense (cf. Section 3.5.1). These facts lead us to name the f-structure grammatical feature corresponding to ta forms "past", while the grammatical feature of base forms is "nonpast". The term "nonpast" includes the future tense of both stative and action verbs, the present and present perfect tenses of stative verbs, and the generic tense. The tense feature of a given sentence can be disambiguated by lexical semantics of the verb (stative or action), adverbial modification, and contextual consideration by the listener.

Examples of Stative-Verb Tense Specification

Stative verbs in their declarative base form project the tense feature "nonpast". For example, in sentence (4.40), a stative verb 'iru' is in the declarative base form, and the tense feature in the f-structure for this sentence is nonpast:

(4.40) Watashiwa ieni iru.

watashi-wa	*pro*	*ie-ni*	*ir-u*
I-TOP	pro-SUBJ	house-OBL	exist-decl.base

"I am in a house."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{`ie; house'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{PRED} & \text{`ir-}\langle \text{SUBJ, OBL} \rangle\text{; exist'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.77: F-structure for (4.40)

In (4.41) the adverbial phrase "kono issuukan" projects the tense feature 'present' and the aspect 'perfect' (Figure 4.78): these features are not derived from any morphological inflections, but specified by the lexical semantics of the adverbial phrase:

(4.41) Kono isshukan watashiwa ieni iru

kono	*issukan*	*watashi-wa*	*ie-ni*	*ir-u*
this	one.week	I-TOP	house-OBL	exist-decl.base

"I have been home for this one week."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{ADJ} & \left\{ \begin{bmatrix} \text{DET} & \left[\text{PRED} \quad \text{`kono; this'}\right] \\ \text{PRED} & \text{`isshukan; one week'} \end{bmatrix} \right\} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{`ie; house'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{PRED} & \text{`ir-}\langle \text{SUBJ, OBL} \rangle\text{; exist'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`present'} \\
\text{ASPECT} & \text{`perfect'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.78: F-structure for (4.41)

By contrast, the adverbial unit "ashitamo" projects the tense feature as 'future' (Figure 4.79):

(4.42) Watashiwa ashitamo ieni iru

watashi-wa	*ashita-mo*	*ie-ni*	*ir-u*
I-TOP	tomorrow-FOC	house-OBL	exist-decl.base

"I will be home tomorrow, too."

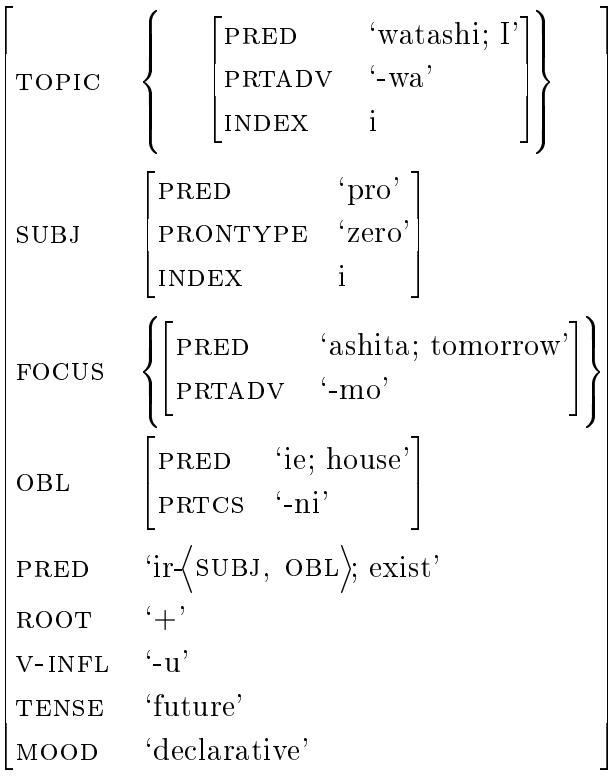

Figure 4.79: F-structure for (4.42)

Examples of Action-Verb Tense Specification

An action verb 'yom-' in the declarative base form projects the tense feature 'nonpast' (Figure 4.80):

(4.43) Watashiwa kono honwo yomu.

watashi-wa	*kono*	*hon-wo*	*yom-u*
I-TOP	this	book-OBJ	read-decl.base

"I read this book."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \left[\text{PRED} \quad \text{`kono; this'}\right] \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.80: F-structure for (4.43)

The tense feature 'nonpast' can be further specified into 'future' by adverbial expressions (Figure 4.81, Figure 4.82, and Figure 4.83):

(4.44) Watashiwa imakara kono honwo yomu.

watashi-wa	*ima-kara*	*kono*	*hon-wo*	*yom-u*
I-TOP	now-from	this	book-OBJ	read-decl.base

"I read this book from now."

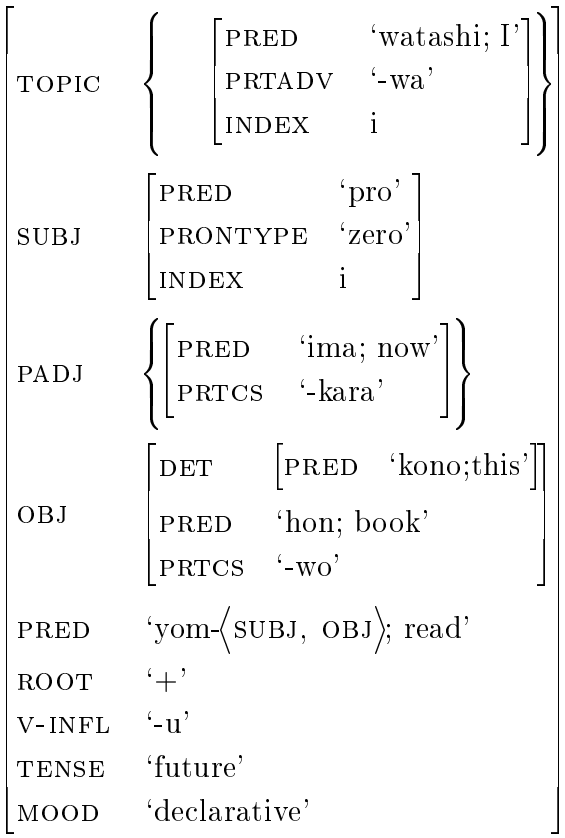

Figure 4.81: F-structure for (4.44)

(4.45) Watashiwa imakara kono honwo yomutsumorida.

watashi-wa	*ima-kara*	*kono*	*hon-wo*
I-TOP	now-from	this	book-OBJ

yom-u-tsumorida
read-decl.base-aux.decl.base
"I will read this book from now."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{`ima; now'} \\ \text{PRTCS} & \text{`-kara'} \end{bmatrix} \right\} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{AUX} & \text{`-tsumorida'} \\
\text{AUX-INFL} & \text{`zero'} \\
\text{TENSE} & \text{`future'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.82: F-structure for (4.45)

(4.46) Watashiwa ashita kono honwo yomutsumorida.

watashi-wa	*ashita*	*kono*	*hon-wo*
I-TOP	tomorrow	this	book-OBJ

yom-u-tsumorida
read-decl.base-aux.decl.base
"I will read this book tomorrow."

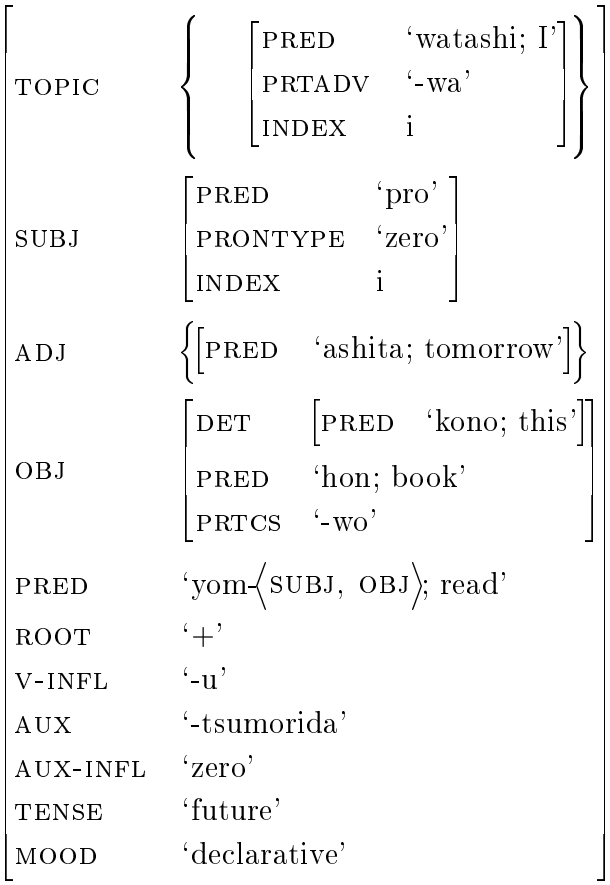

Figure 4.83: F-structure for (4.46)

4.3.2 Aspect

In general, an action is always changing: it starts, continues, sometimes halts, and then finishes. And the result of an action also is important; an action may or may not bring about a change of state on someone or something. The grammatical feature Aspect focuses on these changes or results of an action. There are several values for this feature, and different languages have many different ways of expressing each of these. In English, the progressive and the perfective aspects are syntactically expressed. In Japanese, verbal suffixes are in charge of expressing aspectual meanings, and adverbial modification also plays an important role in specifying aspectual meaning.

"-teiru" Aspect

One of the most frequently used aspectual expressions in Japanese is a verb in its connective ta form plus a verbal suffix "-iru", for example, "yom-te-iru". This expression covers various aspects. The examples below exhibit progressive, perfective, resultative and iterative aspects:

Sentence (4.47) is an example of progressive "-teiru". There is no adverbial modification in this sentence, and the aspectual feature is progressive (Figure 4.84):

(4.47) Watashiwa kono honwo yondeiru.

watashi-wa	*kono*	*hon-wo*	*yom-te-ir-u*
I-TOP	this	book-OBJ	read-conn.ta-suff-decl.base

"I am reading this book."

$$\begin{bmatrix} \text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\ \text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{ROOT} & \text{'+'} \\ \text{V-INFL} & \text{'-te-'} \\ \text{SUF} & \text{'-ir-'} \\ \text{SUF-INFL} & \text{'-u'} \\ \text{TENSE} & \text{'present'} \\ \text{ASPECT} & \text{'progressive'} \\ \text{MOOD} & \text{'declarative'} \end{bmatrix}$$

Figure 4.84: F-structure for (4.47)

Sentence (4.48) is an example of perfective "-teiru". This sentence is modified by an adverbial phrase "senshuu-kara" (since the last week), and the aspectual feature is perfective (Figure 4.85):

(4.48) Watashiwa senshuukara kono honwo yondeiru

watashi-wa *senshuu-kara* *kono* *hon-wo*
I-TOP last.week-from this book-OBJ
yom-te-ir-u
read-conn.ta-suff-decl.base
"I have been reading this book since the last week."

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'senshuu; last week'} \\ \text{PRTCS} & \text{'-kara'} \end{bmatrix} \right\} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \left[\text{PRED} \quad \text{'kono; this'} \right] \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-te-'} \\
\text{SUF} & \text{'-ir-'} \\
\text{SUF-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'present'} \\
\text{ASPECT} & \text{'perfective'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}
$$

Figure 4.85: F-structure for (4.48)

Sentence (4.49) is an example of resultative "-teiru". This sentence is modified by an adverbial phrase "sudeni" (already), and the aspectual feature is resultative (Figure 4.85):

(4.49) Watashiwa sudeni kono honwo yondeiru.

watashi-wa	*sudeni*	*kono*	*hon-wo*	*yom-te-ir-u*
I-TOP	already	this	book-OBJ	read-conn.ta-suff-decl.base

"I have already read this book."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{ADJ} & \left\{ \begin{bmatrix} \text{PRED} & \text{'sudeni; already'} \end{bmatrix} \right\} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{ROOT} & \text{'+'} \\
\text{V-INFL} & \text{'-te-'} \\
\text{SUF} & \text{'-ir-'} \\
\text{SUF-INFL} & \text{'-u'} \\
\text{TENSE} & \text{'present'} \\
\text{ASPECT} & \text{'resultative'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}$$

Figure 4.86: F-structure for (4.49)

What seems to be common among these various aspectual meanings expressed by the V-teiru construction is that something concerning the action is continued. This "something" can be the action itself, or the result of the

action. In the progressive, the action is continued by the referent of the subject. In the perfective, the action has been continued until the time of the utterance. In the resultative, the result of the action in the past has continued to be valid until the present.

Considering the fact that these various aspectual meanings are expressed by one construction, and that they seem to share a certain aspectual meaning best expressed as "continuing", these aspectual meanings can be subsumed under the category "-teiru aspect." In actual language use, this aspectual category can be further specified into subcategories such as the progressive or the resultative, mainly by adverbial modification.

4.3.3 Mood

Mood is the grammatical category which is concerned with the speaker's subjective judgement on a proposition. The values of the grammatical feature Mood include "declarative", "imperative", "subjunctive", "negative", etc. Different moods are expressed in a variety of different ways in different languages, and in English the modal auxiliaries such as "may" or "must" can express mood in a sentence.

(4.50)
Ken must read this book.
Ken may read this book.

A variety of other modal auxiliaries express various modal features in a sentence, and one of the most important axes of modality is the possibility/necessity axis, which is the core of modal logic.

Along with the modal auxiliaries, different adverbs can disambiguate the modality in terms of deontic/epistemic distinctions ([Palmer, 2001]). Deontic mood is concerned with duty or obligation, while epistemic mood is concerned with perception or recognition:

(4.51)
It is possible that Ken reads this book. (epistemic possibility)
It is necessarily the case that Ken reads this book. (epistemic necessity)
It is possible for Ken to read this book. (deontic possibility)
It is necessary for Ken to read this book. (deontic necessity)

In Japanese, the mood feature of a sentence is specified by verbal inflections, verbal suffixes, auxiliaries, adverbial modifications, or a number of idiomatic expressions for modality. This section highlights some typical instances of

modal expressions, and introduces how the corresponding mood features are represented in f-structure. First, I discuss inflectional moods, which are expressed by verbal inflections of a verb (cf. Section 3.5.1). Second, I focus on moods expressed by more than one morpheme, which can be called "non-inflectional moods" in order to make the contrast with moods expressed by verbal inflections.

Inflectional Moods

Inflectional moods are expressed by verbal inflections of a verb, and there is a one-to-one correspondence between the name of the verbal inflection and the modal value. Sentence (4.52) exhibits volitional mood, and the f-structure for this sentence is Figure 4.87. The verb is in the volitional form (cf. Table 3.1 in Section 3.5.1):

(4.52) Watashiwa kono honwo yomo.

watashi-wa	*kono*	*hon-wo*	*yom-o*
I-TOP	this	book-OBJ	read-vol.base

"I will read this book."

$$\begin{bmatrix} \text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{DET} & \left[\text{PRED} \quad \text{'kono; this'}\right] \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\ \text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{ROOT} & \text{'+'} \\ \text{V-INFL} & \text{'-o'} \\ \text{MOOD} & \text{'volitional'} \end{bmatrix}$$

Figure 4.87: F-structure for (4.52)

Sentence (4.53) exhibits imperative mood, and the f-structure for this sentence is Figure 4.88.

(4.53) Kono honwo yome.

kono *hon-wo* *yom-e*
this book-OBJ read-imp.base
"Read this book."

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\ \text{PRED} & \text{'yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\ \text{ROOT} & \text{'+'} \\ \text{V-INFL} & \text{'-e'} \\ \text{MOOD} & \text{'imperative'} \end{bmatrix}$$

Figure 4.88: F-structure for (4.53)

The subordinate clause in Sentence (4.54) exhibits conditional mood, and the f-structure for this sentence is Figure 4.89:

(4.54) Kono honwo yomeba, kimino jinseiwa kawaru.

kono *hon-wo* *yom-eba,* *kimi-no* *jinsei-wa*
this book-obj read-cond.base you-PADJ life-SUBJ
kawar-u.
change-decl.base
"Your life will change if you read this book.

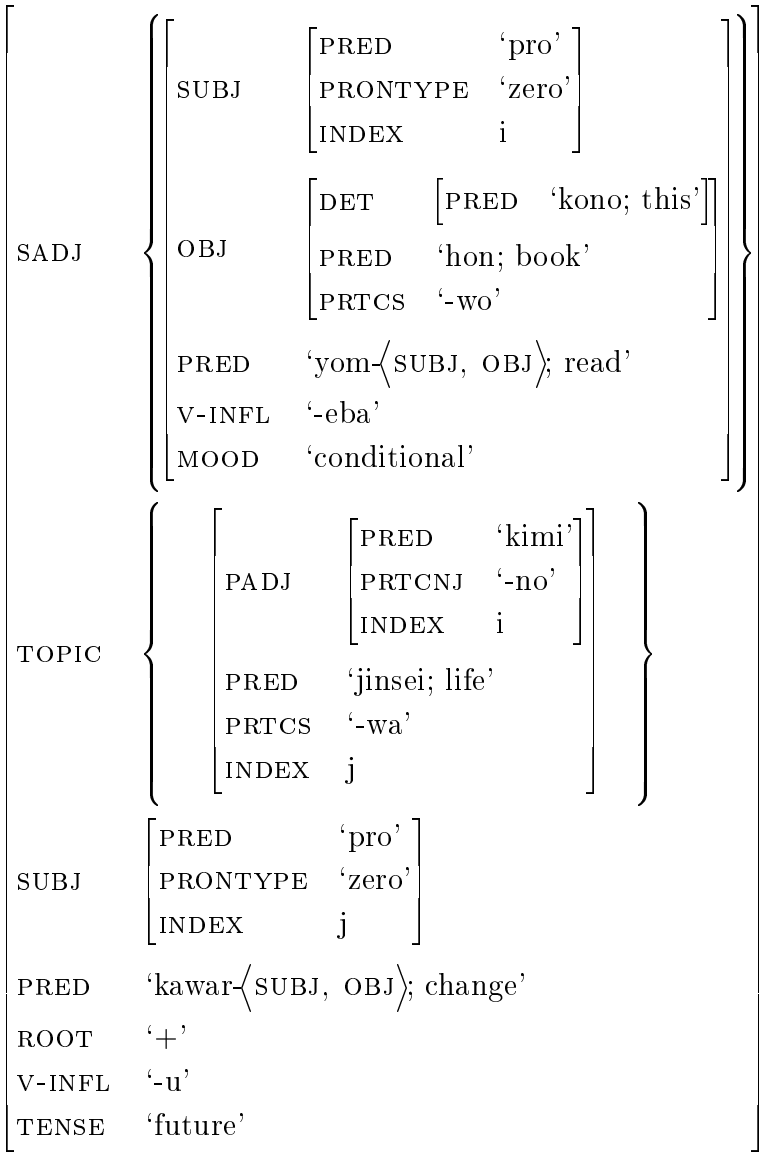

Figure 4.89: F-structure for (4.54)

The subordinate clause in Sentence (4.55) exhibits connective mood, and the f-structure for this sentence is Figure 4.90:

(4.55) Kono honwo yonde jibunno ikikatawo kangaenaoshita.

kono	*hon-wo*	*yom-te*	*jibunno*	*ikikata-wo*
this	book-OBJ	read-conn.ta	onself-of	way.of.life-OBJ
kangaenaos-ta				
reconsider-decl.ta				

"Reading this book, I reflected on my way of life."

$$\begin{bmatrix}
\text{SADJ} & \left\{ \begin{bmatrix}
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \left[\text{PRED} \quad \text{`kono; this'}\right] \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`yom-}\langle \text{SUBJ, OBJ} \rangle\text{; read'} \\
\text{V-INFL} & \text{`-te'} \\
\text{MOOD} & \text{`connective'}
\end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}
\text{PADJ} & \begin{bmatrix} \text{PRED} & \text{`jibun'} \\ \text{PRTCNJ} & \text{`-no'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{PRED} & \text{`ikikata; way of life'} \\
\text{PRTCS} & \text{`-wo'}
\end{bmatrix} \\
\text{PRED} & \text{`kangaenaos-}\langle \text{SUBJ, OBJ} \rangle\text{; reflect on'} \\
\text{V-INFL} & \text{`-ta'} \\
\text{ROOT} & \text{`+'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.90: F-structure for (4.55)

Morphological Moods

Japanese has a variety of non-inflectional ways to express mood, and explaining how each of them is used would require another volume. Here I concentrate on the two major axes of modality: possible/necessary and epistemic/deontic.

Table 4.1 shows one example modal expression for one verb unit for each of the modal categories. As the table shows, the epistemic/deontic and possible/necessary distinctions can be made by verbal morphology:

Table 4.1: Morphological Moods for "kak-" (write)

	Possible	Necessary
Epistemic	kak-ukamoshirenai	kak-unichigainai
Deontic	kak-temoii	kak-anebanaranai

Sentence (4.56) exhibits epistemic possible mood:

(4.56) Naomiwa shousetsuwo kakukamoshirenai.

Naomi-wa *shousetsu-wo*
Naomi-TOP novel-OBJ
kak-u-kamo-shire-na-i
write-decl.base-prtcnj-know-NEG-decl.base
(Lit.)"It cannot be known whether Naomi writes a novel/novels."
"It is epistemically possible that Naomi writes a novel/novels."
"Naomi may write a novel/novels."

The f-structure representation for (4.56) is Figure 4.91:

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`Naomi'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`shosetsu; novel'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`kak-}\langle \text{SUBJ, OBJ} \rangle\text{; write'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{SUF} & \text{`-kamo-shire-na-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`epistemic-possible'}
\end{bmatrix}$$

Figure 4.91: F-structure for (4.56)

Sentence (4.57) exhibits epistemic necessary mood; the expression "chigainai" (lit. no difference) sometimes means "right":

(4.57) Naomiwa shousetsuwo kakunichigainai

Naomi-wa *shousetsu-wo*
Naomi-TOP novel-OBJ
kak-u-ni-chigai-na-i
write-decl.base-prtcnj-different-NEG-decl.base
(Lit.)"There is no difference that Naomi writes a novel."
"It is epistemically necessarily the case that Naomi will write a novel."
"It is certain that Naomi will write a novel/novels."

The f-structure representation for (4.57) is Figure 4.92:

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`Naomi'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`shosetsu; novel'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`kak-}\langle \text{SUBJ, OBJ} \rangle\text{; write'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-u'} \\
\text{SUF} & \text{`-ni-chigai-na-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`epistemic-necessary'}
\end{bmatrix}$$

Figure 4.92: F-structure for (4.57)

Sentence (4.58) exhibits deontic possible mood:

(4.58) Naomiwa shousetsuwo kaitemoii.

Naomi-wa *shousetsu-wo* *kak-te-mo-i-i.*
Naomi-TOP novel-OBJ write-conn.ta-prtcnj-good-decl.base
(Lit.) "It is also good for Naomi to write a novel/novels."
"It is deontically possible for Naomi to write a novel/novels."
"Naomi is allowed to write a novel/novels."

Figure 4.93 is the f-structure representation for (4.58):

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`Naomi'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`shosetsu; novel'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`kak-}\langle \text{SUBJ, OBJ} \rangle\text{; write'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-te'} \\
\text{SUF} & \text{`-mo-i-'} \\
\text{SUF-INFL} & \text{`-i'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`deontic-possible'}
\end{bmatrix}$$

Figure 4.93: F-structure for (4.58)

Sentence (4.59) exhibits deontic necessary mood. Literally, this expression contains double negation:

(4.59) Naomiwa shousetsuwo kakanebanaranai.

Naomi-wa *shousetsu-wo*
Naomi-TOP novel-OBJ
kak-an-eba-nar-ana-i
write-NEG-cond.base-become-NEG-decl.base
(Lit.)"Things doesn't become good if Naomi doesn't write a novel/novels."
"It is deontically necessary for Naomi to write a novel/novels."
"Naomi must write a novel."

Figure 4.94 is the f-structure representation for (4.59):

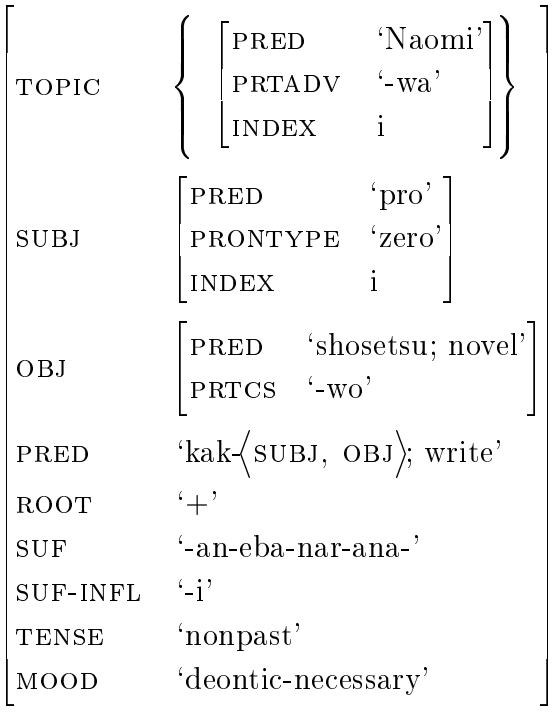

Figure 4.94: F-structure for (4.59)

One modal expression corresponds to one modal value, but not vice versa. For each modal category, there are different expressions with different degree of possibility or necessity. For example, deontic necessity can also be expressed by "kakubekida" or "kakuhougaii", the latter being less demanding. In addition, these modal expression can be combined to express complex modality, such as "verb root + nebanaranai + kamoshirenai (it is epistemically possible that it is deontically necessary)", "verb root + temoii + nichigainai (it is epistemically necessary that it is deontically possible)".

4.3.4 Voice

Causative

Causative voice is expressed by the causative suffixes "-aser-" and "-saser-". Type-I verbs are followed by "-aseru", Type-II verbs are followed by "-saseru". I assume that a verb with a causative suffix is always one word (cf. Section 3.5.4). In sentence (4.60), the root verb "yom-" and the causative suffix "-aser-" constitute one causative verb, and the f-structure for this sentence (Figure 4.95) is monoclausal:

(4.60) Watashiwa ototoni kono honwo yomaseta.

watashi-wa	*ototo-ni*	*kono*	*hon-wo*
I-TOP	younger.brother-OBL	this	book-OBJ

yom-aser-ta
read-caus.-decl.ta
"I made my younger brother read this book."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{'ototo; younger brother'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{'kono; this'} \end{bmatrix} \\ \text{PRED} & \text{'hon; book'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'yom-aser}\langle \text{SUBJ, OBJ, OBL} \rangle\text{; make sb. read'} \\
\text{ROOT} & \text{'+'} \\
\text{SUF} & \text{'-aser-'} \\
\text{SUF-INFL} & \text{'-ta'} \\
\text{VOICE} & \text{'causative'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}$$

Figure 4.95: F-structure for (4.60)

Direct Passive

Passive voice is expressed by the passive suffixes "-areru" or "-rareru". Both follow the root of a verb, and they inflect as Type-II verbs (cf. Section 3.5.4). In direct passive, the object of a verb in the active voice turned into the subject, and the subject of the verb in the active voice is expressed by the oblique-case noun phrase. Only transitive verbs can be in the direct

passive. In sentence (4.61), the root verb "kowas-" and the passive suffix "-are-" constitute a direct passive verb, and Figure 4.96 is the f-structure for it:

(4.61) Watashino tsuboga ototoni kowasareta.

watashi-no	*tsubo-ga*	*ototo-ni*	*kowas-are-ta*
I-of	vase-SUBJ	younger.brother-OBL	break-pass.-decl.ta

"My vase was broken by my younger brother."

$$\begin{bmatrix} \text{SUBJ} & \begin{bmatrix} \text{PADJ} & \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTCNJ} & \text{'-no'} \end{bmatrix} \\ \text{PRED} & \text{'tsubo'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\ \text{OBL} & \begin{bmatrix} \text{PRED} & \text{'ototo; younger brother'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\ \text{PRED} & \text{'kowas-arer}\langle \text{SUBJ, OBL} \rangle\text{;be broken'} \\ \text{ROOT} & \text{'+'} \\ \text{SUF} & \text{'-arer-'} \\ \text{SUF-INFL} & \text{'-ta'} \\ \text{VOICE} & \text{'passive'} \\ \text{TENSE} & \text{'past'} \\ \text{MOOD} & \text{'declarative'} \end{bmatrix}$$

Figure 4.96: F-structure for (4.61)

Indirect Passive from a Transitive Verb

Indirect passive expresses that the referent of the subject is influenced by, or suffered from the action or the result of the action caused by the referent of the oblique. Sentence (4.62) is an example of indirect passive, and the f-structure representation is Figure 4.97:

(4.62) Watashiwa ototoni tsubowo kowasareta.

watashi-wa	*ototo-ni*	*tsubo-wo*	*kowas-arer-ta*
I-TOP	younger.brother-OBL	vase-OBJ	break-pass.-decl.ta

"I suffer from my younger brother's breaking my vase." or "I had my vase broken by my younger brother."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{'ototo; younger brother'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'tsubo; vase'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{PRED} & \text{'kowas-arer}\langle \text{SUBJ, OBJ, OBL} \rangle\text{; suffer from someone's breaking'} \\
\text{ROOT} & \text{'+'} \\
\text{SUF} & \text{'-arer-'} \\
\text{SUF-INFL} & \text{'-ta'} \\
\text{VOICE} & \text{'passive'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}$$

Figure 4.97: F-structure for (4.62)

Indirect Passive from an Intransitive Verb

Unlike direct passive, intransitive verbs can be in the indirect passive. Sentence (4.63) is an example of indirect passive from an intransitive verb, and its f-structure representation is Figure 4.98:

(4.63) Watashiwa imotoni nakareta.

watashi-wa	*imoto-ni*	*nak-arer-ta*
I-TOP	younger.sister-OBL	weep-pass.-decl.ta

"I suffer from my sister's weeping."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{`imoto; younger sister'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{PRED} & \text{`nak-arer}\langle \text{SUBJ, OBL} \rangle\text{; suffer from someone's weeping'} \\
\text{ROOT} & \text{`+'} \\
\text{SUF} & \text{`-arer-'} \\
\text{SUF-INFL} & \text{`-ta'} \\
\text{VOICE} & \text{`passive'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.98: F-structure for (4.63)

Causative-passive

A causative verb can be followed by the passive suffix to constitute a verb which projects "causative-passive" voice. Sentence (4.64) is an example of causative-passive voice, and the f-structure representation of it is Figure 4.99:

(4.64) Aneni kono honwo yomasareta.

Ane-ni	*kono*	*hon-wo*	*yom-as-arer-ta.*
elder.sister-OBL	this	book-OBJ	read-caus-pass-decl.ta

"Someone$_i$ was made to read this book by his(or her)$_i$ sister."

$$\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`pro'} \\ \text{PRONTYPE} & \text{`zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{PRED} & \text{`ane; elder sister'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{PRED} & \text{`yom-as-arer}\langle \text{SUBJ, OBJ, OBL} \rangle\text{; be made to read'} \\
\text{ROOT} & \text{`+'} \\
\text{SUF} & \text{`-as-arer-'} \\
\text{SUF-INFL} & \text{`-ta'} \\
\text{VOICE} & \text{`causative-passive'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.99: F-structure for (4.64)

Benefactive

Benefactives express the speaker's subjective judgement on the benefit of the action (cf. Section 3.5.4). The suffix "-kureru" emphasises the speaker's gratitude that the action by the subject is beneficial to the speaker. Sentence (4.65) is an example of "-kureru" benefactive, and Figure 4.100 is the f-structure representation for it. The gloss for 'hanas-te-kurer<subj,obl,obj>' in Figure 4.100 is 'give someone the favor of telling':

(4.65) Anega watashini shinjitsuwo hanashitekureta.

ane-ga	*watashi-ni*	*shinjitsu-wo*
elder.sister-SUBJ	I-OBL	shinjitsu-OBJ

hanas-te-kurer-ta
tell-conn.ta-give-decl.ta
"My elder sister gave me the favour of telling me the truth."

$$\begin{bmatrix}
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`ane; elder sister'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTCS} & \text{`-ni'} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{`shinjitsu; truth'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`hanas-te-kurer}\langle \text{SUBJ,OBL,OBJ} \rangle\text{;'} \\
\text{ROOT} & \text{`+'} \\
\text{V-INFL} & \text{`-te'} \\
\text{SUF} & \text{`kurer-'} \\
\text{SUF-INFL} & \text{`-ta'} \\
\text{VOICE} & \text{`passive'} \\
\text{TENSE} & \text{`past'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.100: F-structure for (4.65)

The suffix "-morau", on the other hand, emphasises the speaker's gratitude that the action by the referent of the oblique is beneficial to the speaker. Sentence (4.66) is an example of the "-morau" benefacitve, and Figure 4.101 is the f-structure representation for the sentence. The gloss for 'hanas-te-moraw<subj, obl, obj>' in Figure 4.101 is 'receive the benefit of being told from someone':

(4.66) Watashiwa aneni shinjitsuwo hanashitemoratta.

watashi-top	*ane-ni*	*shinjitsu-wo*
I-TOP	elder.sister-OBL	truth-OBJ

hanas-te-moraw-ta
talk-conn.ta-receive-decl.ta
"I received from my sister the benefit of being told the truth".

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{'watashi; I'} \\ \text{PRTADV} & \text{'-wa'} \\ \text{INDEX} & \text{i} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{PRONTYPE} & \text{'zero'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\
\text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'shinjitsu; truth'} \\ \text{PRTCS} & \text{'-wo'} \end{bmatrix} \\
\text{OBL} & \begin{bmatrix} \text{PRED} & \text{'ane; elder sister'} \\ \text{PRTCS} & \text{'-ni'} \end{bmatrix} \\
\text{PRED} & \text{'hanas-te-moraw}\langle \text{SUBJ, OBL, OBJ} \rangle\text{;'} \\
\text{ROOT} & \text{'+'} \\
\text{SUF} & \text{'-moraw-'} \\
\text{SUF-INFL} & \text{'-ta'} \\
\text{TENSE} & \text{'past'} \\
\text{MOOD} & \text{'declarative'}
\end{bmatrix}$$

Figure 4.101: F-structure for (4.66)

Valency-changing Voice

"Possibility" or "spontaneous" suffixes turn a transitive verb into an intransitive verb, and the object of the transitive verb becomes its subject, and the agent of the verb's action is either not expressed or topicalised:

Sentence (4.67) exhibits spontaneous voice, and the f-structure representation for the sentence is Figure 4.102:

(4.67) Kokokara fujisanga mieru.

koko-kara	*fujisan-ga*	*mi-er-u*
here-prtcs	Mt.Fuji-SUBJ	see-suff.-decl.base

"From here, Mt.Fuji is visible." or "We can see Mt.Fuji from here."

$$\begin{bmatrix}
\text{PADJ} & \begin{bmatrix} \text{PRED} & \text{`koko'} \\ \text{PRTCS} & \text{`-kara'} \end{bmatrix} \\
\text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{`fujisan; Mt.Fuji'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\
\text{PRED} & \text{`mi-er}\langle \text{SUBJ} \rangle\text{; be visible'} \\
\text{ROOT} & \text{`+'} \\
\text{SUF} & \text{`-er-'} \\
\text{SUF-INFL} & \text{`-u'} \\
\text{VOICE} & \text{`spontaneous'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{declarative}
\end{bmatrix}$$

Figure 4.102: F-structure for (4.67)

Sentence (4.68) exhibits possible voice and the f-structure representation for the sentence is Figure 4.103:

(4.68) Watashiwa kono honga yomeru.

watashi-wa	*kono*	*hon-ga*	*yom-er-u*
I-TOP	this	book-SUBJ	read-suff.-decl.base

"As for me, this book is readable." or "I can read this book."

$$\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{PRED} & \text{`watashi; I'} \\ \text{PRTADV} & \text{`-wa'} \end{bmatrix} \right\} \\
\text{SUBJ} & \begin{bmatrix} \text{DET} & \begin{bmatrix} \text{PRED} & \text{`kono; this'} \end{bmatrix} \\ \text{PRED} & \text{`hon; book'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\
\text{PRED} & \text{`yom-er}\langle \text{SUBJ} \rangle\text{; be readable'} \\
\text{ROOT} & \text{`+'} \\
\text{SUF} & \text{`-er-'} \\
\text{SUF-INFL} & \text{`-u'} \\
\text{VOICE} & \text{`possible'} \\
\text{TENSE} & \text{`nonpast'} \\
\text{MOOD} & \text{`declarative'}
\end{bmatrix}$$

Figure 4.103: F-structure for (4.68)

Ideally, it is better to be able to make a morpho-syntactic distinction between the spontaneous voice and the possible voice; however, since the distinction is of a semantic nature, morpho-syntactic information in a corpus or parser output is not enough to automatically and unequivocally determine the voice of a sentence in which the possible/spontaneous "-eru" is used.

4.4 Summary

This chapter described core aspects of Japanese grammar based on and encoded in the framework of LFG. In particular, I looked at how core grammatical functions and grammatical features are encoded in Japanese and in the corresponding LFG representations. Japanese sentences are divided into "bunsetsu", or *syntactic units*, and each of the dependencies among these units can be labelled by a unique grammatical function which is specified by the particle at the right periphery of the bunsetsu, or by the part of speech of the head of the unit. It is pointed out that the problem of ambiguity for grammatical function assignment, such as the distinction between REL and COMP, requires more data and study for their resolution. DAG representations of unit dependencies are provided for each of the example sentences, showing that the DAG representation can be used as a basic framework for describing Japanese syntax, as an alternative to phrase-structure trees. For

grammatical features, it is shown that the morphemes in a syntactic unit provide information to determine its grammatical features. Along with the grammatical features mentioned in this chapter (TENSE, ASPECT, MOOD, and VOICE), in further work it will be necessary to represent other grammatical features in the framework of LFG. LFG f-structures go beyond DAG representations in that (1) they show grammatical features not represented in the DAG representations and (2) they show control and raising (anaphoric and functional) not represented in the DAG representations.

Chapter 5

KNP and Kyoto Corpus Ver.4

5.1 Introduction

This chapter introduces KNP and the Kyoto Text Corpus ver.4 (KTC4). KNP ([Kurohashi and Nagao, 1998]) is a dependency parser for Japanese. KNP takes the output of JUMAN ([Kurohashi and Kawahara, 2005]), a Japanese morphological analyser, as its input. The accuracy of both JUMAN and KNP was improved while being used for the development of KTC4, an unlabelled dependency, POS and morphologically annotated corpus of Japanese text ([Kurohashi and Nagao, 1998]). Section 5.2 describes how KNP parses Japanese text and Section 5.3 describes how texts are annotated in KTC4.

5.2 KNP

KNP is a dependency parser for Japanese. KNP takes the output of JUMAN, a morphological analyser for Japanese which was first developed by Sadao Kurohashi in 1992 ([Kurohashi and Kawahara, 2005]). Morphological information is important to identify the boundaries of bunsetsu, or syntactic units, and also to determine the dependency relationships among them. The process of dependency parsing for Japanese using KNP proceeds as follows: first, JUMAN changes each of the input sentences into a sequence of morphemes which are annotated with morphological features, then this sequence of annotated morphemes is sent to KNP to produce an unlabelled dependency tree for the sentence. Before explaining KNP, it is necessary to describe how JUMAN works.

5.2.1 Morphological Analysis of Japanese

One of the characteristics of Japanese text is the absence of white spaces between words. What is more, a sentence ends with punctuation-based delimiters. The first task of morphological analysis for Japanese is to identify the boundaries between syntactic units.

As we have seen in Chapter 3, bunsetsus are the basic unit of syntactic dependency for Japanese sentences. Each unit has one head, and the type of particle following the head determines its grammatical function in the sentence.

Identifying unit boundaries often requires disambiguation, especially when the sentence is written in hiragana script. Using different script systems is the means to make unit or word boundaries explicit; in ordinary Japanese text, particles and verbal inflections must be written in hiragana, and the majority of nouns and the root of verbs are written in Chinese characters or in katakana. It is a formidable task to read an ordinary Japanese text all written in hiragana or in katakana, and it is impossible to write Japanese text in Chinese characters only.

In example (5.1), assume that the noun 'karada'(body) is written in hiragana. In this case, the form happens to be the same as a particle plus a copula, and hence this fragment has two different analyses:

(5.1) hashirukarada

hashir-u-kara-da
run-decl.base-because-copula.decl.base
"It is because someone runs."

hashir-u *karada*
run-decl.base body
"a body that runs"

Semantically, given appropriate context, it is rather easy to choose the intended interpretation. Morphologically, however, it is not obvious which analysis is appropriate, especially when it comes to automatic morphological analysis. JUMAN is one of the answers to these problems.

5.2.2 Morphological Analysis by JUMAN

JUMAN implements the idea of a minimum cost path analysis for resolving ambiguity in identifying morphemes in an input string. The analysis is to

find paths, or a sequence of morphemes, and choose the minimum cost path as the output. Minimum cost path analysis has been implemented in input methods for Japanese ([Hisamitsu and Nitta, 1991]), such as MS-IME.

JUMAN takes a sequence of Japanese characters as input and yields a path (or paths) of morphemes with minimum cost as output. Morphological analysis for each input sequence of characters proceeds as follows ([Kurohashi and Kawahara, 2005]):

1. Dictionary lookup: find all character combinations which are registered in the morpheme dictionary.

2. Connectability check: look up the connectability dictionary, and find possible connections between adjacent morphemes, and the cost of each connection.

3. Cost calculation: calculate the costs of each of the possible morpheme sequences for the input, and choose the minimum cost path as the morphological analysis for the input. Lemma weight, POS costs and connectivity costs contribute to overall costs.

Dictionary Lookup

Dictionary lookup tries to find all the possible morphemes that begin with a certain character in the input sequence. The dictionary of JUMAN has the following format in BNF (5.2) (each term is translated from Japanese):

(5.2)

```
<definition>        ::=  (<#pos><morphemes>)|
                         (<#pos>(<#subpos><morphemes>))
<morphemes>         ::=  <info> | <info><morphemes>
<info>              ::=  (<lemma info><phonology info><inflection info><semantics info>)
<lemma info>        ::=  (lemma <lemma sequence>)
<lemma sequence>    ::=  <lemma content> | <lemma sequence>
<lemma content>     ::=  <#spelling> |
                         (<#spelling> <#value>)
<phonology info>    ::=  (pronunciation <#pronunciation in hiragana characters>)
<inflection info>   ::=  (inflection <#inflection type>) | NIL
<semantics info>    ::=  (semantics <#semantic description>) | NIL
```

`<#pos>` and `<#subpos>` must be defined in the morpho-pos dictionary. If a morpheme is defined in terms of its `<#pos>`, then its `<#subpos>` must also be defined.

`<inflection info>` cannot be omitted if it is an inflectional pos.

`<lemma content>` is a pair of the base declarative form of a morpheme and its weight. The weight is 1 in default, and omitted if it is in default.

As for inflecting morphemes, their base declarative forms are written in `<#pronunciation in hiragana characters>`.

Texts in "" can be written in `<#semantic description>`.

Consider the latinised character sequence below:

(5.3) hashirukarada

The part from the first character 'h' to the seventh character 'u' corresponds to the morpheme 'hashiru' registered in the morpheme dictionary, hence this part is a possible morpheme in a morphological analysis of the sequence. The dictionary also specifies the part of speech of each morpheme. If the morpheme is an inflecting part of speech, then the inflection form dictionary is looked up in order to determine the inflection form of the morpheme. In the example, the morpheme 'hashiru' is a verb in the declarative base form. The entry for this morpheme in the dictionary is (5.4); the terms except for lemma and orthography are translated from Japanese, and 'inflection type-I r' means that this is a type-I verb whose root ends with a consonant 'r':

(5.4)

(verb ((pronunciation hasiru)(lemma 走る はしる)(inflection type−I r)
(semantics ″standard lemma: 走る″)))

Starting with the eighth character 'k', there are two possible ways to analyse the rest of the character sequence in example (5.1): one is to divide the part into two morphemes 'kara' and 'da', and the other is to keep the part as one morpheme 'karada'. The entries for them are (5.5):

(5.5)

(noun (normal ((pronunciation karada)(lemma 身体 体 (からだ 1.6))
(semantics ″standard lemma:身体″))))

(particle (conjunctive ((lemma から)(pronunciation kara))))

(copula ((lemma だ) (pronunciation da) (inflection copula)))

The noun "karada" has three different lemma forms, and the last one, which is written in hiragana, has weight 1.6, while other lemmas have the default weight 1, since it is less common to write this lemma all in hiragana characters. The first lemma is the standard lemma as specified in "semantics".

Each morpheme is given a "part of speech cost" according to which part of speech (POS) the morpheme belongs to. The POS costs in the JUMAN dictionary are shown in Table 5.1:

Table 5.1: Part-of-Speech Costs

pos	subpos	pos cost
special	*	100
verb	N/A	100
adjective	N/A	100
copula	N/A	11
auxiliary	N/A	10
noun	cardinal	40
noun	formal	70
noun	adverbial	70
noun	others	100
demonstrative	adverbial	60
demonstrative	others	40
adverb	*	100
particle	sentence ending	20
particle	others	10
conjunctive	N/A	100
noninflecting adjective	N/A	100
exclamative	N/A	110
prefix	*	50
suffix	nominal predicative	14
suffix	nominal	35
suffix	nominal cardinal	35
suffix	nominal special	35
suffix	adjectival predicative	14
suffix	adjectival nominal	14
suffix	verbal	14
unspecified	katakana	5000
unspecified	alphabet	100
unspecified	others	5000
*	N/A	10

After dictionary lookup of “hashirukarada”, we have two possible morpheme sequences: “hashiru-kara-da” and “hashiru-karada”. These are passed on to the next step.

Connectability Check

Each pair of morphemes in morpheme sequences is checked in terms of its connectability by looking up the connectability dictionary. Two morphemes are connectable if they can appear next to each other in a sequence. Connectability is defined by the morphological environment in which certain types of morphemes can appear side by side, and it is represented in the connectability dictionary in the following BNF (5.6):

(5.6)

```
<connectability rule>  ::=  ((<morphemes>) (<morphemes>))|
                            ((<morphemes>) (<morphemes>) <connectability cost>)
<morphemes>            ::=  <structure>|
                            <structure> <morphemes>
<structure>            ::=  (<#pos>) |
                            (<#pos><#subpos>) |
                            (<#pos><#subpos><#inflection type>) |
                            (<#pos><#subpos><#inflection type><#inflection form>) |
                            (<#pos><#subpos><#inflection type><#inflection form><#lemma>)
```

"*" can be used for <#pos>, <#subpos>, <#inflection type >, <#inflection form>, <#lemma>.
<#connectability cost> is an integer from 0 to 255 (default: 10).

For example, the two morphemes "kara-da" are connectable since the connectability dictionary has the rule (5.7). This rule states that a copula can follow particles as specified, including "kara", and that its cost is 10 by default:

(5.7) (((particle case particle * * made)
(particle adverbial particle * * dake)
(particle adverbial particle * * bakari)
(particle adverbial particle * * bakkari)
(particle adverbial particle * * nomi)
(particle adverbial particle * * nado)
(particle adverbial particle * * nanka)
(particle adverbial particle * * kurai)
(particle adverbial particle * * gurai)
*(particle case particle * * kara)*
(particle conjunctive particle * * ka)
(particle conjunctive particle * * kara)
(particle conjunctive particle * * ka)
(particle conjunctive particle * * no))
((copula)))

Cost Calculation

The cost of an output is the sum of the costs of individual morphemes and of the costs of connectability between each of the adjacent morphemes in a possible morpheme sequence, where POS costs can be multiplied by weights (e.g., a noun "karada" written in hiragana has weight 1.6). Morpheme costs and connectability costs are manually specified and listed in the morpheme dictionary and the connectivity dictionary, and used for cost calculation.

The process of cost calculation is as follows ([Kurohashi et al., 2005]). Dummy morphemes "start" and "end" are postulated at the beginning and the end of the input string, providing the start and end points of a DAG representing possible morphological analyses. The remaining vertices of the graph are the morphemes in the input string, with their POS costs and weights. Among possible graphs for the same input string, the graph with the minimum cost is selected as the morphological analysis of the input string.

The graph from the example "hashirukarada" is given in Figure (5.1):

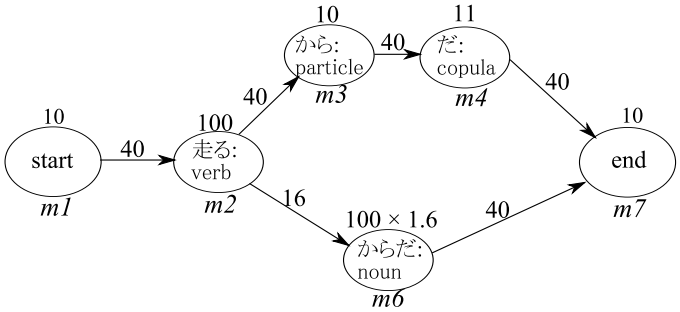

Figure 5.1: DAG for "hashirukarada"

In the graph, each of the vertices represents a morpheme, and each of the arcs represents the connection between two morphemes. The paths from *m1* to *m7* represent the sequences of candidate morphemes. The POS cost of each morpheme is specified above the vertex, and the connectability cost is specified above the arc. All the vertices except for m6 are weighted by 1. The vertex *m6* is weighted by 1.6 because the entry for the noun represented by m6 (5.5) specifies that it should be weighted if it is written all in hiragana characters. The output is the least costly graph among the possible graphs from "start" to "end". In Figure (5.1), the path *m1m2m3m4m7* is less costly than the path *m1m2m6m7*, hence "hashiru(verb)-kara(particle)-da(copula)" is produced as the morphological analysis. The actual output by JUMAN is Figure 5.2; each line corresponds to one morpheme:

走る はしる 走る 動詞 2 * 0 子音動詞ラ行 10 基本形 2 "代表表記:走る"
から から から 助詞 9 接続助詞 3 * 0 * 0 NIL
だ だ だ 判定詞 4 * 0 判定詞 25 基本形 2 NIL

Figure 5.2: JUMAN output for "Hashirukarada"

It is possible that the same string yields more than two output paths, if these paths have the same cost. The cost of morphemes is set based on their part of speech. Therefore, the cost of two homonyms is the same if they belong to the same part of speech, or if they belong to different parts of speech whose costs are the same, hence there can be two outputs for the same string. Ambiguities in morphological analysis due to the same cost can sometimes be resolved in the KNP dependency analysis.

5.2.3 Dependency Analysis by KNP

A dependency analysis of a Japanese sentence can be represented as a DAG for the sentence (cf. Section 3.3.2). Its vertices are the syntactic units (bunsetsu) of the sentence, and its arcs are the dependency relations.

KNP processes each morpheme sequence in the output of JUMAN in the following steps:

1. Disambiguate homonyms (morphemes in the same form but different meanings) if any.
2. Annotate each morpheme with features used for unit construction.
3. Annotate each unit with features.
4. Generate an unlabelled dependency tree according to the features.

In the JUMAN/KNP analysis, which aims at providing deep linguistic resources, the input is morphologically analysed before unit boundary identification takes place; JUMAN analyses the input into a sequence of morphemes, but does nothing about unit boundaries. It is the task of KNP to compute unit boundaries, and to link units in terms of an unlabelled dependency analysis.

As far as Japanese is concerned, it would also be possible to have unit boundary identification (before deep morphological analysis) by finding particles and assuming them as unit boundaries; this would result in a kind of shallow dependency parser, and comparing the performance of dependency analysis of after-morphology deep analysis (like JUMAN/KNP) to that of before-morphology shallow analysis will be one of the topics of future research.

Homonym Disambiguation

First of all, homonyms must be disambiguated, if there are any in the input. Consider the string "aru jinbutsu"; in Section 3.6.3, I pointed out that this sentence has two possible analyses because of the homonym "aru":

(5.8) *aru* *jinbutsu*
certain person
"a certain person"

(5.9) *aru* *jinbutsu*
exist-decl.base person
"a person that exists"

(5.10) is the output by JUMAN for "aru jinbutsu"; the '@' at the beginning of the second line specifies that this morpheme is a homonym of the morpheme specified in the first line:

(5.10)
```
ある ある ある 動詞 2 * 0 子音動詞ラ行 10 基本形 2 "補文ト 代表表記:有る"
@ ある ある ある 連体詞 11 * 0 * 0 * 0 "代表表記:或る"
人物 じんぶつ 人物 名詞 6 普通名詞 1 * 0 * 0 "代表表記:人物"
```

Homonym disambiguation in KNP is rule-based. Rules have the following format:

(5.11)
```
(
(previous morpheme)
([pos features][pos features]*)
features
)
```

The bracketed part after `(previous morpheme)` lists the possible POS features of the homonymous morphemes. The order of these possible POS features reflects the preference among them, and the first feature is given to the morpheme. This preference order is set manually. If other features must be assigned to it, they are specified in `features`.

For example, rule (5.12) states that if there are two possible POS features for a given homonym, one of which is verb and the other is non-inflecting adjective, then this homonym is analysed as a verb with the additional feature 'light verb', regardless of its previously assigned morpheme.

(5.12)

```
(
(*?)
([verb] [non-inflecting adjective])
 light verb
)
```

When this rule is applied to the example "aru jinbutsu", then 'aru' would be incorrectly analysed as a verb. Such misanalyses will be fixed in the next step.

KNP has 62 manually-constructed rules for homonym disambiguation. All the rules are applied to each homonym, until one of the rules applies and the homonym is disambiguated.

Feature Annotation on Morphemes

After homonym disambiguation, morphemes are assigned features used for combining morphemes into syntactic units. These features are assigned based on the environment in which morphemes appear. The environment is specified by rules in the same format as (5.11).

KNP has 351 rules for feature annotation. The main features given at this step are: POS change, compounding, and syntactic unit boundaries.

POS Change

There are cases where the morphological environment is required to reliably determine the POS of certain morphemes. The "aru" in (5.8) is one such case. Rule (5.13) changes the POS of "aru" from verb to non-inflecting adjective, if it appears in the beginning of a sentence:

(5.13)

```
(( ( ?* ) ( [verb * * * aru (beginning of sentence)] ) ( ?* )
&poschange:non-inflecting adjective:* ))
```

Compound nouns

There are a variety of rules for compounds, and many of them reflect lexical or conventional aspects of noun compounding. In general, a morpheme is given a feature "←" if it can constitute a compound with the previous morpheme, and "^←" if not. A morpheme is given a feature "→" if it can constitute a compound with the following morpheme, and "^→" if not.

For example, if a Chinese-origin adverb comes directly before a noun, then it is likely that together they constitute a compound noun; in (5.14), the adverb "futsu" is a part of the compound. There is no unit boundary between "futsu" and "senkyo":

(5.14) Futsusenkyo

futsu-senkyo
ordinary-election
(Lit.)"ordinary election" or
"popular election"

If the same adverb appears before morphemes other than nouns, then they do not constitute a compound; in (5.15), e.g. the adverb "futsu" does not constitute a compound with the following verb. There is a unit boundary between "futsu" and "okonawareru". In addition, an inflected element cannot constitute a nominal compound:

(5.15) Futsu okonawareru senkyo

futsu	*okonaw-arer-u*	*senkyo*
ordinarily	do-passive-decl.base	election

"an election which is held usually"

By means of feature annotation rules, the noun "senkyo" in (5.14) is assigned a "←" feature, indicating that "futsu" and "senkyo" constitute a compound. However, the same noun is not assigned a "←" feature in (5.15), indicating that "okonaw-arer-u" and "senkyo" do not constitute a compound.

Syntactic Unit Boundary
Basically, KNP tries to determine the leftmost morpheme of a syntactic unit, and will assign the feature "leftmost morpheme" if such a morpheme can be found. A prefix or an open-class content word (noun, verb, adjective) can be the leftmost morpheme of a syntactic unit. The syntactic requirements for such a morpheme to constitute the leftmost morpheme in a given morpheme sequence are that it appears at the beginning of a sentence or that it cannot be compounded to the previous morpheme. The requirements can be summarised in rule (5.16); the format is the same as that for other feature assignments:

(5.16) ((*? "^→") ((prefix "^←") (content "^←")) leftmost morpheme)

Applied to example (5.15), the adverb "futsu", the verb "okonawareru", and the noun "senkyo" cannot be compounded to their previous morphemes, hence each of them functions as a single syntactic unit.

The accuracy of unit-boundary identification by the simple rule such as (5.16) depends on the accuracy and coverage of noun-compound rules.

Feature Annotation on Units

Each syntactic unit is assigned with a feature relevant for subsequent coordination checks and dependency resolution. Feature assignment is rule-based, and rules have the following format:

(5.17)
```
(
(previous unit)
(current unit)
(following unit)
feature
)
```

For example, rule (5.18) specifies that a unit has "-wo" case if its head is not an inflecting element and if it is followed by the particle "-wo", regardless of what kind of unit comes before and after itself:

(5.18)
```
(
( ?* )
( < ( ?* [* * * * * ((^inflecting pos))] [particle * * * wo] ) > )
( ?* )
 case: wo
)
```

KNP has 675 manually-constructed rules for feature annotation on units. All rules are applied to each of the units in the input, and when one of the rules matches a unit, then the feature that the rule specifies is annotated to the unit. One match does not break the feature assignment loop, hence one unit can be assigned more than one feature. If none of the rules applies to a unit, then this unit receives a "NONE" feature assignment by default. No unit is left featureless.

Dependency Generation

Dependency relations (mainly unlabelled) among units are determined by rules in the following format:

(5.19)
```
(
(dependendent unit feature)
(dependee unit feature, dependency type )
```

```
boundary
preference
)
```

KNP distinguishes between three coarse-grained dependency types: apposition ('A'), coordination ('C') and others ('D'). KNP uses all the 40 dependency rules for each of the units in an input sentence. If the unit feature of a unit matches the `dependent unit feature` of a rule, then the rule checks whether the input contains any further unit which matches the `dependee unit feature` in the same rule. If there is such a unit, then it is the dependee unit, and the rule is applied to both units, establishing a dependency relationship between them. The dependency assignment loop for a unit breaks when it receives a dependency type feature and its dependee is determined.

The `boundary` feature is a unit feature beyond which a dependency cannot go, i.e., if there is a unit which matches `boundary`, then all the units after this unit cannot be a dependee unit for the current dependent unit. Feature matching stops at boundary units, and feature checking using other rules continues. The presence of stop points for feature matching makes the process shorter, especially when the input sentence contains more than one subordinate clause. The arguments and adjuncts of one subordinate clause are never dependent on other clauses.

`Preference` indicates the preferred distance between the dependent and the dependee units. If, for example, `preference` is set to 1 for a certain rule, and there are more than one possible dependee unit for the current dependent unit, then the nearest among these dependee units is chosen as the dependee unit. If `preference` is set to 2, then the second nearest is chosen, and so on.

For example, the skeletal rule (5.20) specifies that a unit with case "-ga" is dependent on the nearest possible inflecting unit, and that its dependency is D, i.e., neither apposition nor coordination:

(5.20)
```
(
(case: ga)
(inflecting, D)
()
1
)
```

Actual dependency rules carry more detailed feature specifications in order to reliably handle a variety of environments in which a given type of morpheme may appear.

After the steps above, KNP outputs the dependency parse for input sentences. The parser output for a sentence has the following format:

(5.21)

```
# Sentence ID
* target.unit.number <unit tags>*
morpheme pronunciation lemma pos subpos <morpheme tags>*
morpheme pronunciation lemma pos subpos <morpheme tags>*
* target.unit.number <unit tags>*
morpheme pronunciation lemma pos subpos <morpheme tags>*
...
* -1D <unit tags>*
morpheme pronunciation lemma pos subpos <morpheme tags>*
EOS
```

Lines starting with '*' except for the sentence ID line specify unit boundaries. The target unit number (the unit on which a given unit depends) and unit tags are given at this line. Each of the following lines specify each of the morphemes in the unit, until the next unit boundary line. The last syntactic unit does not have a target unit, hence its target unit number is "-1D".

For example, KNP output for sentence (5.22) is (5.23):

(5.22) Watashiga kono honwo yonda.

watashi-ga	*kono*	*hon-wo*	*yom-ta.*
I-SUBJ	this	book-OBJ	read-decl.ta

"I read this book."

(5.23)

```
# S-ID:1 KNP:2008/07/04
* 3D <文頭><一人称><ガ><助詞><体言><係:ガ格><区切:0-0><RID:1093><格要素><連用要素>
私 わたくし 私 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:訓 代表表記:私" <漢字読み:訓><代表表記:私>
<品曖><音訓解消><ALT-私-わたし-私-6-1-0-0-"代表表記:私">
<品曖-普通名詞><品曖-その他><文頭><一人称><漢字><かな漢字><名詞相当語><自立><タグ単位始><文節始>
が が が 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* 2D <連体修飾><連体詞形態指示詞><タグ単位受無視><係:連体><区切:0-4><RID:1448>
この この この 指示詞 7 連体詞形態指示詞 2 * 0 * 0 NIL <かな漢字><ひらがな><自立><タグ単位始><文節始>
* 3D <ヲ><助詞><体言><係:ヲ格><区切:0-0><RID:1099><格要素><連用要素>
本 ほん 本 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 代表表記:本" <漢字読み:音><代表表記:本><漢字><かな漢字>
<名詞相当語><自立><タグ単位始><文節始>
を を を 助詞 9 格助詞 1 * 0 * 0 NIL <かな漢字><ひらがな><付属>
* -1D <文末><時制:過去><句点><用言:動><レベル:C><区切:5-5><ID:(文末)><RID:112><提題受:30>
読んだ よんだ 読む 動詞 2 * 0 子音動詞マ行 9 タ形 8 "代表表記:読む" <代表表記:読む><表現文末><かな漢字>
<活用語><自立><タグ単位始><文節始>
。 。 。 特殊 1 句点 1 * 0 * 0 NIL <文末><英記号><記号><付属>
EOS
```

5.2.4 Summary

This section provides an overview over JUMAN, a morphological analyser for Japanese, and KNP, a dependency parser for Japanese. KNP operates on

the output of JUMAN. Morphological information is important to identify the boundaries of bunsetsu, or syntactic units, and also to determine the dependency relationship among them. The process of dependency parsing for Japanese using KNP is as follows: JUMAN analyses each of input sentences into a sequence of morphemes which are annotated with relevant features, then this sequence of annotated morphemes is sent to KNP to determine unit boundaries and output a (coarse-grained and mostly unlabelled) dependency tree for the sentence.

5.3 Kyoto Corpus Ver.4 (KTC4)

5.3.1 Introduction

KNP is a rule-based dependency parser, and the rules had to be corrected and extended manually in order to achieve optimal coverage and accuracy. In order to obtain accurate and comprehensive rules, it is possible to analyse text data using KNP, and to iteratively improve the rules so that they yield correct parses. This parser improvement goes hand-in-hand with the construction of a parsed corpus. This section introduces Kyoto Corpus version 4.0 (KTC4), a Japanese parsed corpus which was built while improving KNP ([Kurohashi and Nagao, 1998]).

5.3.2 Overview of KTC4

KTC4 is a corpus of Japanese newspaper text which was automatically parsed and annotated by KNP with part-of-speech tags and dependency relations between syntactic units, and then corrected manually. KTC4 contains about 40,000 sentences taken from Mainichi Shimbun, a Japanese newspaper, annotated with POS tags, morphological tags and lemma forms; in addition, each sentence is segmented into syntactic units (bunsetsu). Abstract, unlabelled dependency relations among syntactic units are specified. 5447 sentences are also annotated with case relations, ellipsis, and coreference information. The basic statistics of the KTC4 data are given in Table 5.2:

Table 5.2: Statistics of KTC4

Number of words	972894
Number of units	372130
Number of sentences	38400
Words per unit	2.614
Units per sentence	9.691
Words per sentence	25.336

POS and Dependency Annotation Scheme

As an example, the KTC4 POS and dependency annotation for sentence (5.24) is given in (5.25). The pronunciation is transcribed into English:

(5.24) Somariano shutomogadisiode enjokatsudowo tsudukeru kokurenkikanya hiseifusoshikiwa juusannichi subeteno katsudowo chuushishita.

[Somaria-no]$_{f1}$ *[shuto-mogadishio-de]*$_{f2}$ *[enjo-katsudo-wo]*$_{f3}$
Somalia-PADJ capital-Mogadishu-PADJ aid-activity-OBJ
[tsuzuke-ru]$_{f4}$ *[kokuren-kikan-ya]*$_{f5}$
continue-decl.base UN-organization-and
[hi-seifu-soshiki-wa]$_{f6}$ *[juusan-nichi]*$_{f7}$ *[subete-no]*$_{f8}$
non-government-organization-TOP thirteen-day all-PADJ
[katsudo-wo]$_{f9}$ *[chuushi-shita]*$_{f10}$
activity-OBJ stop-do.decl.ta
"The UN organisations and NGOs working for aid in Mogadishu, the capital of Somalia, stopped all their activities on 13th (of January, 1995)".

(5.25)

```
# S-ID:950115048-002 KNP:99/07/26
* 0 1D
ソマリア Somalia * noun place * *
の no * particle conjunctive * *
* 1 3D
首都 shuto * noun normal * *
モガディシオ Mogadish * noun place * *
で de * particle case * *
* 2 3D
援助 enjo * noun sahen * *
活動 katsudo * noun sahen * *
を wo * particle case * *
* 3 5D
続ける tsudukeru * verb * vowel base
* 4 5P
国連 kokuren * noun organization * *
機関 kikan * noun normal * *
や ya * particle conjunctive * *
* 5 9D
非 hi * prefix na-adjectival * *
政府 seifu * noun normal * *
組織 soshiki * noun sahen * *
は wa * particle adverbial * *
*6 9D
十三 juusan * noun numeral * *
日 nichi * suffix nominal-cardinal * *
、 , * special comma * *
* 7 8D
全て subete * adverb * * *
の no * particle conjunctive * *
* 8 9D
活動 katsudou * noun sahen * *
を wo * particle case * *
* 9 -1D
停止 teishi * noun sahen * *
した shita suru verb * sahen ta
。 . * special period * *
EOS
```

The first line in (5.25) gives the sentence ID. The lines which start with '*' are the first lines of syntactic units. They also specify the unit ID number and the target unit ID number of the unit on which this unit is dependent, and the character after the target unit ID specifies the type of dependency. For example, the lines starting with '*' in (5.25) above specify that the 0th unit is dependent on the 1st unit, the 1st and the 2nd units on the 3rd, and the 3rd and the 4th on the 5th, and so on. If a unit does not have any target unit, then it is the root unit of the sentence, and this is indicated by "-1D".

Each line between a unit ID and the next corresponds to one morpheme in the unit (in the order in which they occur in the string) and the tags relevant to this unit. This format is similar to the format of KNP output.

Case Relation Annotation Scheme

As mentioned before, a subset of 5447 sentences in KTC4 carries additional case, ellipsis and coreference mark-up. An example of KTC4 annotation of case relations for sentence (5.26) is given in (5.27):

(5.26) Nipponto Kankokuwa, kotoshide kokkouseijoukakara sanjuunenwo mukaeta.

[Nippon-to]$_{f1}$ *[Konkoku-wa,]*$_{f2}$ *[kotoshi-de]*$_{f3}$
Japan-and Korea-TOP this.year-PADJ
[kokkou-seijou-ka-kara]$_{f4}$ *[sanjuu-nen-wo]*$_{f5}$
diplomatic.relations-normal-suff-PADJ thirty-years-OBJ
[mukae-ta]$_{f6}$
meet-decl.ta
"Japan and Korea meet thirty years this year since the normalisation of diplomatic relations." or
"Japan and Korea mark this year their 30th anniversary of normalisation of diplomatic relations."

(5.27)

```
# S-ID:950105042-001 KNP:96/11/17 MOD:2003/05/20
* 0 1P
nippon * noun place * *
to * particle case * *
* 1 5D
kankoku * noun place * *
wa * particle adverbial * *
、* special comma * *
* 2 5D
kotoshi * noun temporal * *
de * particle case * *
* 3 5D <rel type="no" target="nippon" sid="950105042-001" tag="0"/>
<rel type="no" target="kankoku" sid="950105042-001" tag="1"/>
kokkou * noun normal * *
+ 4 6D <rel type="ga" target="kokkou" sid="950105042-001" tag="3"/>
seijou seijouda adjective * na root
ka * suffix nominal * *
kara * particle case * *
* 4 5D
sanju * noun numeral * *
nen * suffix nominal * *
wo * particle case * *
* 5 -1D <rel type="ga" target="kankoku" sid="950105042-001" tag="1"/>
<rel type="wo" target="sanjunen" sid="950105042-001" tag="5"/>
<rel type="jikan" target="kotoshi" sid="950105042-001" tag="2"/>
<rel type="kara" target="seijouka" sid="950105042-001" tag="4"/>
<rel type="ga" target="nihon" sid="950105042-001" tag="0"/>
<mode rel="ga">AND</mode>
mukaeru * verbal * vowel base
。* special comma * *
EOS
```

The tags are annotated on relevant syntactic units or morphemes, and they have the following format:

(5.28) <rel type="a" target="b" sid="c" tag="d">

The syntactic unit annotated with (5.28) has a target word "b", the "d"th unit (in the sentence) whose ID is "c", and the relation type is "a". Relation types are the names of case particles. In example (5.27), the root unit "mukaeta" is annotated with five case relation tags (the 6th tag specifies that

the units which have the relation type "ga" are in a coordination relation). The first of these five tags is (5.29):

(5.29) <rel type="ga" target="kankoku" sid="950105042-001" tag="1">

This tag specifies that the unit which carries this annotation (in this example, "mukaeta") has a target "kankoku", the 1st unit in the sentence whose ID is "950105042-001", and the relation type is "ga", the case particle for the grammatical function SUBJ, meaning that "kankoku" is the subject of "mukaeta".

Notice that the unit "kankoku" does not have the case particle "ga", but "wa". This means that the relation type does not specify the case particle which the unit actually has; rather, the relation type specifies the grammatical function of the unit (though it does not use the term "SUBJ" or "OBJ", but the name of case particles).

The sentence ID enables us to specify the inter- and intra-sentential reference. If the sentence ID specified by a given relation tag is not the same as the sentence ID of itself, then the target is not in the sentence.

The identity of an extra-sentential target can be indefinite, such as people in general, or the speaker/writer.

Nominal units are also annotated with case relations. The third unit "kokkouseijouka (normalisation of diplomatic relations)" is annotated with three relation tags, and the last one is annotated with the morpheme "seijouka". The first two of these three tags specify that "nippon" and "kankoku" both have the relation which can be specified by the case particle "no (*of*)". The last of the three specifies that the morpheme "seijou", which is an adjective, can take "kokkou" as its subject.

The Advantage of Common Format of KTC4/KNP

The format of KTC4 annotations strongly resembles that of KNP outputs. This similarity is due to the fact that both have been developed in tandem. The main differences between the two annotation schemes are that syntactic units are annotated with their unit IDs in KTC4, but not in KNP. In addition, KNP contains unit features and morpheme features which are used for morphological analyses, unit boundary analyses, and dependency analyses by JUMAN and KNP. KTC4, on the other hand, contains only the features produced as a result of these analyses.

The resemblance between KTC4 annotations and the outputs of KNP enables us to develop an NLP application which can be applicable to both of them. For example, an application based on KTC4 input can be used also for the output of KNP. The LFG functional equation annotation on KTC4

is an attempt to utilise KTC4-style information for acquiring LFG resources from raw text parsed by KNP, which will be presented in the next chapter.

5.4 Summary

This chapter has introduced KNP and the Kyoto Corpus version 4 (KTC4). KNP is a dependency parser for Japanese, and its accuracy was improved while being used for the development of the Kyoto Corpus, a parsed corpus of Japanese. KNP takes the output of JUMAN, a Japanese morphological analyser, as its input. The output of JUMAN is a sequence of morphemes for an input sentence, and KNP establishes the unit boundaries in the sequence, and the dependency relations among these units. Both JUMAN and KNP are rule-based. KTC4 has been developed while improving the accuracy of KNP rules. In addition, KTC4 representations and KNP output share similar formats, hence it will be easy to develop an NLP application for both of them.

Chapter 6

A Method for the Automatic Annotation of F-Structure Functional Equations to KTC4 Representations and KNP Output

6.1 Introduction

This chapter introduces the LFG annotation method for KTC4 representations and KNP output. I use KTC4 as the corpus from which wide-coverage LFG resources are acquired. The method I introduce implements the idea that the part-of-speech tags on each morpheme and the unlabelled dependency tags on each syntactic unit in KTC4 provide us with enough information for constructing what [Cahill et al., 2003, Cahill et al., 2004, Cahill, 2004] call "proto" f-structures for the texts in the corpus, without employing PCFG-style syntactic trees.

The f-structure functional annotation method assumes the ideas which I have introduced in Chapter 4: each of the syntactic units in a sentence corresponds to a sub-f-structure; they combine to form the f-structure for the sentence; the core aspects of the f-structure (except for grammatical features and control relations) are represented in terms of a DAG.

KTC4/KNP represent unlabelled dependency relations among syntactic units which are annotated with part-of-speech tags on each morpheme, and the information represented in KTC4 annotations enables us to acquire LFG functional equations automatically for a large number of Japanese sentences.

Furthermore, applying the automatic annotation method to KNP output, we can produce f-structure representations for raw text, which can be utilised for other applications. Essentially, the annotation method is to label each of the unlabelled dependency relations among syntactic units, and capture the labels and the local morphological information in terms of LFG functional equations. By resolving these equations, we produced the f-structure representation for the input sentence.

However, this method encounters problems when the representation in KTC4/KNP is used as is; some additional operations for improving and enhancing them are required in order to acquire better LFG representations for the input. The major problems are coordination, one-to-many correspondences between one unit and more than one f-structure, and zero-pronoun identification. Some of them require special treatments since the information which is encoded explicitly by tags in KTC4/KNP is not enough to derive the acquisition of fully appropriate LFG f-structures.

This chapter has the following structure: Section 6.2 gives an overview of the automatic annotation method, its component parts and the flow of information. Section 6.3 explains how appropriate functional equations are assigned to syntactic units. In Section 6.4, the treatment of coordination is explained. Section 6.5 describes how problematic morphological environments which fail to be assigned with a correct grammatical function are treated. In Section 6.6, the process of assigning zero-pronoun equations is introduced. The evaluation of the f-structure representation acquired by the method developed in this chapter is the topic of Chapter 7.

6.2 Overview of the Method

The automatic f-structure annotation method is summarised as follows:

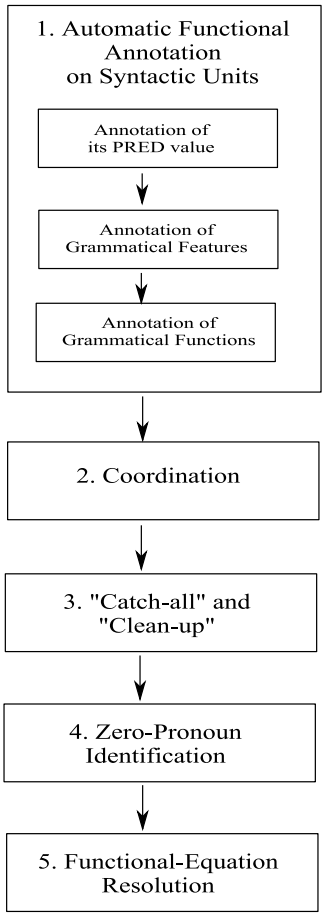

Figure 6.1: The automatic f-structure annotation method for Japanese

6.3 Automatic Functional Annotation on Syntactic Units

In principle, there are three types of functional equations to be annotated for one syntactic unit; its PRED value, its grammatical features, and its grammatical or discourse function. In this study, I make no distinction between grammatical functions and discourse functions.[1]

The policy of automatic annotation implemented in this study is to add everything which is obvious from the tag information in the original corpus

[1]This is because the function assignment of TOPIC and FOCUS in Japanese is in principle the same as the function assignment for other grammatical functions (cf. Section 4.1).

(or parser output), so as to be able to fully reconstruct the original corpus information from the functional equations we acquire automatically. This issue is relevant to e.g. natural language generation from f-structure representation. Therefore, all grammatical features in this study are based on what is explicitly specified morphologically.

Some grammatical-feature specifications, such as tense disambiguation (cf. Section 4.3.1), disambiguation of "-teiru aspect" by adverbial modification (cf. Section 4.3.2) and idiomatic modal expressions (cf. Section 4.3.3), are not addressed in this study; they need special treatments which cover the semantics of adverbial modification and other morpho-syntactic constructions, and a variety of idiomatic expressions should also be handled properly, which are objectives of further study. F-structure representations for KTC4/KNP will be of help for this direction of research.

6.3.1 Annotation of PRED Values

A syntactic unit consists of at least one content word or morpheme (cf. Section 3.3), and the lemma corresponding to this word is (or these words are) the value of the PRED attribute of the f-structure corresponding to this unit. Particles and punctuation marks are not content words, while words of all the other parts of speech are content words. For example, a syntactic unit "watashi-ga (I-SUBJ)" has "watashi" as its content word, hence "watashi" is the value of the PRED attribute of the f-structure which corresponds to this unit, and the case particle just contributes to the value of grammatical feature "CASE".

There are some instances in which one syntactic unit has more than one content word, for example, compound nouns and verbs with a series of suffixes or auxiliaries. In this study, the sequence of the lemmas of these content words as a whole is considered to be the value of the PRED attribute. A more detailed automatic analysis will be required in terms of the intra-unit dependency relationship among the compounded nouns or among the main verb and verbal suffixes, but this is one of the objectives of future research.

As for valence-changing verbal suffixes, the value of the PRED attribute contains the main verb followed by the suffix. For example, a verbal unit "yomaseru (make someone read)" has one main verb "yom-u" and a causative suffix "-ase-ru". For the f-structure Fx which corresponds to this verbal unit, the functional equation for the PRED value is "Fx:PRED===‘yomaseru<SUBJ, OBL, OBJ>"'.

6.3.2 Annotation of Grammatical Features

As we have seen in Section 4.3, grammatical features such as TENSE, ASPECT, MOOD and VOICE are specified morphologically. Part of speech tags on each of these morphemes in KTC4 provide us with information to annotate each of the syntactic units with LFG functional equations for these grammatical features.

Along with the grammatical features, the morphological environment of a given syntactic unit is also taken as the value of a certain attribute in this study. For example, the inflection forms of adjectives, verbs, verbal suffixes, auxiliaries and copulas are included into the attribute-value pairs of f-structure representations. Currently, the sequence of parts of speech are also included. These features will be useful for automatic generation of sentences from f-structure representations.

6.3.3 Annotation of Grammatical Function

The grammatical function of a given syntactic unit is determined according to the morphological information which is provided as KTC4/KNP tags. In order to make the grammatical function assignment process easy to maintain, I divide syntactic units into the following categories:

- Particled inflective units: units which have at least one particle, and which have an inflecting element (verb, adjective, verbal suffix or copula) as their head.
- Particled non-inflective units: units which have at least one particle, and which do not have an inflecting element as their head.
- Non-particled, inflective units: units which have no particle, and which have an inflecting element as their head.
- Non-particled, non-inflective units: units which have no particle, and which do not have an inflecting element as their head.

Every syntactic unit belongs to one of these categories. The units in one category are further divided into subcategories according to the type of the particle they have, or according to the inflection form of the head inflecting element. To each of the subcategories, a particular grammatical function is assigned. The process is summarised as follows. The process proceeds automatically and is implemented with much use of regular expressions:

For each syntactic unit,

Step 1. Check the root.

If the current syntactic unit is the root of the sentence, go to Step 9.

Else, go to Step 2.

Step 2. Check the presence of particles.

If the current syntactic unit has a particle, go to Step 3.

Else, go to Step 4.

Step 3. (for particled units) Check the presence of inflecting elements.

If this syntactic unit has at least one inflecting element, go to Step 5.

Else, go to Step 6.

Step 4. (for non-particled units) Check the presence of inflecting elements.

If this syntactic unit has at least one inflecting element, go to Step 7.

Else, go to Step 8.

Step 5. (for particled, inflective units) Check the particle.

If this unit has the case particle "-to", then the grammatical function of this unit is COMP.

Else, if this unit has a formal noun and either of the case particles "-ga", "-wo", "-ni", then this unit has the grammatical function SUBJ, OBJ, or OBL, respectively.

Else, if this unit has a formal noun and a case particle other than "-ga", "-wo" and "-ni", then this unit is has the grammatical function PADJ. (Complement clause in this unit will be treated later).

Else, if this unit has a particle other than case particle, then this unit has the grammatical function SADJ.

Go to Step 9.

Step 6. (for particled, non-inflective units) Check the particle.

If the dependency type of this unit is coordination, then the grammatical function is COORD.

Else, if the dependency type of this unit is apposition, then the grammatical function is APP.

Else, if the particle is "-ga", then the grammatical function is SUBJ.

Else, if the particle is "-wo", then the grammatical function is OBJ.

Else, if the particle is "-ni", then the grammatical function is OBL.

Else, if the particle is a case particle other than "-ga", "-wo" and "-ni", then the grammatical function is PADJ.

Else, if the particle is "-wa", then the grammatical function is TOPIC.

Else, if the particle is an adverbial particle other than "-wa", then the grammatical function is FOCUS.

Else, the grammatical function is PADJ.

Go to Step 9.

Step 7. (Non-Particled, Inflective)

If the inflection form of the last morpheme of this syntactic unit is the declarative base form, or the declarative ta form, then the grammatical function is REL.

Else, the grammatical function is SADJ.

Go to Step 9.

Step 8. (Non-Particled, Non-Inflective)

If the dependency type of this unit is coordination, then the grammatical function is COORD.

Else, if the dependency type of this unit is apposition, then the grammatical function is APP.

Else, if the part of speech of the head of this unit is determiner, then the grammatical function is DET.

Else, the grammatical function is ADJ.

Go to Step 9.

Step 9. (Root or not)

If the current unit is the root of the sentence, then stop.

Else, move to the next unit and go to Step 1.

6.3.4 A Worked Example

Let us see how the functional annotation process proceeds for a given sentence. Example (6.1) is a sentence from KTC4, and (6.2) is its representation in KTC4 annotated with POS tags but without case relations (both of which have been shown in the previous chapter as (5.24) and (5.25)):

(6.1) Somariano shutomogadisiode enjokatsudowo tsudukeru
kokurenkikanya hiseifusosshikiwa juusannichi subeteno katsudowo
chuushishita.

[Somaria-no]$_{f0}$ *[shuto-mogadishio-de]*$_{f1}$ *[enjo-katsudo-wo]*$_{f2}$
Somalia-PADJ capital-Mogadish-PADJ aid-activity-OBJ
[tsuzuke-ru]$_{f3}$ *[kokuren-kikan-ya]*$_{f4}$
continue-decl.base UN-organization-and
[hi-seifu-soshiki-wa]$_{f5}$ *[juusan-nichi]*$_{f6}$ *[subete-no]*$_{f7}$
non-government-organization-TOP thirteen-day all-PADJ
[katsudo-wo]$_{f8}$ *[chuushi-shita]*$_{f9}$
activity-OBJ stop-do.decl.ta
"The UN organisations and NGOs working for aid in Mogadishu, the capital of Somalia, stopped all their activities on 13th (of January, 1995)".

(6.2)

```
# S-ID:950115048-002 KNP:99/07/26
* 0 1D
ソマリア Somalia * noun place * *
の no * particle conjunctive * *
* 1 3D
首都 shuto * noun normal * *
モガディシオ Mogadish * noun place * *
で de * particle case * *
* 2 3D
援助 enjo * noun sahen * *
活動 katsudo * noun sahen * *
を wo * particle case * *
* 3 5D
続ける tsudukeru * verb * vowel base
* 4 5P
国連 kokuren * noun organization * *
機関 kikan * noun normal * *
や ya * particle conjunctive * *
* 5 9D
非 hi * prefix na-adjectival * *
政府 seifu * noun normal * *
組織 soshiki * noun sahen * *
は wa * particle adverbial * *
*6 9D
十三 juusan * noun numeral * *
日 nichi * suffix nominal-cardinal * *
、 , * special comma * *
* 7 8D
全て subete * adverb * * *
の no * particle conjunctive * *
* 8 9D
活動 katsudou * noun sahen * *
を wo * particle case * *
* 9 -1D
停止 teishi * noun sahen * *
した shita suru verb * sahen ta
。 . * special period * *
EOS
```

Each bunsetsu unit corresponds to a fragment f-structure in the f-structure for the whole sentence.

The f-structure functional equations are divided into three types: equations for the value of PRED attribute (semantic form), grammatical functions, and grammatical features.

Let us see how each of the ten units in (6.1) is annotated with functional equations.

0^{th} unit:

Line "* 0 1D" in (6.2) specifies that this unit is the 0th unit of this sentence and this unit directly depends on the 1st unit.

PRED: The content word in this unit is the noun "Somalia", hence we generate a functional equation for the PRED value of this unit as follows:

(6.3)

F0:PRED===‘Somalia’

Grammatical function: This unit has a particle and no inflecting element, and the particle is not a particle including a SUBJ, OBJ, OBL, TOPIC

or FOCUS function. Therefore the grammatical function of this unit is PADJ. We generate the functional equation for the grammatical function of this unit as follows:

(6.4)
F0 elm F1:PADJ

The operator "x elm y" states that x is an element of a set y.

Grammatical features: The second line in this unit in (6.2) specifies that this unit has a noun "Somalia", which is a place name. The third line in this unit specifies that this unit has a conjunctive particle "-no'. We generate the functional equations for the grammatical features of this unit as follows[2]:

(6.5)
F0:SUBPOS===‘place|prtcnj’,
F0:PRTCNJ===‘-no’

1^{st} unit:

The line "* 1 3D" in (6.2) specifies that this unit is the 1st unit of the sentence and that this unit directly depends on the 3rd unit.

PRED: This unit has two content words "shuto (capital city)" and "Mogadish", and they constitute a compound noun. They are therefore taken to be a single value of the PRED attribute for the f-structure corresponding to this unit. The functional equation for PRED is

(6.6)
F1:PRED===‘shuto|Mogadish; the capital Mogadish’

Grammatical function: This unit has a particle and no inflecting element, and the particle is not for SUBJ, OBJ, OBL, TOPIC or FOCUS. Therefore the grammatical function of this unit is PADJ. We generate the functional equation for the grammatical function of this unit as follows:

(6.7)
F1 elm F3:PADJ

Grammatical features: The part-of-speech tags specify that "Shuto" is a normal noun,"Mogadish" is a place name, and the particle "-de" is a

[2]The vertical line between ‘place’ and ‘prtcnj’ means that they are concatenated.

case particle. We generate the functional equations for the grammatical features of this unit as follows:

(6.8)
F1:SUBPOS===‘norm|place|prtcs’,
F1:PRTCS===‘-de’

2^{nd} unit:

Line “* 2 3D” in (6.2) specifies that this unit is the 2nd unit of the sentence and that this unit directly depends on the 3rd unit.

PRED: This unit has two content words “enjo (help)” and “katsudo (activity)”, and together they constitute a compound noun. They are taken to be a single value of the PRED attribute for the f-structure corresponding to this unit. The functional equation for PRED is as follows:

(6.9)
F2:PRED===‘enjo|katsudo; aid activity’

Grammatical function: This unit has a particle and no inflecting element, and the particle is “-wo”, which is for OBJ. We generate the functional equation for the grammatical function of this unit as follows:

(6.10)
F3:OBJ===F2

Grammatical features: The part-of-speech tags specify that “enjo” and “katsudo” are sahen nouns, and that “-wo” is a case particle. We generate the functional equations for the grammatical features of this unit as follows:

(6.11)
F2:SUBPOS===‘sahen|sahen|prtcs’,
F2:PRTCS===‘-wo’

3^{rd} unit:

Line “* 3 5D” in (6.2) specifies that this unit is the 3rd unit of the sentence and that it directly depends on the 5th unit.

PRED: The content word in this unit is “tsudzukeru (continue)”. This verb is a transitive verb. The functional equation for PRED is as follows:

(6.12)
F3:PRED===‘tsuzukeru<subj, obj>; continue’

Grammatical function: This unit has an inflecting element with no particle, and the inflection form is the declarative base form, and the unit is analysed as a relative clause modifier of the 5th unit:

(6.13)
F3 elm F5:REL

Grammatical features: The part-of-speech of “tsudukeru” is “Type-II verb” and its inflection form is the base form. There is no other morpheme, so it is not necessary to specify a sequence of morphemes. The tense is present, and the mood is declarative. The functional equations for this unit are as follows:

(6.14)
F3:INFL===‘base’
F3:TENSE===‘present’
F3:MOOD===‘declarative’

4^{th} unit:

Line “* 4 5P” in (6.2) specifies that this unit is the 4th unit of the sentence and it depends on the 5th unit as a coordinate.

PRED: This unit has two content words “kokuren (abbreviation for the United Nations)” and “kikan (organisation)”, and they constitute a compound noun. They are taken to be a single value of the PRED attribute for the f-structure corresponding to this unit. The functional equation for PRED is as follows:

(6.15)
F4:PRED===‘kokuren|kikan; UN organisations’

Grammatical function: The dependency type of this unit is coordination, and the grammatical function is COORD.

(6.16)
F4 elm F5:COORD

Grammatical features: The part-of-speech tags specify that "kokuren" and "kikan" are normal nouns, and that "-ya" is a conjunctive particle. We generate the functional equations for the grammatical features of this unit as follows:

(6.17)
F4:SUBPOS===‘norm|norm|prtcnj’,
F4:PRTCNJ===‘-ya’

5^{th} unit:

Line "* 5 9D" in (6.2) specifies that this unit is the 5th unit of this sentence and that it directly depends on the 9th unit of this sentence.

PRED: This unit has three content words "hi (non-)", "seifu (government)" and "soshiki (organization)", and they constitute a compound noun. They are taken to be a single value of the PRED attribute for the f-structure corresponding to this unit. The functional equation for PRED is as follows:

(6.18)
F5:PRED===‘hi|seifu|soshiki; non-government organisations’

Grammatical function: This unit has a particle and no inflecting element, and the particle is "-wa". The grammatical function of this unit is TOPIC:

(6.19)
F5 elm F9:TOPIC

Grammatical features: The part-of-speech tags specify that "hi-" is an adjectival prefix, "seifu" is a normal noun, "soshiki" is a sahen noun, and "-wa" is an adverbial particle. The prefix is "hi-", and the adverbial particle is "-wa".

(6.20)
F5:SUBPOS===‘adjectivalprefix|normal|sahen|prtadv’
F5:PRF===‘hi’
F5:PRTADV===‘-wa’

6^{th} unit:

Line "* 6 9D" in (6.2) specifies that this unit is the 6th unit of this sentence and that it directly depends on the 9th unit of this sentence.

PRED: This unit has two content words 'juusan (thirteen)' and 'nichi (day)', and they constitute a compound noun "the 13th day of a month". The month is not specified explicitly, since it is considered to be known by the reader. The functional equation for PRED is as follows:

(6.21)
F6:PRED===‘juusannichi; the 13th day’

Grammatical function This unit has no particle and no inflecting part-of-speech, hence its grammatical function is ADJ.

(6.22)
F6 elm F9:ADJ

Grammatical features The part-of-speech tags specify that 'juusan' is a numeral noun and 'nichi' is a nominal-cardinal suffix, which is a type of unit classifier for dates (cf. Section 3.6.1). This unit has a comma:

(6.23)
F6:SUBPOS===‘numeral|nominal-cardinal|comma’
F6:NUMERAL===‘juusan; thirteen’
F6:SUF===‘nichi;date’
F6:COMMA===‘,’

7^{th} unit:

Line "* 7 8D" in (6.2) specifies that this unit is the 7th unit of this sentence and that it directly depends on the 8th unit of this sentence.

PRED: The content word in this unit is 'subete (all)':

(6.24)
F7:PRED===‘subete; all’

Grammatical function: This unit has no inflecting part-of-speech and one particle, and this particle is not for SUBJ, OBJ, OBL, TOPIC or FOCUS. The grammatical function of this unit is PADJ:

(6.25)
F7 elm F9:PADJ

Grammatical features: 'Subete' is an adverb and '-no' is a conjunctive particle. Adverbs are not divided into sub parts of speech, hence "*":

(6.26)
F7:SUBPOS===‘*|prtcnj’
F7:PRTCNJ=‘-no’

8^{th} unit:

Line "* 8 9D" in (6.2) specifies that this unit is the 8th unit of this sentence and that it directly depends on the 9th unit of this sentence.

PRED: The content word is 'katsudou (activity)':

(6.27)
F8:PRED===‘katsudou; activity’

Grammatical function: This unit has the case particle '-wo' and does not have inflectional part-of-speech, hence the grammatical function is OBJ:

(6.28)
F9:OBJ===F8

Grammatical features: The part-of-speech of 'katsudou(activity)' is 'sa-hen noun' and '-wo' is a case particle:

(6.29)
F8:SUBPOS===‘sahen|prtcs’
F8:PRTCS===‘-wo’

9^{th} unit:

Line "* 9 -1D" in (6.2) specifies that this unit is the 9th unit of this sentence and that it depends on no other unit, hence it is the root unit of this sentence.

PRED: The content word is 'teishi' and 'shita'. The lemma form of 'shita' is 'suru'.

(6.30)
F9:PRED===‘teishisuru<subj, obj>; stop’

Grammatical function: This unit is the root unit of this sentence. Hence it does not have grammatical function.

Grammatical features: The part of speech of 'teishi' is 'sahen noun'. 'Suru' is a verb, and verbs are not divided into sub parts of speech. This unit has a period. The inflection form of 'suru' in this unit is the declarative -ta form:

(6.31)

F9:SUBPOS===‘sahen|*|period’
F9:INFL===‘decl.ta’
F9:PERIOD===‘.’
F9:TENSE===‘past’
F9:MOOD===‘declarative’

We have now produced the complete set of functional equations to compute the labelled dependency among syntactic units. The functional equations represent the dependency among the syntactic units as a DAG:

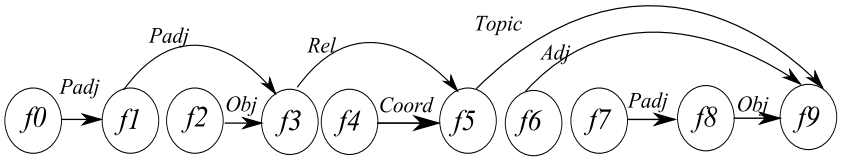

Figure 6.2: DAG representation for the example sentence (6.1)

The corresponding f-structure (Figure 6.3) adds the PRED grammatical feature specifications to the main dependency representation:

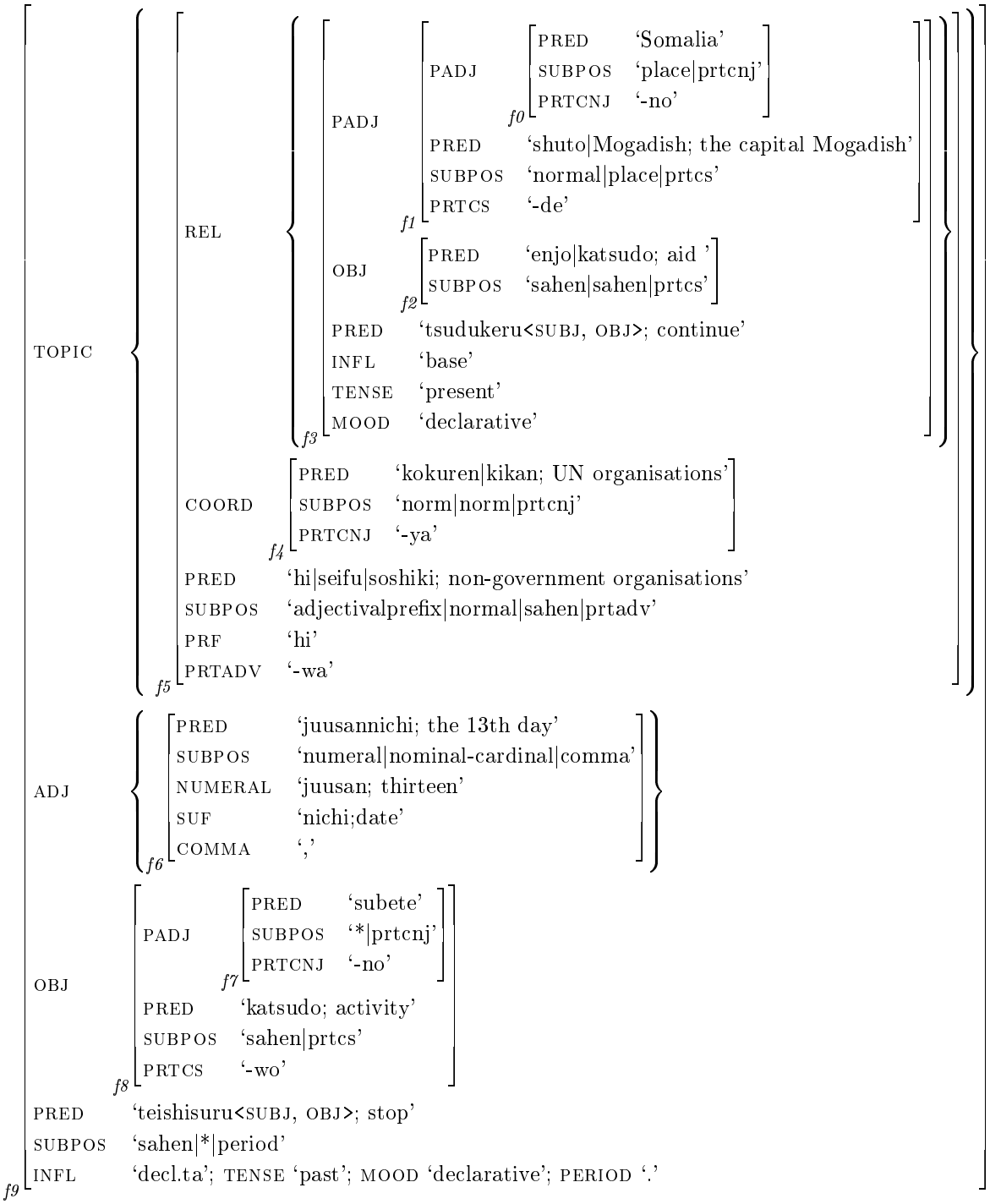

Figure 6.3: F-structure for the example sentence (6.1)

Notice that this f-structure is not complete; the grammatical function SUBJ is subcategorised for by the verbal predicate 'teishisuru' in the relative clause REL function, but it is absent in this f-structure; hence this f-structure is incomplete. For this f-structure to be complete, the SUBJ zero pronoun for the predicate 'teishisuru' must be added into this f-structure. This issue is the topic of Section 6.6.

6.4 Coordination

Section 4.2.12 points out that the dependency representation in KTC4 and the dependency parse output by KNP do not treat coordination properly; that is, in KTC4/KNP representations, one coordinate phrase depends on another coordinate phrase (see $f4$ and $f5$ in Figures 6.2 and 6.3). This leads to an incorrect structure in which one coordinate is embedded into another coordinate. In order to avoid generating this type of structure for coordination, we need to create a dummy unit which has the coordinates as its elements for the set-valued COORD-feature.

6.4.1 Coordination Fixing

The following description summarises the process of creating dummy units for coordination in terms of LFG functional equations:

For each sentence,

Step 1. Search coordinates in the equations for a sentence.

If there are coordinates, then go to Step 2.

If there is not, then go to Step 10.

Step 2. Initialise the dummy unit number d. Go to Step 3.

Step 3. Put the equations into an array; each cell contains one equation. Go to Step 4.

Step 4. Search an equation "Fx elm Fy:COORD" (y is not equal to d) in the array.

If there is, go to Step 5.

Else, go to Step 9.

Step 5. Search an equation "Fz:GF===Fy" or "Fy elm Fz:GF". (Search the last coordinate which has a GF function)

If there is one, go to Step 6.

Else, go to Step 7.

Step 6. Rewrite and add functional equations.

Rewrite the equation "Fz:GF===Fy" into "Fz:GF===Fd", or "Fy elm Fz:GF" into "Fd elm Fz:GF".

Add a new equation "Fy elm Fd:NCOORD" into the array.

Rewrite the equation "Fx elm Fy:COORD" into "Fx elm Fd:COORD".

Increment d with one.

Go to Step 4, after the cell in which the equation "Fx elm Fy:COORD (now Fx elm Fd:NCOORD)" is stored.

Step 7. Search an equation "Fy elm Fz:COORD". (Search a coordinate depended on by another coordinate)

When there is one, go to Step 8.

Else, go to Step 4, starting from the cell after the cell in which the equation "Fx elm Fy:COORD" is stored.

Step 8. Rewrite "Fy elm Fz:COORD" into "Fy elm Fd:COORD", and go to Step 4, after the cell in which the equation "Fx elm Fy:COORD" is stored.

Step 9. Rewrite "NCOORD" into "COORD". Go to Step 10.

Step 10. Stop.

The dummy unit number must be initially set more than the number of the units in the input sentence. In this case, it is set 10. The interim function name "NCOORD" is required to prevent an infinite loop.

Sentence (6.1) contains functional equations "F4 elm F5:COORD" and "F5 elm F9:TOPIC". The process finds "F4 elm F5:COORD" at Step 4, and "F5 elm F9:TOPIC" at Step 5. These functional equations are rewritten, and a new functional equation is added at Step 6 as follows:

1 "F4 elm F5:COORD" → "F4 elm F10:COORD": the first coordinate depends on the dummy unit.

2 "F5 elm F9:TOPIC" $\rightarrow$ "F10 elm F9:TOPIC": the dummy unit is the TOPIC of F9.

3 "F5 elm F10:NCOORD"(a new equation): the second coordinate depends on the dummy unit.

The process goes back to Step 4, but this time there is no more "Fx elm Fy:COORD", therefore the process goes to Step 9 and "NCOORD" is written into "COORD", then stops.

The DAG representation for example sentence (6.1) after coordination fixing is given in Figure 6.4. Compared with the DAG before coordination fixing (Figure 6.2), this DAG has a dummy vertex $f10$ which has two coordinate unit dependents.

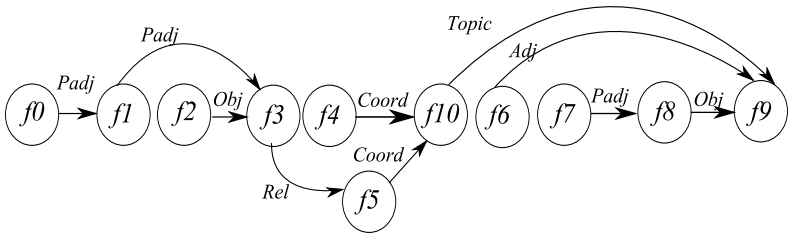

Figure 6.4: DAG representation for sentence (6.1) after coordination fixing

The f-structure for Figure 6.4 is Figure 6.5; the relative clause modifies only one of the two coordinates:

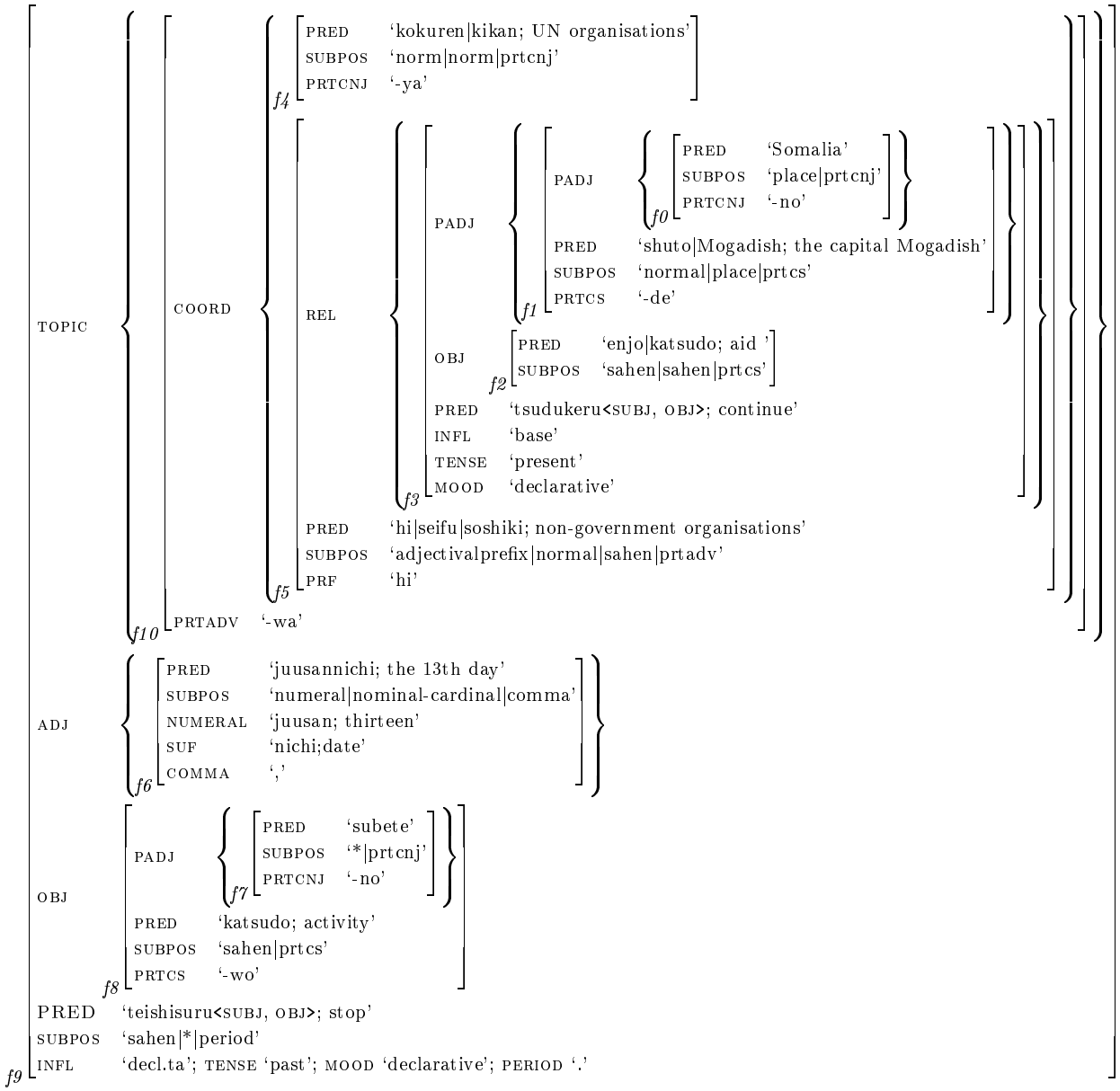

Figure 6.5: F-structure for the example sentence (6.1) after fixing coordination

6.4.2 Adjunct Modification of Coordinates

If one coordinate is modified by an adjunct, there are two possibilities in terms of what this phrase modifies. One possibility is that it modifies both of the coordinate phrases, and the other is that it modifies the coordinate

on which it directly depends. For example, line "* 3 5D" in the KTC4 representation in (6.2) for sentence (6.1) specifies that the 3^{rd} unit depends on the 5^{th} unit. Semantically, however, the relative clause modifies both of the coordinates, hence the 3^{rd} unit must depend on the 10^{th} and modify both of the coordinate units, as in Figure 6.6:

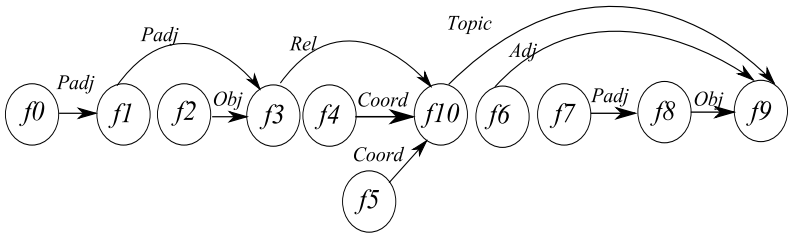

Figure 6.6: DAG representation for sentence (6.1) after fixing coordination and adjunct modification

Therefore, if a coordinate is depended on by adjuncts, their dependencies must be fixed so that they depend on the dummy syntactic unit for coordinates, as shown in e.g. Figure 6.6.

The f-structure for sentence (6.1) after coordination fixing is given in Figure 6.7:

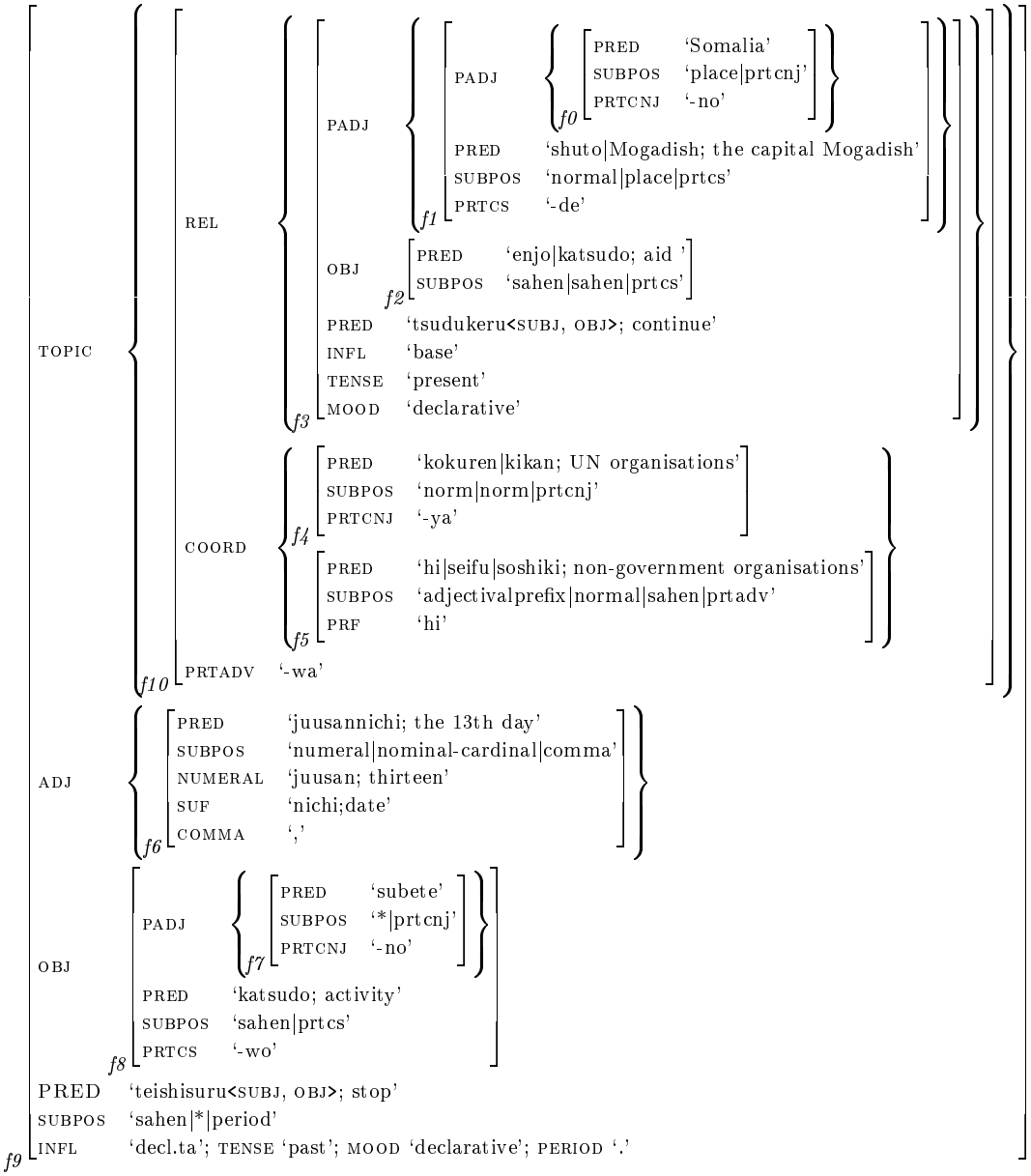

Figure 6.7: F-structure for the example sentence (6.1) after fixing coordination and adjunct modification

It is difficult to automatically determine whether one adjunct modifies both (or all) of the coordinate phrases, or it modifies the coordinate on which it directly depends. In the coordination-fixing module of this study, I assume that an adjunct depends on the dummy unit, as shown in Figure 6.7. This assumption leads to an inclusive analysis when one adjunct only

modifies the coordinate on which it directly depends. For example, consider sentence (6.32):

(6.32) Akai honto jishoga teeburuno ueni aru.

Aka-i$_{f0}$	*hon-to*$_{f1}$	*jisho-ga*$_{f2}$	*teeburu-no*$_{f3}$
red-decl.base	book-COORD	dictionary-SUBJ	table-of
ue-ni$_{f4}$	*ar-u*$_{f5}$		
above-OBL	be-decl.base		

"A red book and dictionary are on the table." or "A red book and a dictionary are on the table"

Sentence (6.32) has two different meanings; one is that there are a book and a dictionary on the table and both of them are red, and the other is that there are a red book and a dictionary on the table. The second meaning sounds more natural than the first because the adjective "akai" directly depends on "hon", but we cannot exclude the possibility that the first meaning better describes the situation. Since this problem of ambiguous modification for coordinates can be solved by taking into consideration the context in which a given sentence is used, I leave it to future study.

6.5 "Catch-All" and "Clean-Up"

There are a variety of instances in which the annotation process described above fails to properly capture the morphological characteristics of a given unit, and annotates an incorrect grammatical function. In particular, there are instances in KTC4/KNP output in which one unit has more than one head. Such units fail to be assigned proper grammatical functions by an automatic method which assumes that one unit has only one head. Therefore, these instances must be detected and treated properly. This section describes some of these instances and how they are treated.

6.5.1 COMP-Taking Formal Nouns

"No" as a formal noun takes a COMP (cf. Section 3.6.1):

(6.33) Kenga Naomiga rikonshita*no*wo shiranakatta.

Ken-ga$_{f0}$	*[Naomi-ga*$_{f1}$	*rikonshi-ta]-no-wo*$_{f2}$
Ken-SUBJ	[Naomi-SUBJ	divorce-decl.ta]-that-OBJ
shir-ana-katta$_{f3}$		
know-NEG-decl.ta		

"Ken didn't know that Naomi had divorced."

The incorrect DAG representation for (6.33) is shown below:

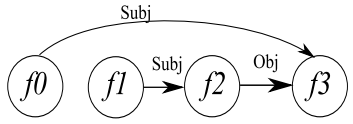

Figure 6.8: DAG for (6.33)

Figure 6.9 is the f-structure representation for (6.33) from the automatic annotation method (on KTC4/KNP representations):

$$
\begin{bmatrix}
\text{SUBJ} & _{f0}\begin{bmatrix} \text{PRED} & \text{`Ken'} \\ \text{PRTADV} & \text{`-ga'} \end{bmatrix} \\
\text{OBJ} & _{f2}\begin{bmatrix} \text{SUBJ} & _{f1}\begin{bmatrix} \text{PRED} & \text{`Naomi'} \\ \text{PRTCS} & \text{`-ga'} \end{bmatrix} \\ \text{PRED} & \text{`rikonshita}\langle \text{SUBJ} \rangle\text{-no'} \\ \text{PRTCS} & \text{`-wo'} \end{bmatrix} \\
\text{PRED} & \text{`shir-}\langle \text{SUBJ, OBJ} \rangle\text{'} \\
\text{TENSE} & \text{`past'} \\
_{f3}\text{NEG} & \text{`+'}
\end{bmatrix}
$$

Figure 6.9: F-structure for (6.33)

The f-structure above does not represent the grammatical function COMP of the formal noun 'no'. This results from an incorrect functional equation "(f3 PRED === 'rikonshita<SUBJ>-no')". Since syntactic unit $f3$ has two content words (the verb 'rikonshita' and the formal noun 'no'), the automatic annotation method concatenates the PRED value of the verb and that of the formal noun, hence it outputs an incorrect functional equation.

This incorrect annotation must be fixed automatically, as soon as it is detected in the KTC4 representation or KNP parsed output. There are three things to be done: the PRED form for the verb must be fixed, the functional equations for the PRED value of 'no' must be added, and a new unit for the formal noun is added to the original representation, so that the verb depends on this new unit with the grammatical function COMP. The

DAG representation after automatic fixing is shown in Figure 6.10; the unit $f4$ is the new unit which is added for the formal noun:

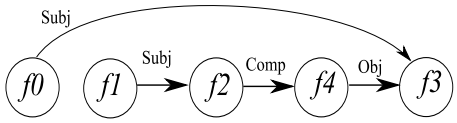

Figure 6.10: DAG for (6.33) after automatic fixing

Figure 6.11 below is the f-structure for the sentence after automatic fixing:

$$\begin{bmatrix}
\text{SUBJ} & f0\begin{bmatrix}\text{PRED} & \text{`Ken'}\\ \text{PRTADV} & \text{`-ga'}\end{bmatrix} \\
\text{OBJ} & \begin{bmatrix}
\text{COMP} & \begin{bmatrix}\text{SUBJ} & f1\begin{bmatrix}\text{PRED} & \text{`Naomi'}\\ \text{PRTCS} & \text{`-ga'}\end{bmatrix}\\ & f2\begin{bmatrix}\text{PRED} & \text{`rikonshita}\langle \text{SUBJ}\rangle\text{'}\end{bmatrix}\end{bmatrix}\\
\text{PRED} & \text{`no'}\\
\text{PRTCS} & \text{`-wo'}
\end{bmatrix} f4\\
\text{PRED} & \text{`shir-}\langle \text{SUBJ, OBJ}\rangle\text{'}\\
\text{TENSE} & \text{`past'}\\
\text{NEG} & \text{`+'}
\end{bmatrix} f3$$

Figure 6.11: F-structure for (6.33) after fixing the equation

6.5.2 Formal Nouns and Adverbs Taking an Appositional COMP

A verbal unit ending with an inflecting morpheme in its base form or the ta form and depending on a noun receives the grammatical function REL (cf. Section 6.3.3). The problem is that this fails to distinguish a relative clause from an appositional clause. As pointed out in Section 4.2.10, the distinction between REL and appositional COMP depends on zero-pronoun resolution for a missing argument of the verbal head of the clause dependent on a noun.

Fortunately, there are some nouns and adverbs which often take an appositional complement and do not function as the argument of the verb in the complement clause. One such instance is the formal noun 'koto' (meaning 'a

thing') and a set of other formal nouns which constitute adverbial units (cf. Section 3.6.1). If it is found that a clause is dependent on such a noun or adverb, and the grammatical function REL is given to this clause, then this assignment is rewritten into COMP. In order to make a distinction between COMPs of this kind and COMPs with the case particle "-to", the former are called APCOMP (APpositional COMPlement) in this study.

6.6 Zero-Pronoun Identification

6.6.1 LFG Representation of Zero Pronouns

Japanese is a language with a high prevalence of zero pronouns, unrealised arguments which can be inferred from context. Zero-pronoun identification is required for an f-structure to satisfy the well-formedness conditions, viz., Completeness and Coherence. Consider the following sentence (6.34) in which the OBJ of the verb is a zero pronoun which is coindexed with the TOPIC:

(6.34) Kono madowa watashiga ake-ta.

$[kono]_{f1}$	$[mado\text{-}wa]_{f2}$	$[watashi\text{-}ga]_{f3}$	$[pro]_{f4}$	$[ake\text{-}ta]_{f5}$
this	window-TOP	I-SUBJ	pro	open-decl.ta

"As for this window, I opened it."

An f-structure satisfies the Completeness condition iff it contains values for all grammatical functions that are subcategorised for by its predicates (cf. Section 2.3.3). If we ignore zero pronouns, then the f-structure in Figure 6.12 for (6.34) will violate the Completeness condition. The verb "ake-<SUBJ, OBJ>" in (6.34) subcategorises for an OBJ, but the f-structure does not contain an OBJ attribute, hence this f-structure is not well-formed:

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{DET} & {}_{f1}\begin{bmatrix}\text{PRED} & \text{`kono;this'}\end{bmatrix} \\ \text{PRED} & \text{`mado;window'} \\ \text{PRTADV} & \text{`wa'} \end{bmatrix}_{f2} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix}\text{PRED} & \text{`watashi;I'} \\ \text{PRTCS} & \text{`-ga'}\end{bmatrix} \\
\text{PRED} & \text{`ake-}\langle \text{SUBJ, OBJ} \rangle\text{'; open} \\
\text{TENSE} & \text{-ta} \\
\text{MOOD} & \text{declarative}
\end{bmatrix}_{f5}
$$

Figure 6.12: An incomplete f-structure for (6.34)

The complete f-structure for (6.34) is shown in Figure 6.13:

$$
\begin{bmatrix}
\text{TOPIC} & \left\{ \begin{bmatrix} \text{DET} & {}_{f1}\begin{bmatrix}\text{PRED} & \text{`kono;this'}\end{bmatrix} \\ \text{PRED} & \text{`mado;window'} \\ \text{PRTADV} & \text{`wa'} \\ \text{INDEX} & \text{i} \end{bmatrix}_{f2} \right\} \\
\text{SUBJ} & {}_{f3}\begin{bmatrix}\text{PRED} & \text{`watashi;I'} \\ \text{PRTCS} & \text{`-ga'}\end{bmatrix} \\
\text{OBJ} & {}_{f4}\begin{bmatrix}\text{PRED} & \text{`pro'} \\ \text{INDEX} & \text{i}\end{bmatrix} \\
\text{PRED} & \text{`ake-}\langle \text{SUBJ, OBJ} \rangle\text{'; open} \\
\text{TENSE} & \text{-ta} \\
\text{MOOD} & \text{declarative}
\end{bmatrix}_{f5}
$$

Figure 6.13: The complete f-structure for (6.34)

F-structure in Figure 6.13 represents an unrealised OBJ with PRED 'pro', whose referential index is coindexed with the referential index of the TOPIC function. The zero pronoun in 6.13 is coindexed with something which appears somewhere in the local context, but this is not the case in sentence (6.35), in which the OBJ zero pronoun refers to something beyond the sentence boundary:

(6.35) Watashiga ake-ta.

[watashi-ga]$_{f1}$	*[pro]$_{f2}$*	*[ake-ta]$_{f3}$*
I-SUBJ	pro	open-decl.ta

"I opened it."

The f-structure for (6.35) does not include anything coindexed with the OBJ zero pronoun:

$$\begin{bmatrix} \text{SUBJ} & {}_{f1}\begin{bmatrix} \text{PRED} & \text{'watashi;I'} \\ \text{PRTCS} & \text{'-ga'} \end{bmatrix} \\ \text{OBJ} & {}_{f2}\begin{bmatrix} \text{PRED} & \text{'pro'} \\ \text{INDEX} & \text{i} \end{bmatrix} \\ \text{PRED} & \text{'ake-}\langle \text{SUBJ, OBJ} \rangle\text{'; open} \\ \text{TENSE} & \text{-ta} \\ {}_{f3}\text{MOOD} & \text{declarative} \end{bmatrix}$$

Figure 6.14: The complete f-structure for (6.35)

Next, consider the following sentence (6.36) in which the verb "ak-" is an intransitive verb:

(6.36) Kono madoga kazede aita.

kono	*mado-ga*	*kaze-de*	*ak-ta*
this	window-SUBJ	wind-by	open-decl.ta

"This window opened by the wind."

The Coherence condition requires that an f-structure should be non-redundant; for example, core grammatical functions (SUBJ, OBJ, OBJ2, OBL2 for English) must be subcategorised for by the predicate (i.e., the predicate requires these functions) in the f-structure for a clause (cf. Section 2.3.2). The f-structure for (6.36) is shown in Figure 6.15 below:

$$
f5\left[\begin{array}{ll}
\text{SUBJ} & f2\left[\begin{array}{ll}
\text{DET} & f1\left[\text{PRED} \quad \text{`kono;this'}\right] \\
\text{PRED} & \text{`mado;window'} \\
\text{PRTADV} & \text{`ga'}
\end{array}\right] \\
\text{PADJ} & \left\{ f3\left[\begin{array}{ll}
\text{PRED} & \text{`kaze; wind'} \\
\text{PRTCS} & \text{`-de'}
\end{array}\right] \right\} \\
\text{PRED} & \text{`ak-}\langle \text{SUBJ} \rangle\text{'; open} \\
\text{TENSE} & \text{-ta} \\
\text{MOOD} & \text{declarative}
\end{array}\right]
$$

Figure 6.15: The coherent f-structure for (6.36)

If we incorrectly assume the presence of a zero pronoun (Figure 6.16), then the resulting f-structure will violate the Coherence condition:

$$
f5\left[\begin{array}{ll}
\text{SUBJ} & f2\left[\begin{array}{ll}
\text{DET} & f1\left[\text{PRED} \quad \text{`kono;this'}\right] \\
\text{PRED} & \text{`mado;door'} \\
\text{PRTADV} & \text{`ga'}
\end{array}\right] \\
\text{PADJ} & \left\{ f3\left[\begin{array}{ll}
\text{PRED} & \text{`kaze; wind'} \\
\text{PRTCS} & \text{`-de'}
\end{array}\right] \right\} \\
\text{OBJ} & \left[\text{PRED} \quad \text{`pro'}\right] \\
\text{PRED} & \text{`ak-}\langle \text{SUBJ} \rangle\text{'; open} \\
\text{TENSE} & \text{-ta} \\
\text{MOOD} & \text{declarative}
\end{array}\right]
$$

Figure 6.16: An incoherent f-structure for (6.36)

Therefore, zero pronouns must be properly identified in order to acquire a complete and coherent f-structure for each sentence in the input text. The key to do this is to have access to subcategorisation frames of verbs, in other words, information about which grammatical function a given verb subcategorises for.

6.6.2 Zero-Pronoun Identification for Japanese

Zero-pronoun identification for Japanese will be processed in the following steps:

For each verb in the input sentence,

Step 1. Check whether its subject is overtly expressed.
If it is, go to Step 2.
If it is not, then this verb's subject is a zero pronoun. Put a subject zero pronoun dependent on this verb into the f-structure, then go to Step 2.

Step 2. Check whether this verb should take an object, using a list of object-taking verbs.
If this verb belongs to this list, then got to Step 3.
If it is not, go to Step 4.

Step 3. Check whether its direct object is overtly expressed.
If it is, go to Step 4.
If it is not, then this verb's object is a zero pronoun. Put an object zero pronoun dependent on this verb in the f-structure. Go to Step 4.

Step 4. Check whether this verb should take an oblique, using a list of oblique-taking verbs.
If this verb belongs to this list, then go to Step 5.
If not, go to Step 6.

Step 5. Check whether its oblique is overtly expressed.
If it is, go to Step 6.
If not, then this verb's oblique is a zero pronoun. Put an oblique zero pronoun dependent on this verb in the f-structure, then go to Step 6.

Step 6. Stop.

In the process described above, we need to have a list of object-taking verbs and a list of oblique-taking verbs. There are two approaches to construct these lists automatically for zero-pronoun identification: a morphology-based approach and a probabilistic approach.

6.6.3 A Morphology-Based Approach to Zero-Pronoun Identification

As described in Section 3.5.2, some verbs constitute Transitive-Intransitive pairs. The categories of transitive-intransitive pairs are shown again in Table 6.1 from [Yoshikawa, 1989] and can be identified based on word endings:

Table 6.1: Transitive-Intransitive Pairs

	Intransitive ending	Transitive ending	Examples	
1	-ARU	-U	husag*ar-u*	husag-*u*
2	-ARU	-ERU	ag*ar-u*	ag*e-ru*
3	-U	-ERU	ak-*u*	ak*e-ru*
4	-ERU	-U	tor*e-ru*	tor-*u*
5	-ERU	-ASU	nur*e-ru*	nur*as-u*
6	-RERU	-SU	tao*re-ru*	tao*s-u*
7	-U	-ASU	kawak-*u*	kawak*as-u*
8	-IRU	-ASU	nob*i-ru*	nob*as-u*
9	-IRU	-OSU	och*i-ru*	ot*os-u*
10	-RU	-SU	noko*r-u*	noko*s-u*
11	-RU	-SERU	no*r-u*	no*se-ru*
12	-IERU	-ESU	k*ie-ru*	k*es-u*

Transitive-intransitive pairs are automatically extracted from the content-word dictionary of JUMAN by a Ruby script. This script yields the verb pairs belonging to each of the twelve categories. As for the category U-ERU and category ERU-U, the automatic extraction does not distinguish between them. The verb pairs belonging to these categories are the same. Among them, some of ERU-ending verbs are transitive verbs with U-ending intransitive partners (belonging to the category U-ERU), while other ERU-ending verbs are intransitives with U-ending transitive partners (belonging to the category ERU-U). Therefore, this distinction must be made manually.

The number of verbs belonging to each category is shown in Table 6.2 below:

Table 6.2: Number of Transitive-Intransitive Pairs

	Intransitive ending	Transitive ending	Number
1	-ARU	-U	16
2	-ARU	-ERU	99
3	-U	-ERU	74
4	-ERU	-U	46
5	-ERU	-ASU	27
6	-RERU	-SU	21
7	-U	-ASU	33
8	-IRU	-ASU	7
9	-IRU	-OSU	4
10	-RU	-SU	21
11	-RU	-SERU	7
12	-IERU	-ESU	1
			Sum: 356

Among these pairs, the majority of verbs ending with "SU" paired with no other verbs are transitive verbs. Manually checking 628 "-SU" ending verbs in the JUMAN dictionary, I found that 609 are transitive. So far, we have a list of 965 transitive verbs, which consist of the paired verbs (356 types) and "su" transitive verbs (609 types). We also have a list of 356 intransitive verb types.

The list of transitive verbs can be used for zero-pronoun identification. If a verb appears without an overt object, then check whether this verb belongs to the list of transitive verbs. If it does, then it has a zero-pronominalised object in the context. If it does not, then this verb is intransitive and does not have zero-pronominalised object.

The main problem with the Morphological approach is its low coverage. The JUMAN dictionary lists a total of 4171 verbs, and the transitive verb list covers only 30% of them. The rest of them are non-paired verbs not ending with "su". Their valency must be determined somehow.

6.6.4 A Probabilistic Approach to Zero-Pronoun Identification

The Probabilistic approach is a measure to overcome the problem of low coverage in the Morphological approach. The probabilistic approach looks at the syntactic environment in which verbs appear. I estimate the probability that each of the verbs takes an overt object (a noun unit with the case

particle "-wo") as its argument. This probability is estimated by the number of occurrences of a verb with an overt object V'_o divided by the number of occurrences of this verb appearing in the training corpus V'_s. I call this probability "transitivity rate" Tr:

$$Tr = \frac{V'_o}{V'_s} \tag{6.37}$$

If the transitivity rate of a verb is near 100%, then this verb has almost always been used as a transitive verb with its overt object in the training corpus, hence it is highly likely that this verb is used as a transitive verb also in the unseen part of the corpus. If this verb is used without any overt object, then it is also highly likely that there is an object zero pronoun.

If, on the other hand, the transitivity rate of a verb is low, then it is unlikely to encounter this verb used as a transitive verb with its overt object in the unseen part of the corpus. In addition, it is unnatural to suppose that a verb is a transitive verb if the object is not overt. In particular, if the transitivity rate is zero, we do not have to worry too much about this verb.

Figure 6.17 below shows the transitivity rates of verbs in KTC4:

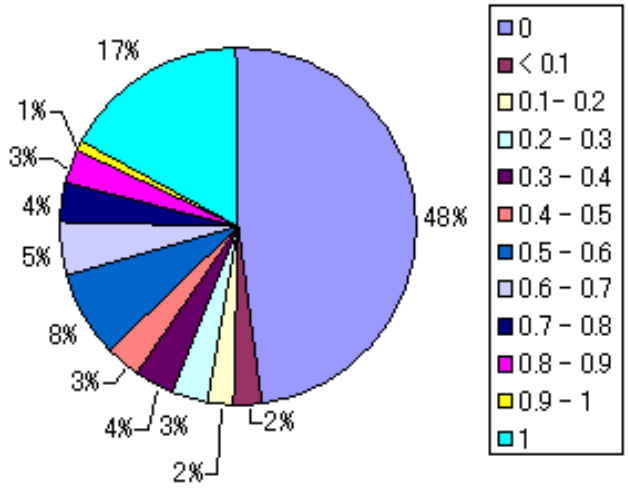

Figure 6.17: Transitivity Rates of Verbs in KTC4

As Figure 6.17 shows, use of transitivity rates in determining the valency of verbs can help solve the problem of low coverage encountered in the Morphological approach; now we know the verbs with zero transitivity rate (intransitive verbs) and the verbs with 100% transitivity rate (transitive verbs), and together they cover about 65% of all verbs.

The problem here is to determine the middle ground; in other words, how high should the transitivity rate be to conclude that a given verb is a transitive verb or an intransitive verb?

6.7 Summary

This chapter has introduced the LFG annotation method for KTC4 representations and KNP output. I use KTC4 as the corpus from which wide-coverage LFG resources are acquired. The method I introduced implements the idea that the part-of-speech tags on each morpheme and the unlabelled dependency tags on each syntactic unit in KTC4 provide us with enough information for constructing what [Cahill et al., 2003, Cahill et al., 2004, Cahill, 2004] call "proto" f-structures for texts in a corpus, without employing PCFG-style syntactic trees.

The method is to label each of the unlabelled dependency relations among syntactic units according to the case tags and other information given by KTC4/KNP, and to represent these labels and the morphological information in terms of LFG functional equations. By resolving these equations, we produce the f-structure representation for the input sentence.

The problems of this method (coordination, one-to-many correspondence between one unit and more than one f-structure, and zero-pronoun identification) are treated by some some additional operations in order to acquire improved LFG representations for the input. Some of them require special treatments (such as an externally supplied or computed subcategorisation list) since the information which is encoded explicitly by tags in KTC4/KNP is not enough to solve the particular problem. The evaluation of the f-structure representations acquired by the method is the topic of Chapter 7.

Chapter 7

Evaluation of the LFG Annotation for KTC4/KNP

7.1 Introduction

This chapter explains how the automatic annotation of LFG functional equations for KTC4/KNP is evaluated, presents and discusses the evaluation results, and outlines future research. First, the f-structures acquired by the automatic annotation of KTC4 representations are evaluated against Gold-Standard f-structures for 500 sentences randomly extracted from KTC4. These Gold-Standard f-structures are automatically annotated and then manually corrected. Different methods for zero-pronoun identification are compared in terms of precision, recall and f-scores for grammatical functions. It is shown that zero-pronoun identification using both morphological and probabilistic methods yields the best result among other methods. Second, the same 500 sentences (but without KTC4 annotations) are parsed with KNP, then the KNP output is automatically annotated with LFG functional equations, and the acquired f-structures are compared against the Gold Standard f-structures. The results are lower than those of KTC4, due to the noise introduced by the parser compared to the KTC4 treebank representations.

7.2 Procedure

The quality of the f-structures automatically acquired from KTC4 is evaluated against gold standard f-structures for 500 sentences randomly chosen from one half of KTC4. To create the gold standard, f-structure functional equations are annotated automatically by the method developed in my research without zero-pronoun identification and then the f-structures are man-

ually corrected and extended.

The types of corrections and extensions are as follows:

- Zero pronoun annotation
- Correction of dependency relations (grammatical functions)
- Correction of morphological analysis (grammatical features)

Zero pronouns in the 500 Gold Standard f-structures are added manually, based on the context in which each of them appears in the original text, verbal morphology, and A Japanese Lexicon ([Ikehara et al., 1999]), a hand-coded Japanese case-frame dictionary.

Table 7.1 shows the numbers of core arguments in the 500 Gold Standard f-structures, and the numbers of zero pronouns for each core argument. Almost all subjects are zero-pronouns, and most obliques are not overtly realised:

Table 7.1: The numbers of SUBJ, OBJ and OBL arguments and the numbers of zero pronouns for each core grammatical function argument in the Gold Standard f-structures:

Grammatical functions	Token numbers	Token numbers of pro
SUBJ	1720	1667 (approx. 96% of all SUBJ)
OBJ	526	198 (approx. 37% of all OBJ)
OBL	740	456 (approx. 61% of all OBL)

Each of 500 f-structures acquired by the method described in the previous chapter (for both the original KTC4 treebank representations and KNP parser output) are converted into dependency triples, and these triples are compared with the dependency triples of the Gold Standard f-structures, and precision, recall and f-score are calculated using the software of [Crouch et al., 2002].

7.3 Overall Results Using KTC4 Treebank Representations Without Zero Pronoun Identification

The overall results for all grammatical functions and features using KTC4 treebank representations are shown in Table 7.2:

Table 7.2: Overall Results for All Grammatical Functions and Features Using KTC4 Treebank Representations

Feature	precision	recall	f-score
All	23144 / 23412 = 0.9885	23144 / 25285 = 0.9153	0.9505
subj	300 / 300 = 1.0	300 / 1720 = 0.1744	0.2970
obj	414 / 417 = 0.9928	414 / 526 = 0.787	0.8780
obl	397 / 398 = 0.9974	397 / 740 = 0.5364	0.6977
adj	456 / 457 = 0.9978	456 / 457 = 0.9978	0.9978
padj	995 / 997 = 0.9979	995 / 997 = 0.9979	0.9979
sadj	423 / 427 = 0.9906	423 / 444 = 0.9527	0.9712
rel	300 / 410 = 0.7317	300 / 307 = 0.9771	0.8368
comp	155 / 160 = 0.9687	155 / 155 = 1.0	0.9841
apcomp	69 / 73 = 0.9452	69 / 165 = 0.4181	0.5798
topic	285 / 285 = 1.0	285 / 285 = 1.0	1.0
focus	98 / 98 = 1.0	98 / 98 = 1.0	1.0
det	73 / 73 = 1.0	73 / 73 = 1.0	1.0
tense	1339 / 1341 = 0.9985	1339 / 1343 = 0.9970	0.9977
mood	1504 / 1507 = 0.9980	1504 / 1507 = 0.9980	0.9980
aspect	62 / 120 = 0.5166	62 / 120 = 0.5166	0.5166
voice	112 / 112 = 1	112 / 112 = 1	1
pos	4787 / 4797 = 0.9979	4787 / 4796 = 0.9981	0.9980
prtcs	1699 / 1703 = 0.9976	1699 / 1703 = 0.9976	0.9976
prtcnj	788 / 790 = 0.9974	788 / 790 = 0.9974	0.9974
prtadv	596 / 597 = 0.9983	596 / 597 = 0.9983	0.9983
prtend	18 / 18 = 1	18 / 18 = 1	1
infl	1652 / 1656 = 0.9975	1652 / 1656 = 0.9975	0.9975
v-infl	1187 / 1191 = 0.9966	1187 / 1191 = 0.9966	0.9966
suf-infl	427 / 431 = 0.9907	427 / 431 = 0.9907	0.9907
aux-infl	110 / 111 = 0.9909	110 / 111 = 0.9909	0.9909
adj-infl	245 / 245 = 1	245 / 245 = 1	1
copula-infl	109 / 109 = 1	109 / 109 = 1	1
app	28 / 28 = 1	28 / 28 = 1	1
prf	129 / 129 = 1	129 / 129 = 1	1
numeral	315 / 315 = 1	315 / 315 = 1	1
aux	110 / 111 = 0.9909	110 / 111 = 0.9909	0.9909
suf	927 / 931 = 0.9957	927 / 931 = 0.9957	0.9957
copula	109 / 109 = 1	109 / 109 = 1	1
style	224 / 224 = 1	224 / 224 = 1	1
negative	150 / 152 = 0.9868	150 / 152 = 0.9868	0.9868
nform	136 / 136 = 1	136 / 136 = 1	1
interrogative	41 / 41 = 1	41 / 41 = 1	1
coord	351 / 351 = 1	351 / 351 = 1	1
r-parenthesis	138 / 140 = 0.9857	138 / 140 = 0.9857	0.9857
l-parenthesis	139 / 140 = 0.9928	139 / 140 = 0.9928	0.9928
comma	694 / 698 = 0.9942	694 / 698 = 0.9942	0.9942
period	510 / 510 = 1	510 / 510 = 1	1

Most of the grammatical features are properly analysed, while the performance for grammatical functions SUBJ, OBJ, OBL and APP are relatively low, hence additional procedures are required to improve the performance.

The overall results for all grammatical functions and features using KNP parser output are shown in Table 7.3:

Table 7.3: Overall Results for All Grammatical Functions and Features Using KNP parser output

Feature	precision	recall	f-score
All	20222 / 23697 = 0.8533	20222 / 25284 = 0.7997	0.8257
subj	238 / 297 = 0.8013	238 / 1720 = 0.1383	0.2359
obj	358 / 417 = 0.8585	358 / 526 = 0.6806	0.7592
obl	316 / 391 = 0.8081	316 / 740 = 0.4270	0.5587
adj	236 / 461 = 0.5119	236 / 457 = 0.5164	0.5141
padj	780 / 1016 = 0.7677	780 / 997 = 0.7823	0.7749
sadj	307 / 426 = 0.7206	421 / 444 = 0.9481	0.7090
rel	260 / 416 = 0.625	260 / 307 = 0.8469	0.7192
comp	143 / 538 = 0.2657	143 / 155 = 0.9225	0.4126
apcomp	57 / 61 = 0.9344	57 / 165 = 0.3454	0.5044
topic	226 / 285 = 0.7929	226 / 285 = 0.7929	0.7929
focus	75 / 95 = 0.7894	75 / 98 = 0.7653	0.7772
det	62 / 77 = 0.8051	62 / 73 = 0.8493	0.8266
tense	1339 / 1341 = 0.9985	1339 / 1343 = 0.9970	0.9977
mood	1451 / 1498 = 0.9686	1451 / 1507 = 0.9628	0.9657
aspect	62 / 120 = 0.5166	62 / 120 = 0.5166	0.5166
voice	110 / 121 = 0.9090	110 / 112 = 0.9821	0.9442
pos	4184 / 4790 = 0.8734	4184 / 4796 = 0.8723	0.8729
prtcs	1554 / 1733 = 0.8967	1554 / 1703 = 0.9125	0.9045
prtcnj	696 / 777 = 0.8957	696 / 790 = 0.8810	0.8883
prtadv	562 / 598 = 0.9397	562 / 597 = 0.9413	0.9405
prtend	9 / 9 = 1.0	9 / 18 = 0.5	0.6666
infl	1570 / 1647 = 0.9532	1570 / 1656 = 0.9480	0.9506
v-infl	1137 / 1198 = 0.9490	1137 / 1191 = 0.9546	0.9518
suf-infl	415 / 429 = 0.9673	415 / 431 = 0.9628	0.9651
aux-infl	101 / 111 = 0.9099	101 111 = 0.9099	0.9099
adj-infl	240 / 243 = 0.9876	240 / 245 = 0.9795	0.9836
copula-infl	93 / 97 = 0.9587	93 / 109 = 0.8532	0.9029
app	14 / 29 = 0.4827	14 / 28 = 0.5	0.4912
prf	97 / 133 = 0.7293	97 / 129 = 0.7519	0.7404
numeral	56 / 316 = 0.1772	56 / 315 = 0.1777	0.1774
aux	101 / 111 = 0.9099	101 / 111 = 0.9099	0.9099
suf	643 / 935 = 0.6877	643 / 931 = 0.6906	0.6891
copua	93 / 97 = 0.9587	93 / 109 = 0.8532	0.9029
style	204 / 210 = 0.9714	204 / 224 = 0.9107	0.9400
negative	136 / 145 = 0.9379	136 / 152 = 0.8947	0.9157
nform	136 / 136 =1	136 / 136 = 1	1
interrogative	39 / 68 = 0.5735	39 / 41 = 0.9512	0.7155
coord	246 / 336 = 0.7321	246 / 351 = 0.7008	0.7161
r-parenthesis	133 / 140 = 0.95	133 / 140 = 0.95	0.95
l-parenthesis	130 / 140 = 0.9285	130 / 140 = 0.9285	0.9285
comma	639 / 696 = 0.9181	639 / 698 = 0.9154	0.9167
period	489 / 516 = 0.9476	489 / 510 = 0.9588	0.9532

The performance for grammatical functions and grammatical features in KNP output is lower than that in KTC4 treebank representations. This is due to the noise in the morphological analysis by JUMAN, for e.g. homophonic particles (e.g., "-ka" as a conjunctive particle and a sentence-ending particle), homophonic verbs (e.g., when "aw-" (meet) in past tense is written in hiragara, it is homophoric with "ar-" in past tense; "atta"), and incorrect syntactic unit chunking by KNP (e.g., for some instances, words without particle functioning as adverbial units are incorrectly analysed as part of the next following unit). In order to obtain higher performance for the automatic annotation of LFG functional equations on KNP output, we need to improve the performance of JUMAN and KNP. This is a topic of future research.

7.4 Zero Pronoun Identification

7.4.1 Experiment 1: Zero Pronoun Identification for KTC4

For the evaluation of zero-pronoun identification, the KTC4 treebank representations corresponding to the 500 gold standard f-structures are automatically converted into f-structures with the following different methods for zero pronoun identification:

Method 1 Null Method: ignore all zero pronouns.

Method 2 Simplistic Method: add pro-SUBJ, pro-OBJ and pro-OBL whenever full NPs with the particle "-ga", "-wo" or "-ni" are missing for local verbs, regardless of the case frame of the verb.

Method 3 Morphological Method: use the list of verbs whose morphology unambiguously specifies their transitivity. The list is automatically constructed from KTC4 (except for the Gold Standard sentences), based on the morphology of the verbs as described in Section 3.5.1.

Method 4 Probabilistic Method: use the list of morphologically ambiguous verbs with high transitivity rate. The list of verbs and their transitivity rate is automatically acquired from one half of KTC4, which contains no Gold Standard sentences. In this experiment, the threshold of the transitivity rate was 0.3; that is, the verbs whose transitivity rates are above 0.3 are assumed to be transitve verbs, hence included into the list.

Method 5 Combination Method: add to the list of verbs in the third method those verbs whose morphology does not specify their transitivity but whose transitivity rate is high (as in Method 4), and use this combined list.

Method 2 serves as the lower bound of the zero-pronoun identification method.

It is expected that Method 2 will yield a good result in terms of recall, but that precision will be quite low. Method 3 is expected to show an improvement in terms of precision, but at the same time lower coverage because of the limited number of morphologically identifiable transitive verbs. The performance of Method 4 and 5 are more difficult to predict.

Table 7.4 shows the results of the five methods. The figures in parentheses are the recall, precision and f-score for zero pronouns only. "Pred-only" in the table means the result includes the precision, recall and f-score of dependency triples of the predicates, arguments and adjuncts in the 500 test sentences, but not atomic features such as tense, mood, aspect features:

Table 7.4: Results of Experiment 1

		Precision	Recall	F-score
Method 1 (Null)	Pred-only	95.71	75.18	84.22
	SUBJ	100.0(0)	17.44(0)	29.70(0)
	OBJ	99.28(0)	78.7(0)	87.80(0)
	OBL	99.74(0)	53.64(0)	69.77(0)
Method 2 (Simplistic)	Pred-only	78.22	95.51	86.01
	SUBJ	98.64(98.91)	97.38(97.60)	98.00(98.25)
	OBJ	39.47(14.13)	97.67(95.80)	56.22(24.62)
	OBL	39.25(19.10)	89.94(88.46)	54.65(31.41)
Method 3 (Morphological)	Pred-only	95.75	92.69	94.20
	SUBJ	98.64(98.91)	97.38(97.60)	98.00(98.25)
	OBJ	92.83(71.55)	88.01(58.04)	90.35(64.09)
	OBL	92.48(88.05)	68.28(28.36)	78.55(42.90)
Method 4 (Probabilistic)	Pred-only	95.19	93.31	94.24
	SUBJ	98.64(98.91)	97.38(97.60)	98.00(98.25)
	OBJ	97.97(87.09)	77.99(18.88)	86.94(31.03)
	OBL	82.17(63.92)	82.32(67.30)	82.24(65.56)
Method 5 (Combination)	Pred-only	95.08	94.37	94.72
	SUBJ	98.64(98.91)	97.38(97.60)	98.00(98.25)
	OBJ	93.26(76.29)	91.59(72.02)	92.41(74.09)
	OBL	84.46(68.65)	81.97(66.34)	83.19(67.47)

In all methods except for Method 1, SUBJ-pro is added trivially; since every verb subcategorises for a subject, if a clause lacks a subject NP, then SUBJ-pro is added to the clause. However, this does not yet yield 100% accuracy because of incorrect annotation of functional equations, especially those on nominal predicates functioning as sentential adjuncts, hence more cleaning-up operations are required.

Trivially, Method 2 yields higher results for SUBJ-pro than for OBJ and OBL. Compared to Method 2, Method 3 yields better results for OBJ than for OBL, due to the fact that passive, causative and benefactive voices, which are projected by verbal suffixes (cf. Section 4.3.4), are less frequent than other voices. Passive, causative and benefactive voices are expressed morphologically, but there is no other morphological clues to show that a given verb subcategorises for OBL. Hence, in order to identify OBL zero pronouns, morphological information might not be as helpful as the syntactic environment. This claim can be supported by the relatively higher result for OBL zero pronoun identification in Method 4, which uses the rate that a given verb appears with an oblique NP. The lower recall on OBJ in Method 4 reflects the small size of the verb list extracted from KTC4; extraction from a larger amount of text might improve the result. Method 5 yields the best pred-only f-score among these methods, due to the slight improvement in OBJ.

The results show that Method 5 is currently the best for OBJ zero pronoun identification. The results of zero pronoun identification for OBL is lower than that for OBJ, because of the ambiguity of "-ni" marked NPs. This particle can only be used as the OBL case marker, or as a postposition which functions as a temporal or a locative adverbial. This ambiguity of the case-marker "-ni" can only be resolved by a more fine-grained semantic-based improvement of the annotation algorithm.

7.4.2 Experiment 2: Zero Pronoun Identification for KNP Output

The next experiment explores how the methods evaluated in Experiment 1 can identify zero pronouns in raw texts, using the KNP dependency parser. For the experiment, I removed the annotations on the 500 Gold Standard sentences, and parsed these raw sentences with KNP. The parser output is automatically annotated with f-structure functional equations, and zero pronouns are identified using the same methods as in Experiment 1. The output f-structures are converted into dependency triples and compared to the original Gold Standard triples. Table 7.5 shows the results of each zero pronoun identification method.

Table 7.5: Results of Experiment 2

		Precision	Recall	F-score
Method 1 (null)	Pred-only	81.91	64.82	72.37
	SUBJ	80.13(0)	13.83(0)	23.59(0)
	OBJ	85.85(0)	68.06(0)	75.92(0)
	OBL	80.81(0)	42.70(0)	55.87(0)
Method 2 (Simplistic)	Pred-only	68.01	83.62	75.01
	SUBJ	89.60(92.16)	88.84(90.33)	89.21(92.04)
	OBJ	35.88(12.88)	88.90(87.41)	51.12(22.45)
	OBL	34.68(16.63)	79.89(78.36)	48.36(27.43)
Method 3 (Morphological)	Pred-only	83.41	80.91	82.14
	SUBJ	89.60(92.16)	88.84(90.33)	89.21(92.04)
	OBJ	85.26(65.95)	77.63(43.35)	81.26(52.31)
	OBL	84.96(82.85)	61.69(27.88)	71.47(41.72)
Method 4 (Probabilistic)	Pred-only	82.68	81.53	82.10
	SUBJ	89.60(92.16)	88.84(90.33)	89.21(92.04)
	OBJ	89.31(84.00)	70.30(14.68)	78.67(24.99)
	OBL	72.82(53.33)	72.44(57.69)	72.62(55.42)
Method 5 (Combination)	Pred-only	82.63	82.14	82.38
	SUBJ	89.60(92.16)	88.84(90.33)	89.21(92.04)
	OBJ	85.82(68.91)	77.99(44.75)	81.71(54.23)
	OBL	72.82(56.33)	72.44(57.69)	72.62(57.00)

Table 7.5 shows that the overall result is lower than Experiment 1. The five methods do not fix incorrect dependencies in the KNP output. As pointed out in Section 7.2, we need to improve the parse quality of KNP in order to obtain higher results in the automatic annotation of LFG functional equations to KNP output.

Similar to Experiment 1, Method 5 yields the best pred-only f-score compared to the other methods. The experiments show that the morphology-based approach and the probability-based approach improve the f-scores of the annotation algorithm in terms of the pred-only f-scores of the sentence as a whole.

However, these two approaches do not properly identify zero pronouns as precisely as expected: for example, the f-score for zero-pronoun OBJ in Method 5 for KNP parser output is only about 54%.

[Kawahara and Kurohashi, 2004b] constructed case frames for 23,000 predicates from KNP parser output of newspaper articles of 20 years (about 21,000,000 sentences), and used them for zero-pronoun identification and

resolution for 100 articles (the number of sentences is not specified). The precision and recall of zero-pronoun identification are 87.1% and 74.8%, respectively. Their case frames include OBJ nouns and OBL nouns which are used as arguments of each verb in the original articles; if a given verb has a case frame with a "-wo"-marked OBJ noun_1 and a "-ni"-marked OBL noun_2, and the test text contains the same verb used with an OBL noun_2 but not with an OBJ, then it is assumed that this verb has an OBJ zero pronoun. Besides, if the noun_1 appears somewhere near the sentence in which the verb appears, then this noun is more probable than other nouns to function as the unrealised OBJ of this verb.

There are a number of possible ways to improve each of the zero-pronoun methods: for example, for the morphological approach, it is possible to construct a more precise and wide-coverage list of transitive verbs and use it for our purpose. However, this task implies manual extraction of transitive verbs which do not have their intransitive counterparts, and employing the extracted list for zero pronoun identification. This process itself might be linguistically interesting, even though it will be time-consuming and runs against the general aim of automatic extraction of linguistic resources. For the probabilistic method, calculating the transitivity rate of more verbs would improve the result, but this implies that we need a Japanese Treebank which is larger and has wider coverage than KTC4, which is still not available to date. In order to yield better results for zero pronoun identification, it seems to be necessary to try alternative approaches along with improving the methods now available, which is one of the objectives of my future work.

7.5 Summary

This chapter explained how the automatic annotation of LFG functional equations for KTC4/KNP is evaluated, discussed the results, and outlined future research. First, the f-structures acquired by the automatic annotation of KTC4 representations are evaluated against Gold-Standard f-structures for 500 sentences randomly extracted from KTC4. These Gold-Standard f-structures are automatically annotated and then manually corrected. Different methods for zero-pronoun identification are compared in terms of precision, recall and f-scores for grammatical functions. It is shown that zero-pronoun identification using both morphological and probabilistic methods yields the best result so far. Second, the same 500 sentences (but without KTC4 annotations) are parsed with KNP, then the KNP output is automatically annotated with LFG functional equations, and the acquired f-structures are compared against the Gold Standard f-structures. The results are lower

than those of KTC4, due to the noise introduced by the parser compared to the KTC4 treebank representations.

Chapter 8

Conclusion

The objective of this thesis is to design, implement and evaluate the automatic acquisition of wide-coverage treebank-based deep linguistic resources for Japanese, as part of GramLab project which focuses on the automatic treebank-based induction of multilingual resources in the framework of Lexical-Functional Grammar (LFG).

Chapter 1 introduced the motivation of this thesis and the methodology of Treebank-Based Automatic Acquisition of Deep Linguistic Resources based on LFG, and its application to Japanese. I argued that the morphology tags and unlabelled dependency tags provided by Kyoto Text Corpus ver.4 (KTC4) provide us with sufficient information to automatically construct "proto-" f-structures for the text in the corpus.

Chapter 2 described the basic framework of LFG in more detail, including the correspondence between different levels of linguistic representation, functional well-formedness, subcategorisation frames of verbal predicates, long-distance dependency, control, and anaphora.

Chapter 3 provided a general description of core syntactic and morphological aspects of Japanese: non-configurationality; the idea of "bunsetsu" or syntactic units and their dependency relationship represented as Directed Acyclic Graphs (DAG); topicalisation by a particular particle; and frequent use of zero-pronouns with or without overt antecedents. Inflecting parts-of-speech and non-inflecting parts-of-speech of Japanese are also described with examples.

Chapter 4 gives the linguistic representation of core grammatical features and functions of Japanese in the framework of LFG. In this chapter, I used Directed Acyclic Graphs (DAG) as a framework for surface syntactic representation of Japanese and provided more fine-grained LFG f-structure analyses. The morphological information in one syntactic unit is combined with that in another unit through labelling the dependency arc between them,

and this combination continues until all the information is gathered at the root unit, which corresponds to the f-structure for the sentence as a whole.

Chapter 5 introduces KNP and the Kyoto Corpus and describes KNP's algorithm for parsing Japanese text, and the Kyoto Corpus representation format. Since KNP is a rule-based parser, and the rules are all hand-coded, it has taken a lot of time and effort to complete it. It will be possible to learn a new parser based on the deep linguistic resources acquired by the method I have presented in this thesis. One way to do this is to employ the DAG-style representation of dependency among syntactic units in Japanese sentences in a tagged corpus or raw text for the development of a probabilistic dependency-based parsing model, which provides clues to determine what head a given unit with certain morphological characteristics is most likely to depend on.

Chapter 6 introduces the LFG annotation method for KTC4 and KNP output. This thesis concentrated on the issue of zero-pronoun identification, while other issues such as tense-mood-aspect disambiguation, the distinction of relative clauses and appositional complement, as well as correction of incorrect morphological analyses remain for further work.

Evaluations of the performance of the LFG annotation method for Japanese are presented and discussed in Chapter 7, and it is shown that zero-pronoun identification using both morphological and probabilistic methods yields the best result so far. Using KTC4 treebank trees, currently my method achieves a pred-only dependency f-score of 94.72%. The parsing experiments using KNP output yield a pred-only dependency f-score of 82.38%. I conclude that more sophisticated methods for zero-pronoun identification and more text data are required to obtain a more accurate, wider-coverage automatic acquisition algorithm for deep linguistic resources for Japanese in the framework of Lexical-Functional Grammar.

Bibliography

[Alsina, 1996] Alsina, A. (1996). *The Role of Argument Structure in Grammar: Evidence from Romance*. CSLI Publications, Stanford, CA.

[Bresnan, 1978] Bresnan, J. (1978). A realistic transformational grammar. In Halle, M., Bresnan, J., and Miller, G., editors, *Linguistic Theory and Psychological Reality*, pages 1–59. The MIT Press, Cambridge, MA.

[Bresnan, 1982a] Bresnan, J. (1982a). Control and complementation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 282–390. The MIT Press, Cambridge, MA.

[Bresnan, 1982b] Bresnan, J., editor (1982b). *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA.

[Bresnan, 2001] Bresnan, J. (2001). *Lexical-Functional Syntax*. Blackwell Publishers, Oxford, UK.

[Bresnan and Kanerva, 1989] Bresnan, J. and Kanerva, J. (1989). Locative inversion in Chichewa: A case study of factorization in grammar. *Linguistic Inquiry*, 20(1):1–50.

[Bresnan and Kaplan, 1982] Bresnan, J. and Kaplan, R. M. (1982). Lexical-Functional Grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge, MA.

[Burke et al., 2004] Burke, M., Lam, O., Cahill, A., Chan, R., O'Donovan, R., Bodomo, A., van Genabith, J., and Way, A. (2004). Treebank-based acquisition of a Chinese Lexical-Functional Grammar. In *Proceedings of the PACLIC-18 Conference*, pages 161–172, Waseda University, Tokyo, Japan.

[Butt et al., 2002] Butt, M., Dyvik, H., King, T. H., Masuichi, H., and Rohrer, C. (2002). The Parallel Grammar project. In *COLING-02 on Grammar engineering and evaluation*, pages 1–7.

[Cahill, 2004] Cahill, A. (2004). *Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations*. PhD thesis, Dublin City University, Ireland.

[Cahill et al., 2003] Cahill, A., Burke, M., O'Donovan, R., van Genabith, J., and Way, A. (2003). Treebank-based multilingual unification-grammar development. In *Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development*, Vienna, Australia.

[Cahill et al., 2004] Cahill, A., Burke, M., O'Donovan, R., van Genabith, J., and Way, A. (2004). Long-distance dependency resolution in automatically acquired wide-coverage PCFG-based LFG approximations. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 320–327, Barcelona, Spain.

[Cahill et al., 2005] Cahill, A., Forst, M., Burke, M., McCarthy, M., O'Donovan, R., Rohrer, C., van Genabith, J., and Way, A. (2005). Treebank-based acquisition of multilingual unification grammar resources. In Bender, E., Flickinger, D., Fouvry, F., and Siegel, M., editors, *Journal of Research on Language and Computation*, pages 247–279. Kluwer Academic Press.

[Cahill et al., 2002] Cahill, A., McCarthy, M., van Genabith, J., and Way, A. (2002). Automatic annotation of the penn-treebank with lfg f-structure information. In *Proceedings of the LFG '02 Conference*, Athens, Greece.

[Chamberlain, 1907] Chamberlain, B. H. (1907). *A Handbook of Colloquial Japanese*. Crosby Lockwood and Son and Kelly and Walsh, ltd., Yokohama, Japan.

[Chomsky, 1981] Chomsky, N. (1981). *Lectures on Government and Binding*. Foris, Dordrecht.

[Chomsky, 1995] Chomsky, N. (1995). *The Minimalist Program*. The MIT Press, Cambridge, MA.

[Chrupala and van Genabith, 2006] Chrupala, G. and van Genabith, J. (2006). Using machine-learning to assign function labels to parser output for spanish. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 136–143, Sydney, Australia.

[Crouch et al., 2002] Crouch, R., Kaplan, R., King, T. H., and Riezler, S. (2002). A comparison of evaluation metrics for a broad-coverage stochastic parser. In *Proceedings of the Workshop on "Parseval and Beyond" at*

the 3rd International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Spain.

[Dalrymple, 2001] Dalrymple, M. (2001). *Syntax and Semantics vol.34, Lexical Functional Grammar*. Academic Press, London, UK.

[Debusmann, 2003] Debusmann, R. (2003). Dependency grammar as graph description. In Duchier, D., editor, *Prospects and Advances of the Syntax/Semantics Interface*, pages 79–84, Nancy, France.

[Dowty and Brodie, 1984] Dowty, D. R. and Brodie, B. (1984). The semantics of 'floated' quantifiers in a transformationless grammar. In *Proceedings of the 4th West Coast Conference on Formal Linguistics*, pages 75–90, University of California, San Diego.

[Duchier and Debusmann, 2001] Duchier, D. and Debusmann, R. (2001). Topological Dependency Trees: A constraint-based account of linear precedence. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 180–187, Toulouse, France.

[Falk, 2001] Falk, Y. (2001). *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. CSLI Publications, Stanford, CA.

[Fujimoto et al., 2002] Fujimoto, T., Ikehara, S., Murakami, J.-i., and Omote, K. (2002). Fukubun-ni okeru soko-no meishi-to shuushokubu-no uchi-to soto-no kankei-no handankisoku, "The criteria for the internal and external relationships between the head noun and the dependent clause in complex sentences". In *The 8th Annual Conference of Natural Language Processing*, Kyoto, Japan. In Japanese.

[Grishman et al., 1994] Grishman, R., Macleod, C., and Meyers, A. (1994). Comlex syntax: Building a computational lexicon. In *Proceedings of COLING 94*, pages 268–272, Kyoto, Japan.

[Guo et al., 2007] Guo, Y., van Genabith, J., and Wang, H. (2007). Treebank-based acquisition of LFG resources for chinese. In *Proceedings of the LFG07 Conference*, pages 214–232, Stanford,CA.

[Haggarty, 2002] Haggarty, R. (2002). *Discrete Mathematics for Computing*. Pearson Education Limited, Essex, UK.

[Hashimoto, 1948] Hashimoto, S. (1948). *Kokugohou Kenkyu, "A Study of Japanese Grammar"*. Iwanami Shoten, Tokyo, Japan. (in Japanese).

[Hisamitsu and Nitta, 1991] Hisamitsu, T. and Nitta, Y. (1991). Morphological analysis of Japanese sentences by minimum connective-cost method. In *Proceedings of the 42nd Conference of Information Processing Society in Japan*, pages 1–2, Hachioji, Tokyo.

[Hockenmaier and Steedman, 2002] Hockenmaier, J. and Steedman, M. (2002). Acquiring compact lexicalized grammars from a cleaner treebank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1974–1981, Las Palmas, Spain.

[Ikehara et al., 1999] Ikehara, S., Miyazaki, M., Shirai, S., A.Yokoo, Nakaiwa, H., Ogura, K., Ooyama, Y., and Hayashi, Y. (1999). *Nihongo Goi Taikei, "A Japanese Lexicon"*. Iwanami Shoten, Tokyo, Japan. (in Japanese).

[Kaplan and Zaenen, 1989] Kaplan, R. and Zaenen, A. (1989). Long-distance dependencies, constituent structure, and functional uncertainty. In Baltin, M. R. and Kroch, A., editors, *Alternative Conceptions of Phrase Structure*, pages 17–42. University of Chicago Press, Chicago, IL.

[Kawahara and Kurohashi, 2002] Kawahara, D. and Kurohashi, S. (2002). Fertilization of case frame dictionary for robust Japanese case analysis. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 425–431, Taipei, Taiwan.

[Kawahara and Kurohashi, 2004a] Kawahara, D. and Kurohashi, S. (2004a). Improving Japanese zero pronoun resolution by global word sense disambiguation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING2004)*, Geneva, Switzerland.

[Kawahara and Kurohashi, 2004b] Kawahara, D. and Kurohashi, S. (2004b). Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 334–341, Hainan Island, China.

[Kawahara and Kurohashi, 2005] Kawahara, D. and Kurohashi, S. (2005). Gradual fertilization of case frames. *Journal of Natural Language Processing*, 12(2):109–131.

[Kuno, 1983] Kuno, S. (1983). *Ninhongo Bunpo Kenkyu, "A Study of Japanese Grammar"*. Daishukan, Tokyo, Japan.

[Kurohashi and Kawahara, 2005] Kurohashi, S. and Kawahara, D. (2005). *Japanese Morphological Analysis System JUMAN 5.1 Users Manual*. University of Tokyo.

[Kurohashi et al., 2005] Kurohashi, S., Kawahara, D., and Shibata, T. (2005). JUMAN/KNP wo mochiita keitaiso kaiseki koubunnkaiseki jisshu "A practice of morphological analysis and parsing using JUMAN/KNP". Available at http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp/20050830-practice.ppt.

[Kurohashi and Nagao, 1997] Kurohashi, S. and Nagao, M. (1997). Kyoto university text corpus project. In *3rd Anuual Meeting of the Association for Natural Language Processing*, pages 115–118. (in Japanese).

[Kurohashi and Nagao, 1998] Kurohashi, S. and Nagao, M. (1998). Building a Japanese parsed corpus while improving the parsing system. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, pages 719–724, Grenade, Spain.

[Levin, 1993] Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, IL.

[Magerman, 1995] Magerman, D. (1995). Parsing as statistical parsing recognition. Technical Report No.19443, IBM.

[Masuichi et al., 2003] Masuichi, H., Ohkuma, T., Yoshimura, H., and Harada, Y. (2003). Japanese parser on the basis of the Lexical-Functional Grammar formalism and its evaluation. In *Language, Information and Computation : Proceedings of the 17th Pacific Asia Conference*, pages 298–309, Sentosa, Singapore.

[Masuoka and Takubo, 1992] Masuoka, T. and Takubo, Y. (1992). *Kiso Nihongo Bumpo, "Basic Japanese Grammar"*. Kuroshio Shuppan, Tokyo, Japan.

[Matsumoto, 1996] Matsumoto, Y. (1996). *Complex Predicates in Japanese*. CSLI Publications, Stanford, CA.

[Miyagawa, 1989] Miyagawa, S. (1989). *Syntax and Semantics 22: Structure and Case Marking in Japanese*. Academic Press, New York.

[Miyao and Tsujii, 2005] Miyao, Y. and Tsujii, J. (2005). Probabilistic disambiguation models for wide-coverage HPSG parsing. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 83–90, Ann Arbor, Michigan.

[Nomura and Koike, 1992] Nomura, M. and Koike, K. (1992). *Nihongo Jiten, "Dictionary of Japanese Language"*. Tokyodo Shuppan, Tokyo, Japan.

[Noro et al., 2005] Noro, T., Koike, C., Hashimoto, T., Tokunaga, T., and Tanaka, H. (2005). Evaluation of a Japanese CFG derived from a syntactically annotated corpus with respect to dependency measures. In *The 5th Workshop on Asian Language Resources*, pages 9–16, Jeju Island, Korea.

[O'Donovan, 2006] O'Donovan, R. (2006). *Automatic Extraction of Large-Scale Multilingual Lexical Resources.* PhD thesis, Dublin City University, Ireland.

[O'Donovan et al., 2005] O'Donovan, R., Cahill, A., van Genabith, J., and Way, A. (2005). Automatic acquisition of Spanish LFG resources from the CAST3LB Treebank. In *Proceedings of the 10th International Conference of LFG*, pages 334–352, Bergen, Norway.

[Okuma et al., 2006] Okuma, T., Masuichi, H., and Yoshioka, T. (2006). Adapting Japanese LFG grammar to generation. *IPSJ SIG Technical Report.*

[Okutsu, 1978] Okutsu, K. (1978). *Bokuwa Unagidano Bunpo, "The Grammar of 'I am an eel"'.* Kuroshio Shuppan, Tokyo, Japan.

[Ota, 2007] Ota, K. (2007). MEni yoru nihongo kakariuke kaiseki, "dependency analysis of Japanese using ME. (in Japanese), Avalable at http://kzk9.net/publications/enshu3-dependency-analysis.pdf.

[Oya and van Genabith, 2007] Oya, M. and van Genabith, J. (2007). Automatic acquisition of lexical-functional grammar resources from a Japanese dependency corpus. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation.*

[Palmer, 2001] Palmer, F. (2001). *Mood and Modality 2nd ed.* Cambridge University Press, Cambridge, UK.

[Pollard and Sag, 1994] Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar.* University of Chicago Press, Chicago, IL.

[Postal, 1970] Postal, P. (1970). On coreferntial complement subject deletion. *Linguistic Inquiry*, 1:439–500.

[Rehbein and van Genabith, 2007] Rehbein, I. and van Genabith, J. (2007). Treebank annotation schemes and parser evaluation for german. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 630–639, Prague.

[Rubin, 1992] Rubin, J. (1992). *Gone Fishin': New Angles on Perennial Problems*. Kodansha International, Tokyo, Japan.

[Schluter and van Genabith, 2008] Schluter, N. and van Genabith, J. (2008). Treebank-based acquisition of lfg parsing resources for french. In *Proceedings of LREC 08*, Marrakech, Morocco.

[Shibatani, 1990] Shibatani, M. (1990). *The Languages of Japan*. Cambridge University Press, Cambridge, UK.

[Steedman, 2000] Steedman, M. (2000). *The Syntactic Process*. The MIT Press, Cambridge, MA.

[Suzuki et al., 2003] Suzuki, J., Hirao, T., Sasaki, Y., and Maeda, E. (2003). Hierarchical directed acyclic graph kernel: Methods for structured natural language data. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 32–39, Sapporo, Japan.

[Tsujimura, 2006] Tsujimura, N. (2006). *An Introduction to Japanese Linguistics* 2^{nd} *ed.* Blackwell Publishers, Oxford, UK.

[Uchimoto et al., 2000] Uchimoto, K., Murata, M., Sekine, S., and Isahara, H. (2000). Dependency model using posterior context. *Natural Language Processing*, 7(5):3–17.

[van Genabith et al., 1999] van Genabith, J., Way, A., and Sadler, L. (1999). Semi-automatic generation of f-structures from tree banks. In *Proceedings of the fourth International Conference on Lexical-Functional Grammar*, Manchester University.

[Wanner and Maratsos, 1978] Wanner, E. and Maratsos, M. (1978). An ATN approach to comprehension. In Halle, M., Bresnan, J., and Miller, G., editors, *Linguistic Theory and Psychological Reality*, pages 119–159. The MIT Press, Cambridge, MA.

[Yoshikawa, 1989] Yoshikawa, T. (1989). *Nihongo Bunpo Nyumon, "An Introduction to Japanese Grammar"*. ALC, Tokyo, Japan.

[Yoshioka et al., 2003] Yoshioka, T., Yoshimura, H., Masuichi, H., and Okuma, T. (2003). A proposal for Experience Knowledge Recycle system. *the 17th Annual conference of the Japanese Society for Artificial Intelligence.*